Faculté de Médecine
Aix*Marseille Université

SESSTIM UMR1252
Sciences Economiques et Sociales de la Santé & Traitement de l'Information Médicale
Inserm / IRD / Aix-Marseille Université

Principe de la Régression Linéaire

Plan

1. Question pratique
2. Définition de la régression
3. Estimation de la droite de régression
4. Test de la pente
5. Précision de la droite de régression
6. Adéquation du modèle
7. Régression Multiple

I. Question pratique

- Lien entre la taille et l'âge ?
- Quand l'âge ↑, la taille ↑ ?
- Connaissant l'âge, peut-on prédire la taille?
- **But médical**: détecter les retards de croissances

Exercice

- Exemple: étude du lien entre la taille et l'âge des filles (en mois), Echantillon de 637 filles

- Importer le fichier de données *FILLES.xls*

transformer le fichier filles.xls en FILLES.csv

```
ATF<-read.csv2("D:\\BIOSTAT\\FILLES.csv", header=TRUE)
```

- Moyenne globale de l'AGE
- m= ? mois
- Variance globale de l'AGE
- s^2= ? $mois^2$
- Graphiques

```
attach(ATF)
mean(AGE)
```

```
var(AGE)
```

```
par(mfrow=c(1,2))
hist(AGE, col="blue")
boxplot(AGE, col="blue")
```

Exercice

- Exemple: étude du lien entre la taille et l'âge des filles (en mois), Echantillon de 637 filles

- Importer le fichier de données *filles.xls*

transformer le fichier filles.xls en filles.csv

ATF<-read.csv2("D:\\BIOSTAT\\filles.csv", header=TRUE)

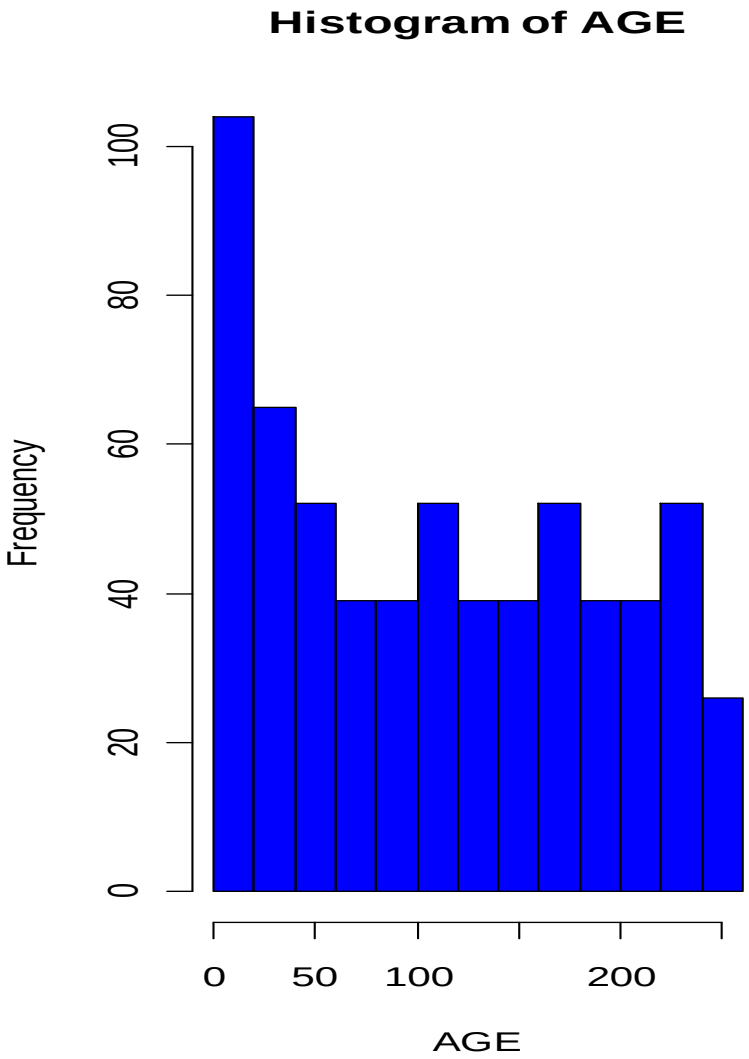
- Moyenne globale de l'AGE
- m= **112,12** mois

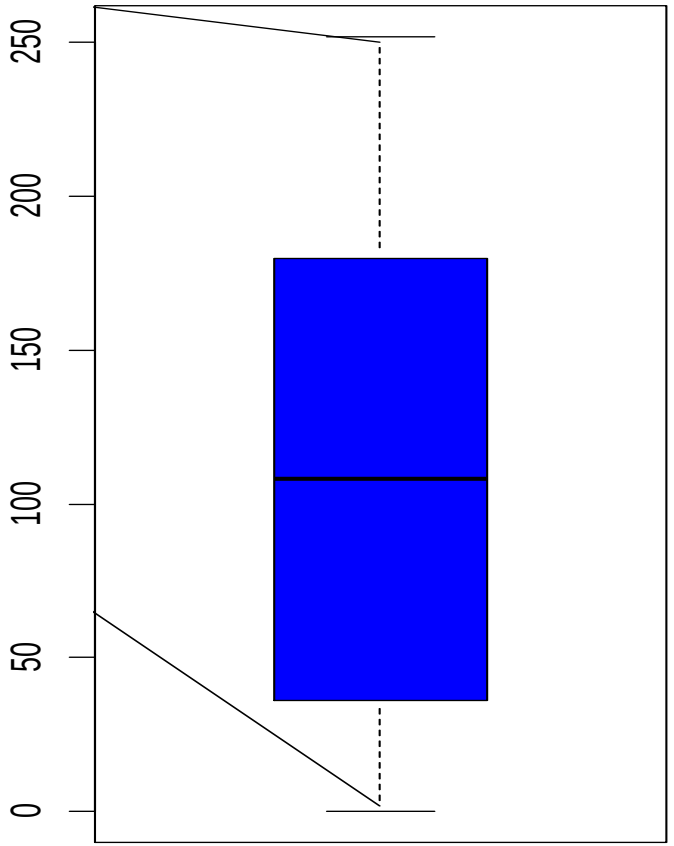
attach(ATF)
mean(AGE)

- Variance globale de l'AGE
- s^2= **6265,86** $mois^2$ *var(AGE)*
- Graphiques

par(mfrow=c(1,2))
hist(AGE, col="blue")
boxplot(AGE, col="blue")

Exercice


Histogram of AGE
Frequency
0
20
40
60
80
100
0
50
100
200
AGE


0
50
100
150
200
250

Exercice

- Exemple: étude du lien entre la taille et l'âge des filles (en mois), Echantillon de 637 filles

- Moyenne globale de la Taille (TAIL)
- m= ? cm
- Variance globale de la Taille (TAIL)
- s^2= ? cm^2
- Graphiques

mean(TAIL)

var(TAIL)

par(mfrow=c(1,2))
hist(TAIL, col="red")

boxplot(TAIL, col="red")

Exercice

- Exemple: étude du lien entre la taille et l'âge des filles (en mois), Echantillon de 637 filles

- Moyenne globale de la Taille (TAIL)
- m= **122,83** cm

mean(TAIL)

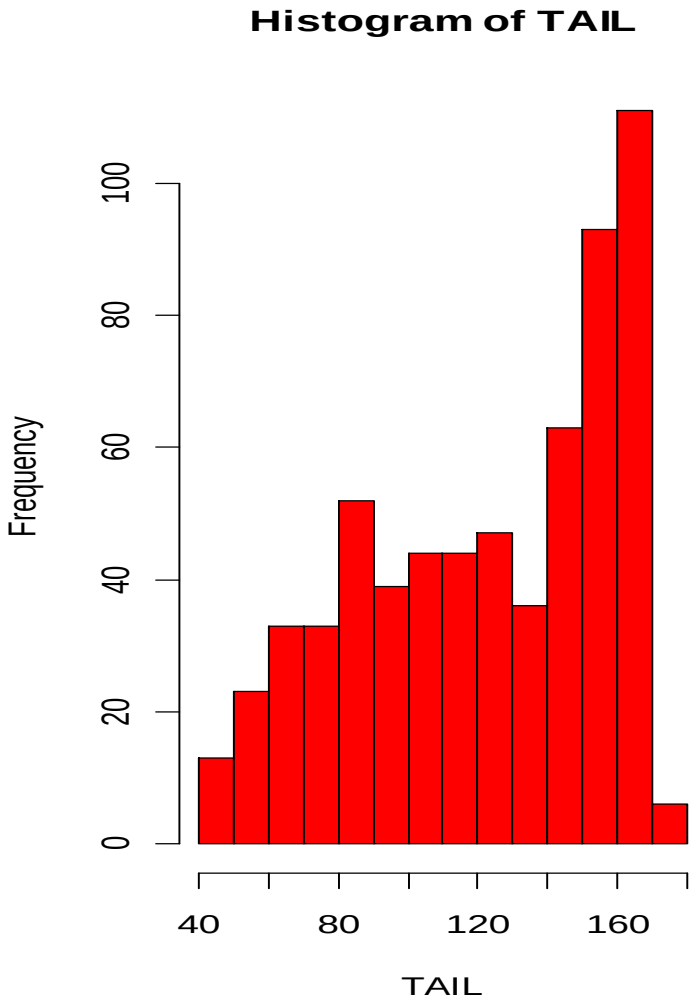
- Variance globale de la Taille (TAIL)
- s^2= **1317,43** cm^2

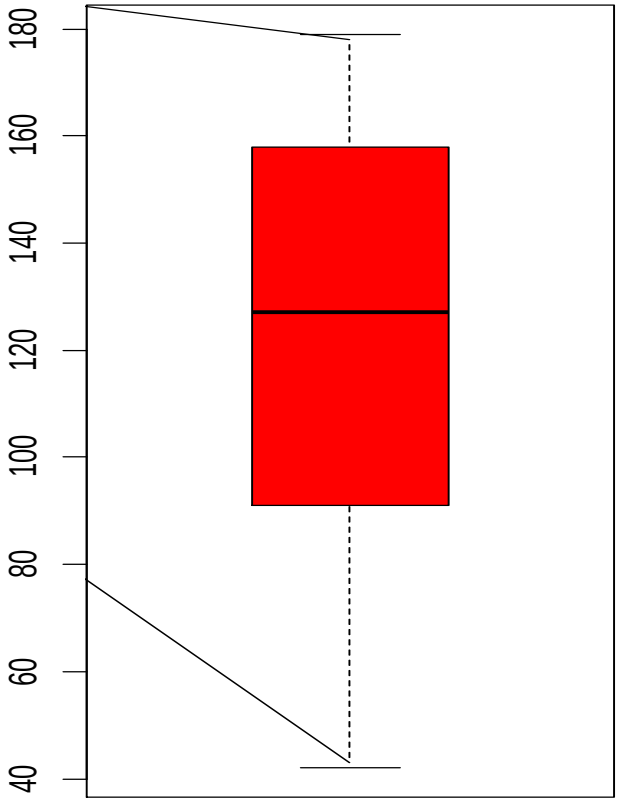
var(TAIL)

- Graphiques

par(mfrow=c(1,2))
hist(TAIL, col="red")
boxplot(TAIL, col="red")

Exercice



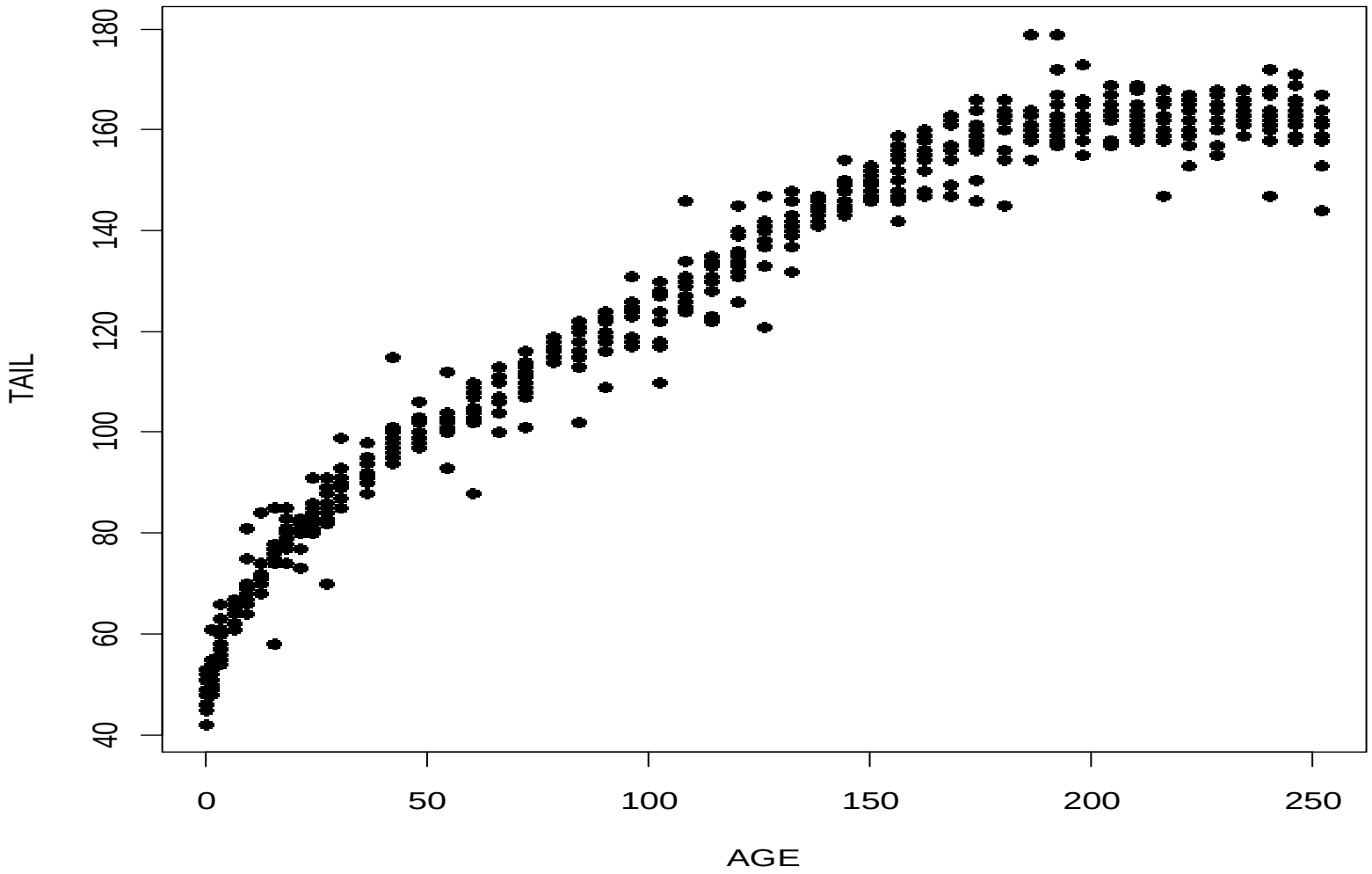


Exercice

- représenter la taille en fonction de l'age

plot(AGE, TAIL)

Exercice



II. Définition

- Régression de Y en X:
  - Y= taille (cm)
  - X= âge (mois)

Comment la Taille évolue **en fonction** de l'Age ?

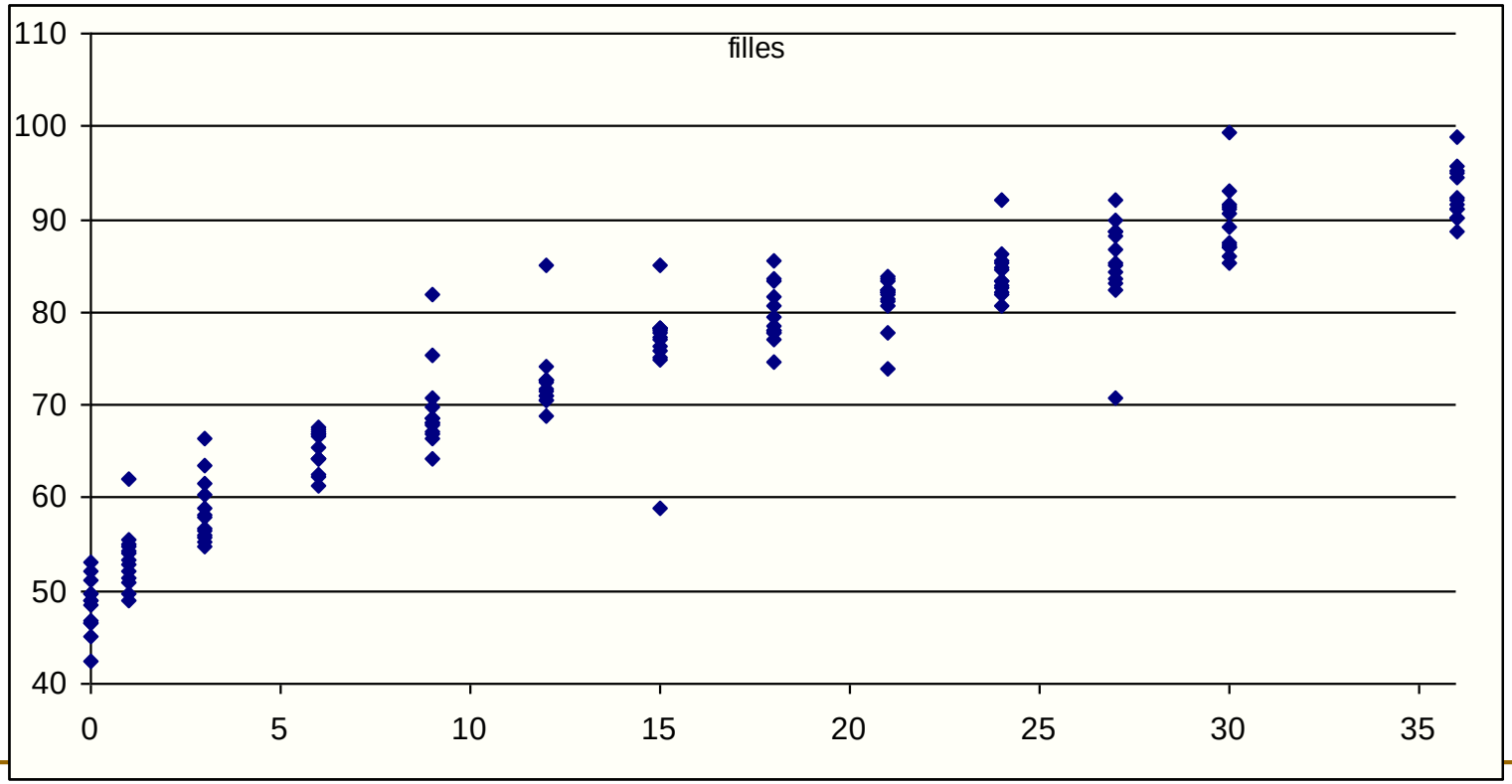
Taille= *f*(Age)

Comment évolue la Taille?

= Quelle valeur de la Taille ?

=>**Pour chaque Age**

=> **Sachant l'âge**

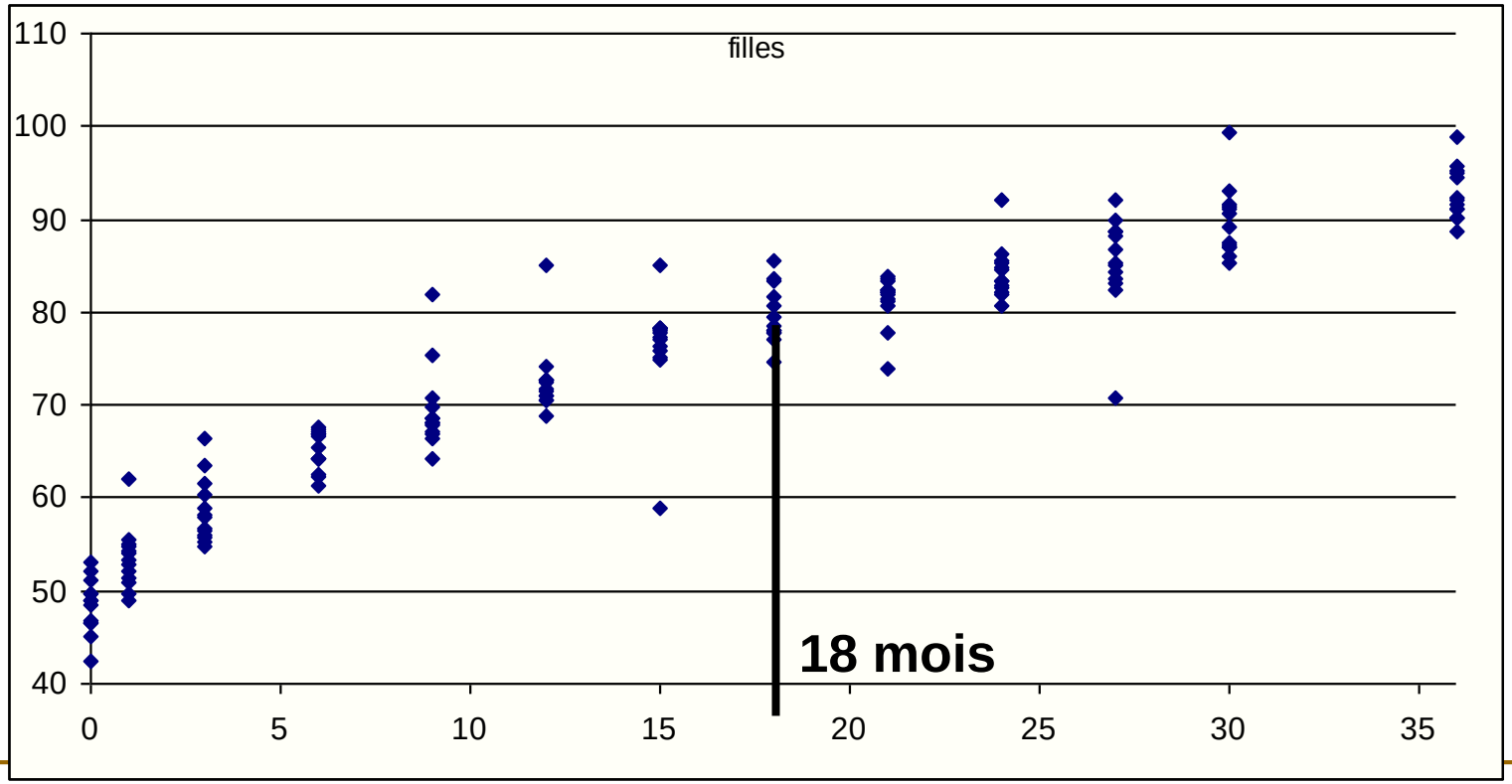


Comment évolue la Taille?

= Quelle valeur de la Taille ?

=>**Pour chaque Age**

=> **Sachant l'âge**

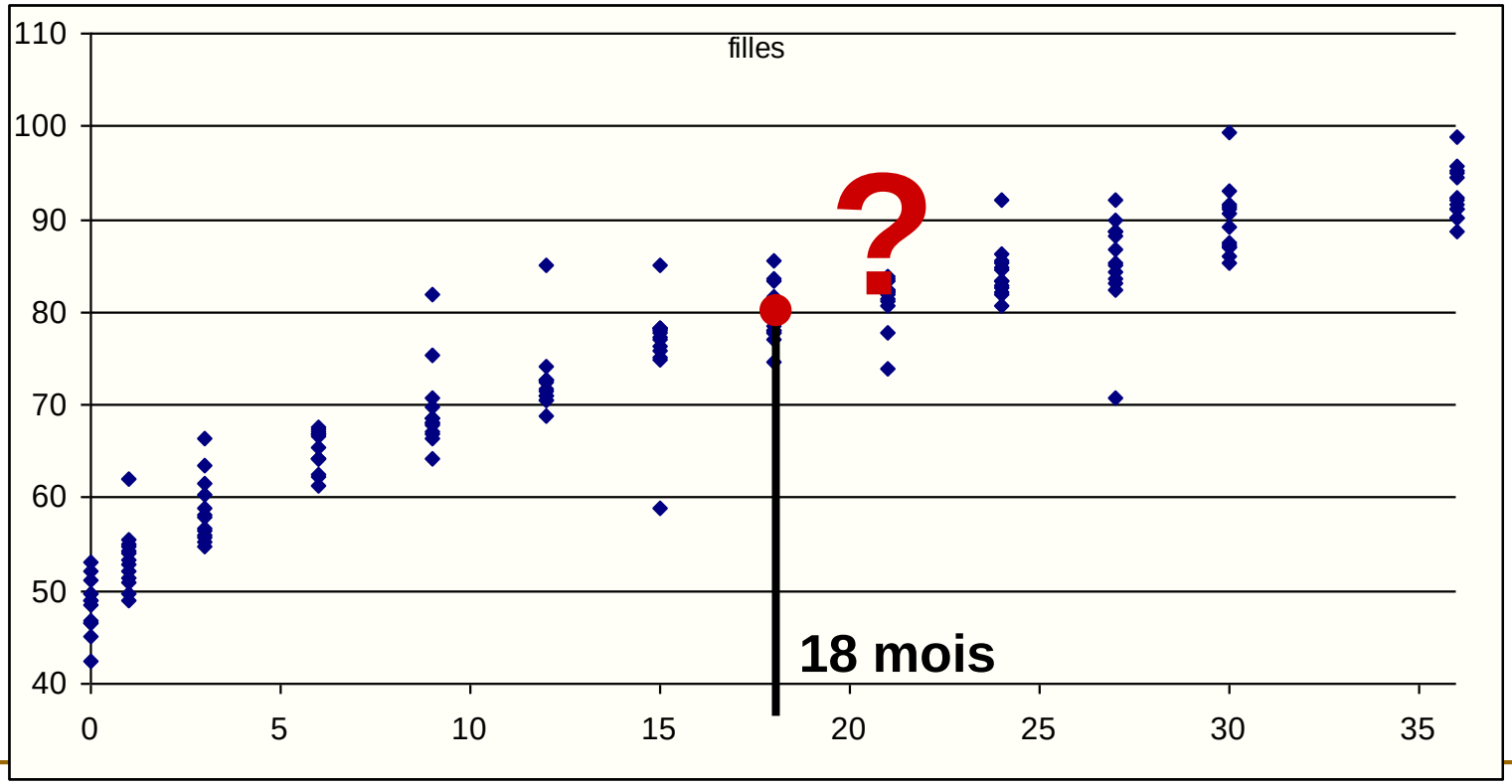


Comment évolue la Taille?

= Quelle valeur de la Taille ?

=>**Pour chaque Age**

=> **Sachant l'âge**



Exercice

- Chez les filles de 18 mois,
  - Qu'elle est la taille moyenne?

    mean(TAIL[AGE==18])
  - Qu'elle est la variance de la taille ?

    var(TAIL[AGE==18])
  - Qu'elle est la distribution ?

    hist(TAIL[AGE==18],col="magenta")

Exercice

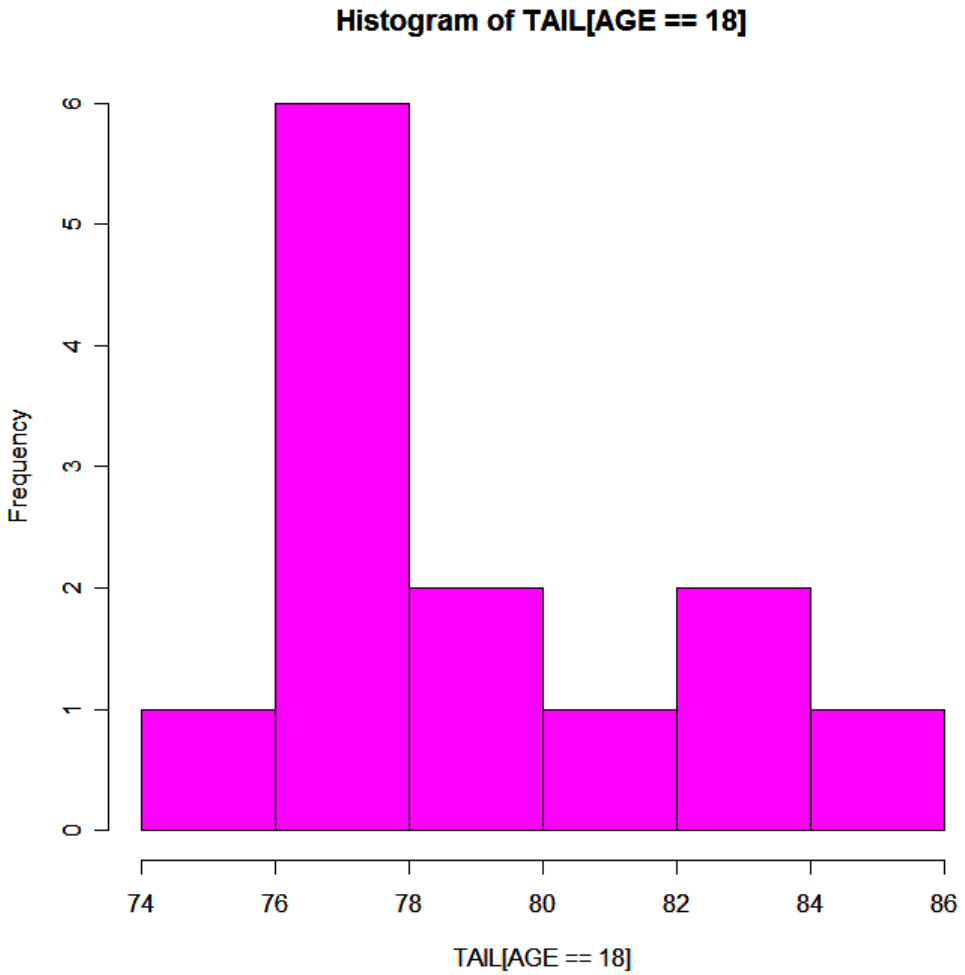
- 18 mois: quelle Taille?

  - Moyenne observée:

M(*T/A=18*)=79,23 cm

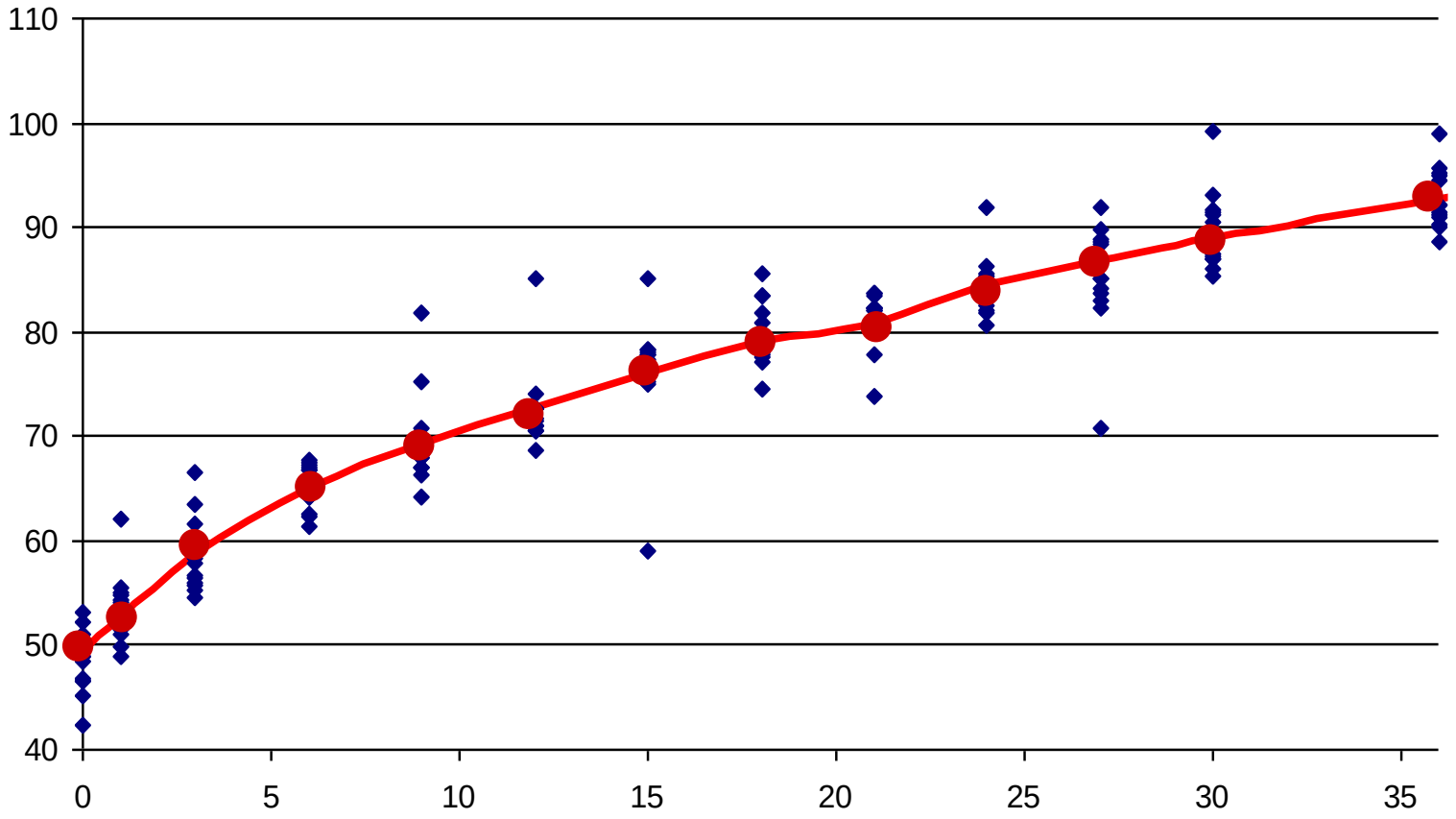
  - Variance observée:

V(*T/A=18*)=9,36 cm²



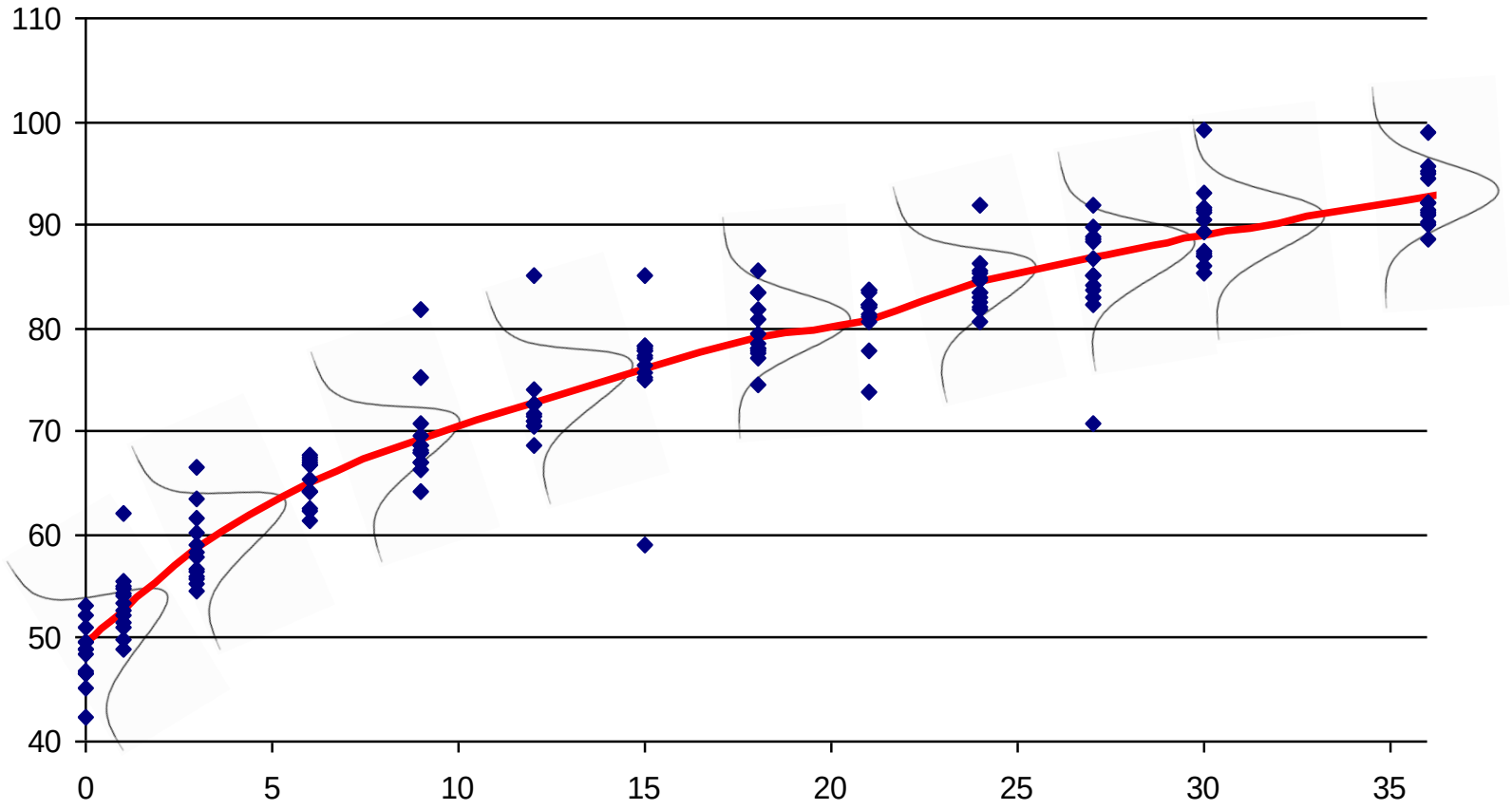
Distribution conditionnelle

$$E(Taille / Age)$$



Distribution conditionnelle

$$L(Taille / Age)$$



Fonction de régression

- Taille fonction de l'âge:

M(Taille/Age)= f(Age)

- Fonction ***f()***: droite

$$E(Taille / Age) = \alpha + \beta \times Age$$

Fonction de régression

- Taille fonction de l'âge:

M(Taille/Age)= f(Age)

- Fonction ***f()***: droite

$$E(Taille / Age) = \alpha + \beta \times Age$$

- Pour chaque sujet

$$Taille = \alpha + \beta \times Age + \varepsilon$$

Fonction de régression

- Taille fonction de l'âge:

E(Taille/Age)= f(Age)

- Fonction ***f()***: droite

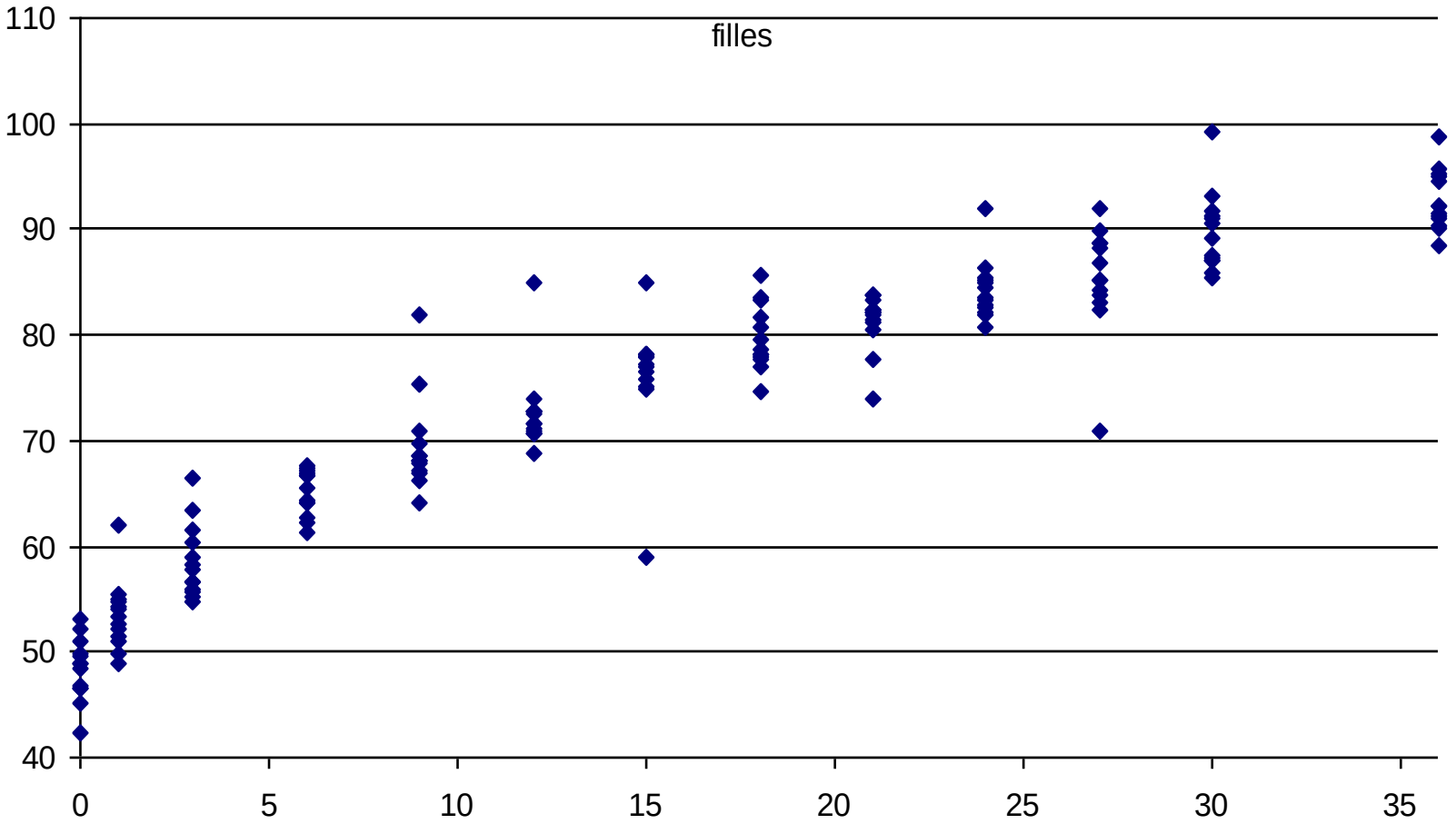
$$E(Taille / Age) = \alpha + \beta \times Age$$

- Pour chaque sujet

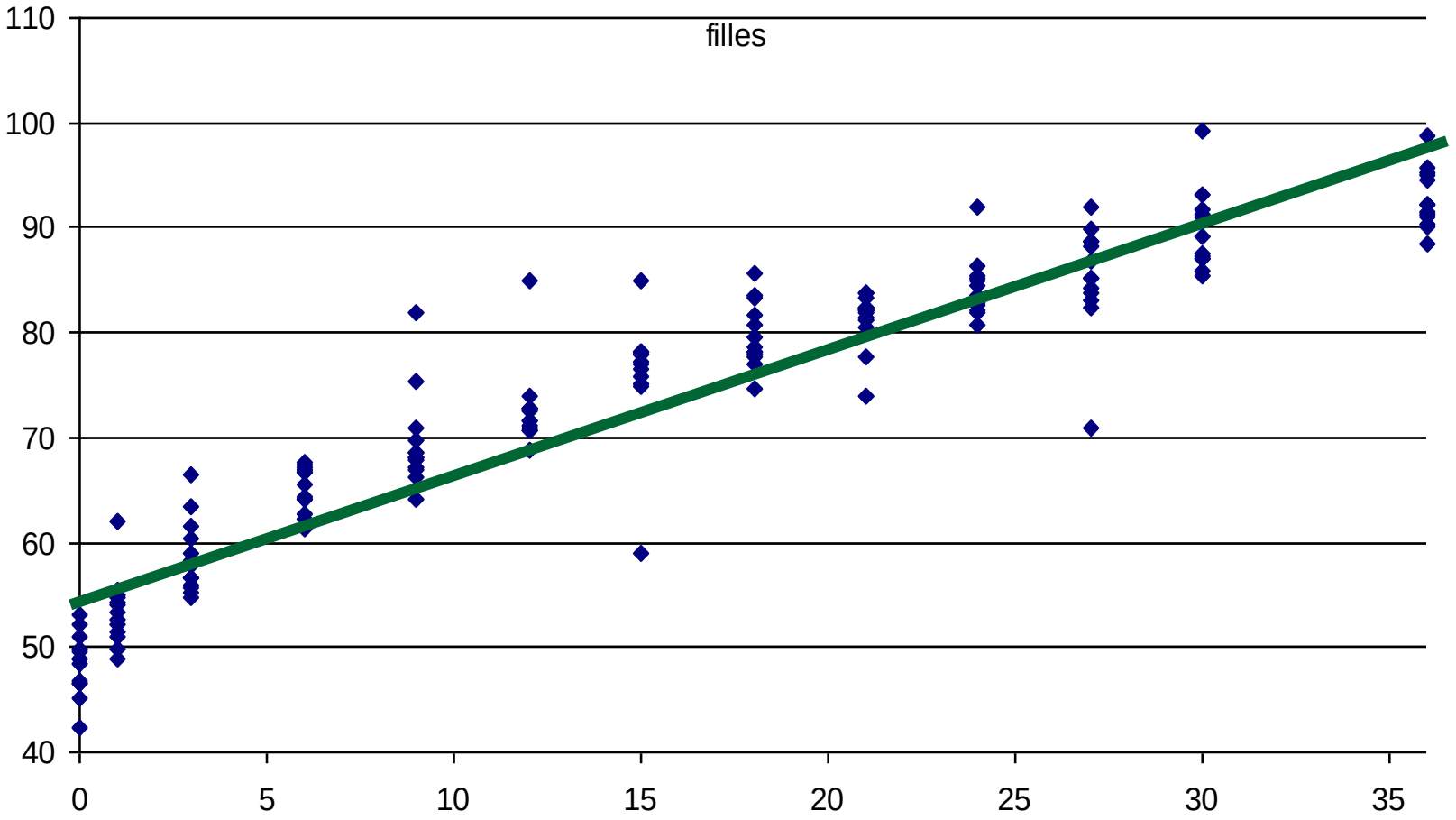
$$Taille = \alpha + \beta \times Age + \varepsilon$$

erreur individuelle

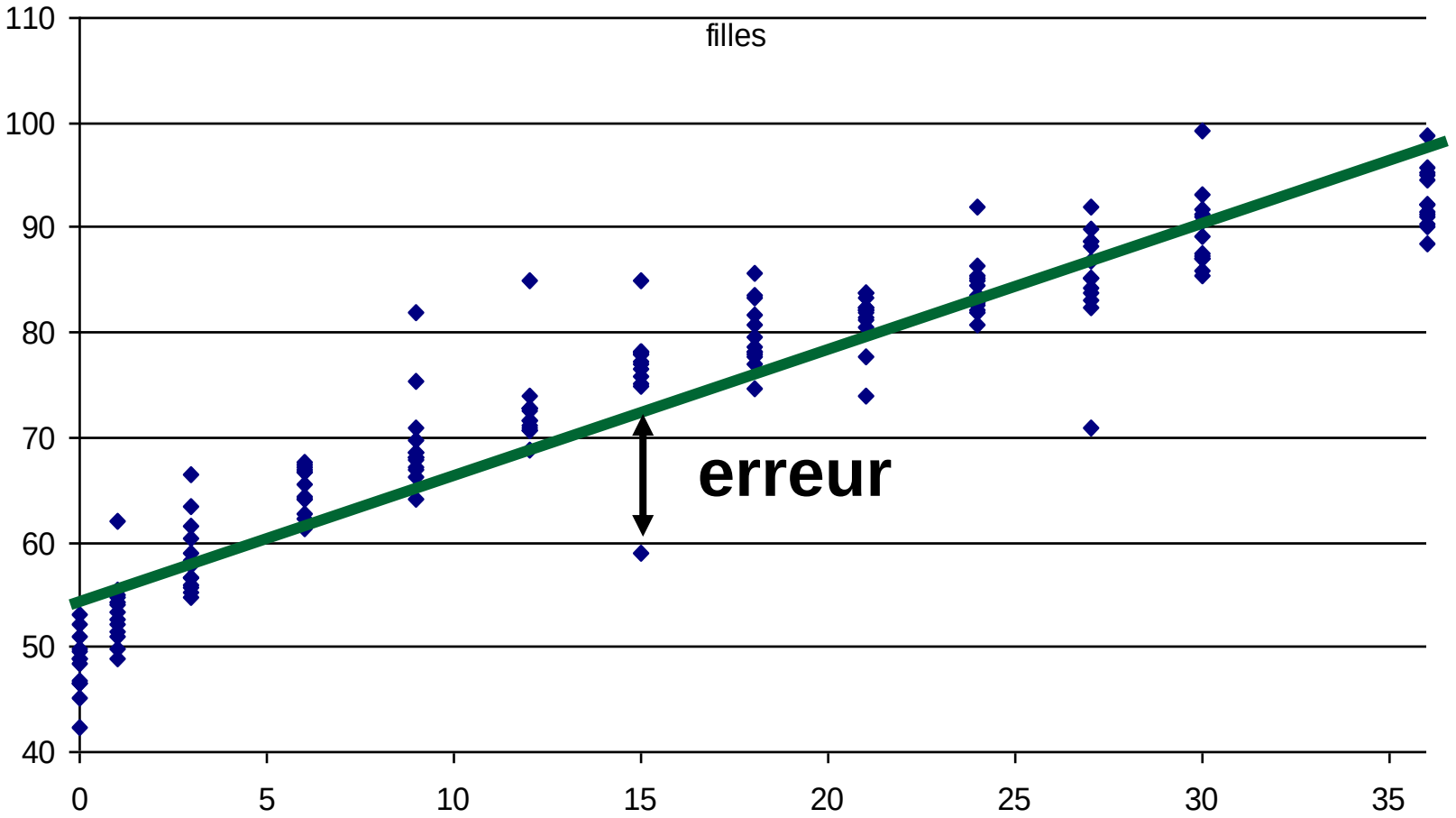
Erreur individuelle



Erreur individuelle

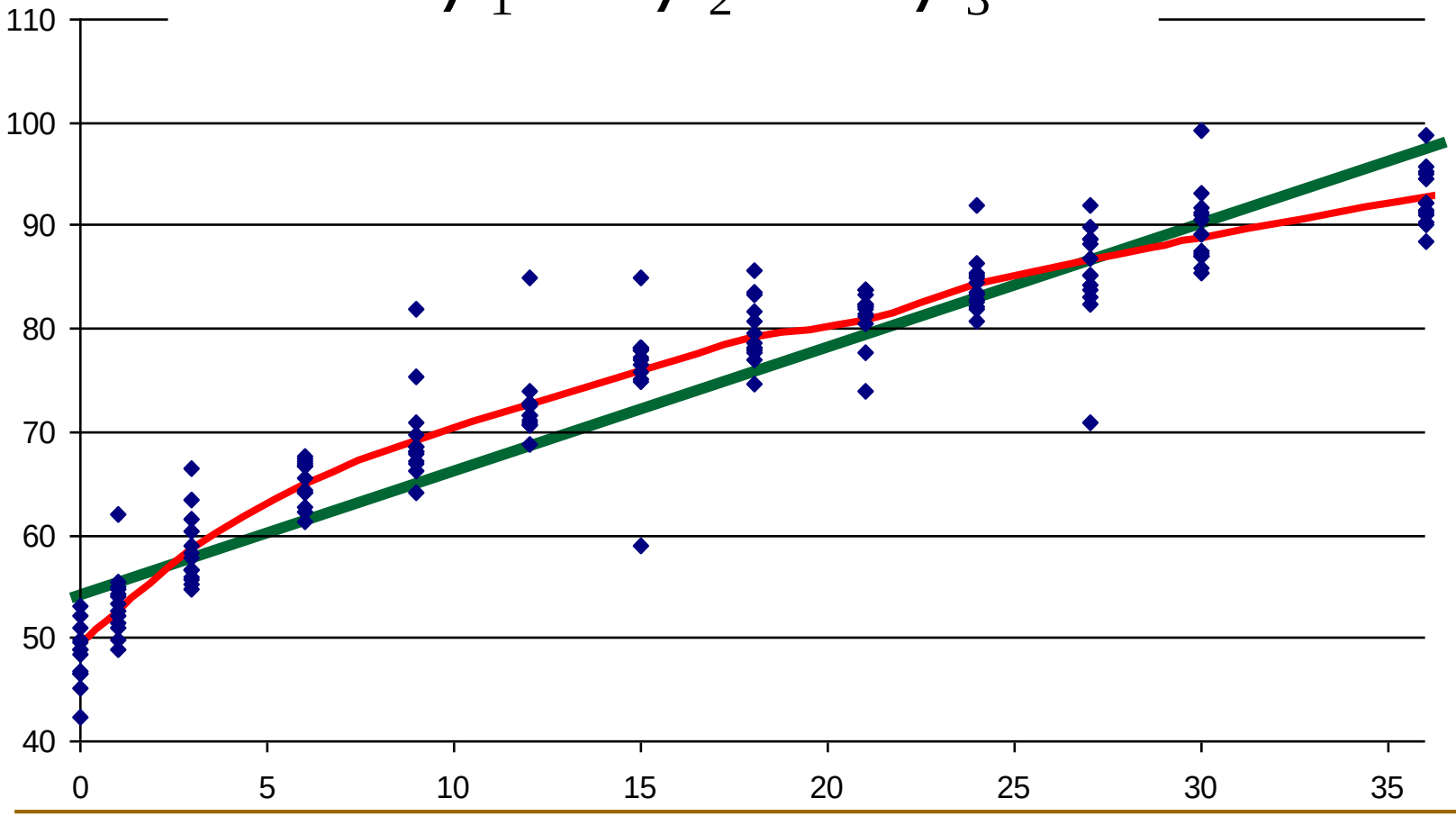


Erreur individuelle



Pourquoi **Linéaire** et pas un **Polynôme**?

$$Y = \alpha + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 \ldots$$



- Régression linéaire:

  modèle le plus **simple**:

  - Interprétation
  - Estimations des paramètres
  - Prédiction

III. Estimation

- Droite de régression:
  - Résume le **mieux** le nuage de point

  => La plus proche de tous les points

  => Erreurs ε petits +++

Principe de l'estimation

- Estimer α et β tel que **ε** petits +++

- **ε_i**: écart entre la droite et le point i

$$y_i = \alpha + \beta \times x_i + \varepsilon_i$$

$$E(Y / X) = \alpha + \beta \times X$$

Principe de l'estimation

- Estimer α et β tel que **ε** petits +++

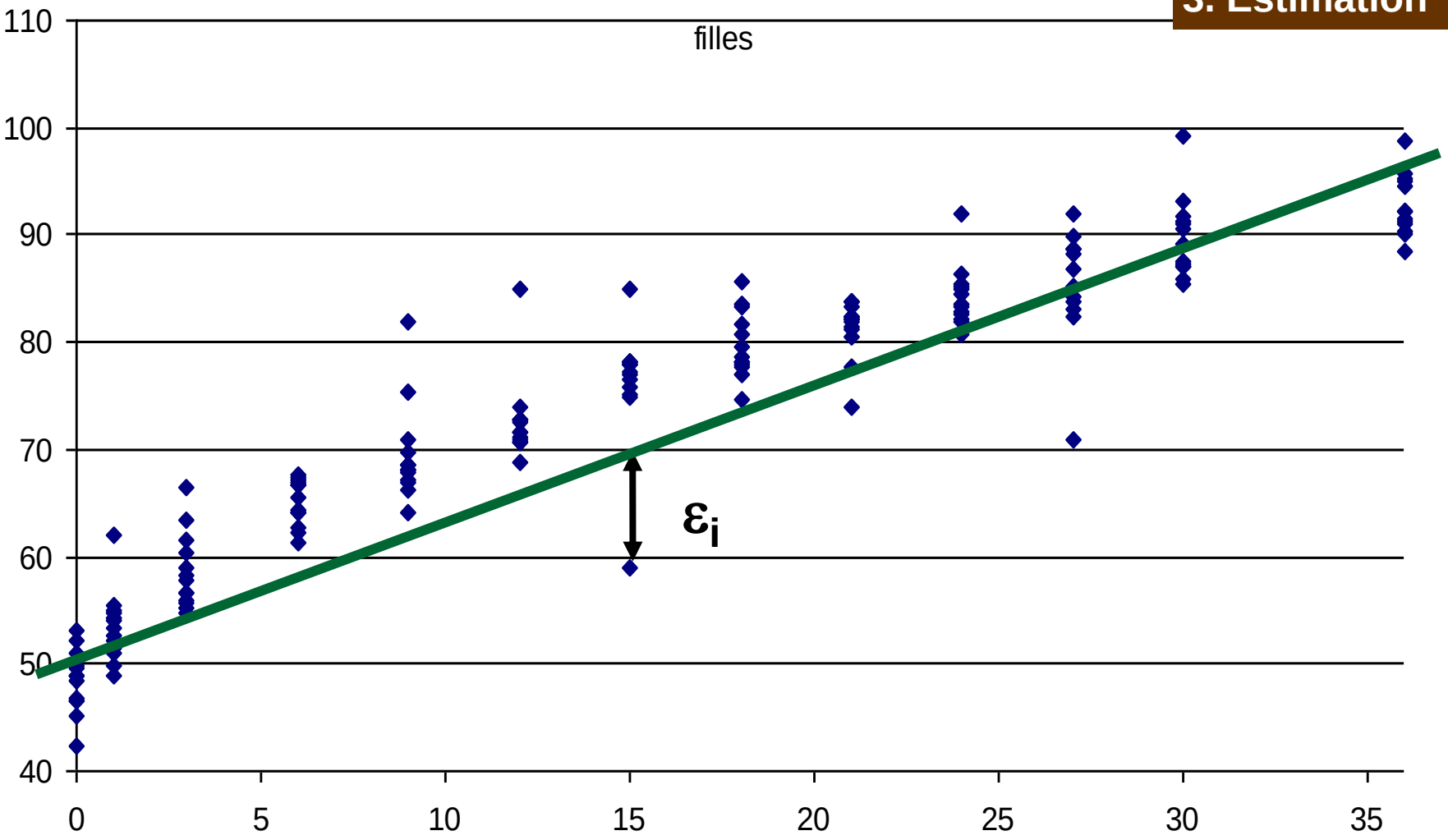
- **ε_i**: écart entre la droite et le point i

$$y_i = \alpha + \beta \times x_i + \varepsilon_i$$

$$E(Y / X) = \alpha + \beta \times X$$

$$\Rightarrow \varepsilon_i = y_i - E(Y / X)$$

Erreur individuelle



Principe de l'estimation

- Somme des Carrés des Ecarts

$$SCE = \sum_{i=1}^{n} (\varepsilon_i)^2$$

- Estimer α et β tel que:

SCE minimum

- Estimation de le pente β

$$b = \frac{cov(XY)}{var(X)}$$

- Estimation de le pente β

$$b = \frac{cov(XY)}{var(X)}$$

- Estimation de le pente β

$$b = \frac{cov(XY)}{var(X)}$$

Rappels

- Estimation de Variance de X:

$$S^2(X) = \frac{\sum_{i=1}^{n}(x_i^2) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}$$

- Estimation de la covariance de XY

$$\widehat{cov(XY)} = \frac{\sum_{i=1}^{n}(x_i y_i) - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n-1}$$

Exercice

- Covariance de la taille et de l'âge:

cov(TAIL,AGE)

- Variance de l'âge

var(AGE)

- Estimation de β

b<-cov(TAIL,AGE)/var(AGE)

b

Exercice

- Covariance de la taille et de l'âge:

cov(TAIL,AGE) = 2742.587

- Variance de l'âge

var(AGE)

- Estimation de β

b<-cov(TAIL,AGE)/var(AGE)

b =0.437703

- Estimation de α :
  - La droite passe par m_Y et m_X

$$m_Y = a + b m_X$$

- Estimation de α :
  - La droite passe par m_Y et m_X

$$m_Y = a + b m_X$$

$$a = m_Y - b m_X$$

Exercice

- Estimation de α :

*a<-mean(TAIL)-b*mean(AGE)*

a =73.729

Exercice

- Estimation de α :

*a<-mean(TAIL)-b*mean(AGE)*

a =73.729

- l'équation s'écrit donc:

Exercice

- Estimation de α :

*a<-mean(TAIL)-b*mean(AGE)*

a =73.729

- l'équation s'écrit donc:

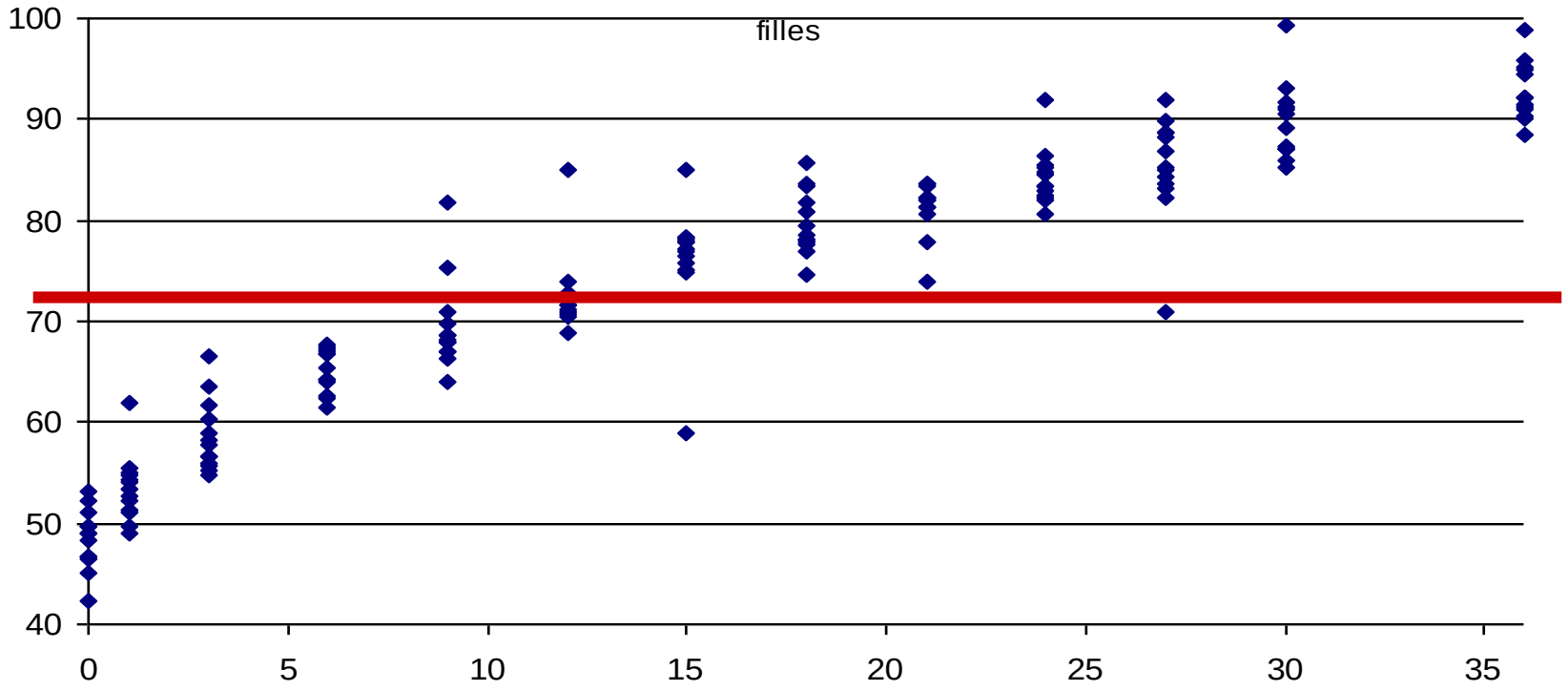
Taille=73.73 +0.44 Age + ε

- ou

E(Taille/Age)=73.73 +0.44 Age

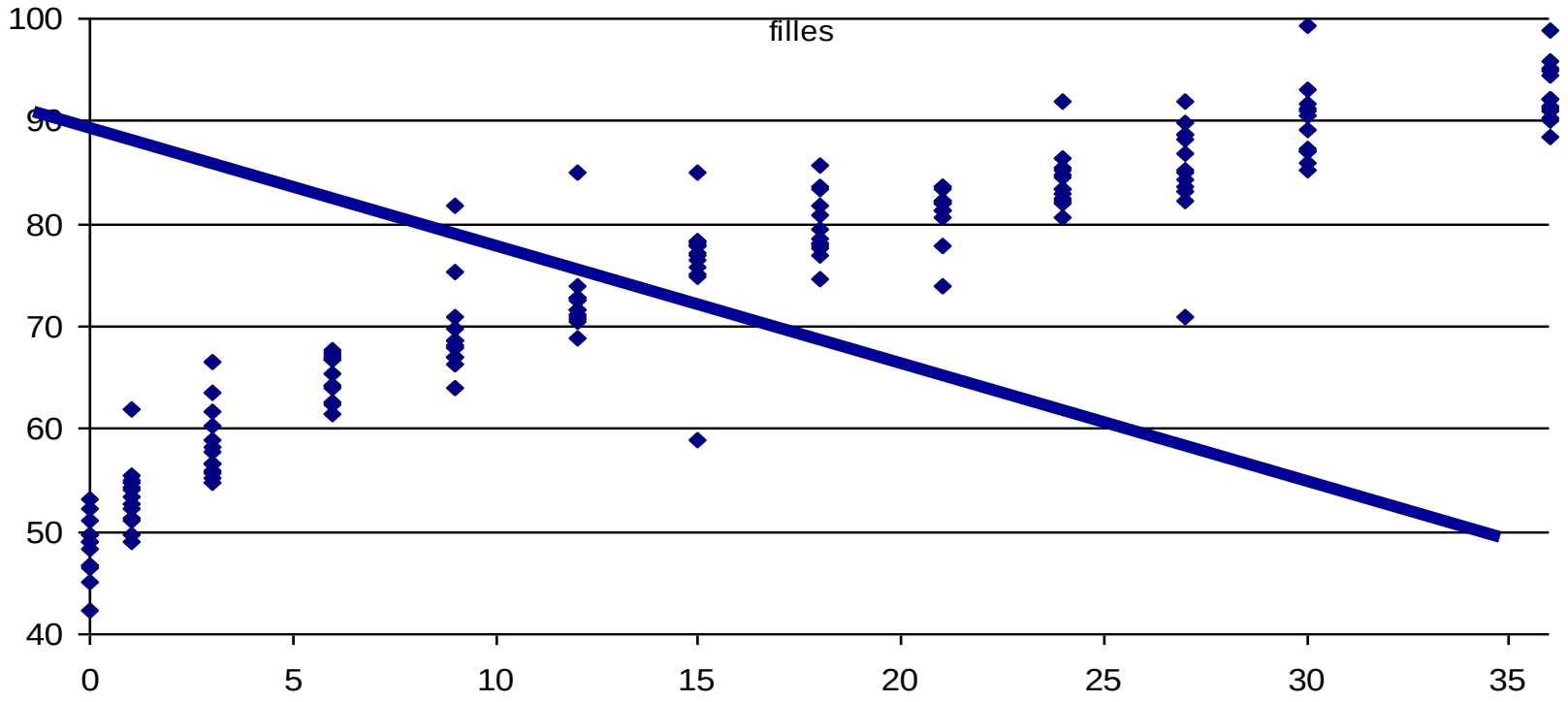
Interprétation

- Pente β:
  - **β=0**: pas de lien, évolutions indépendantes



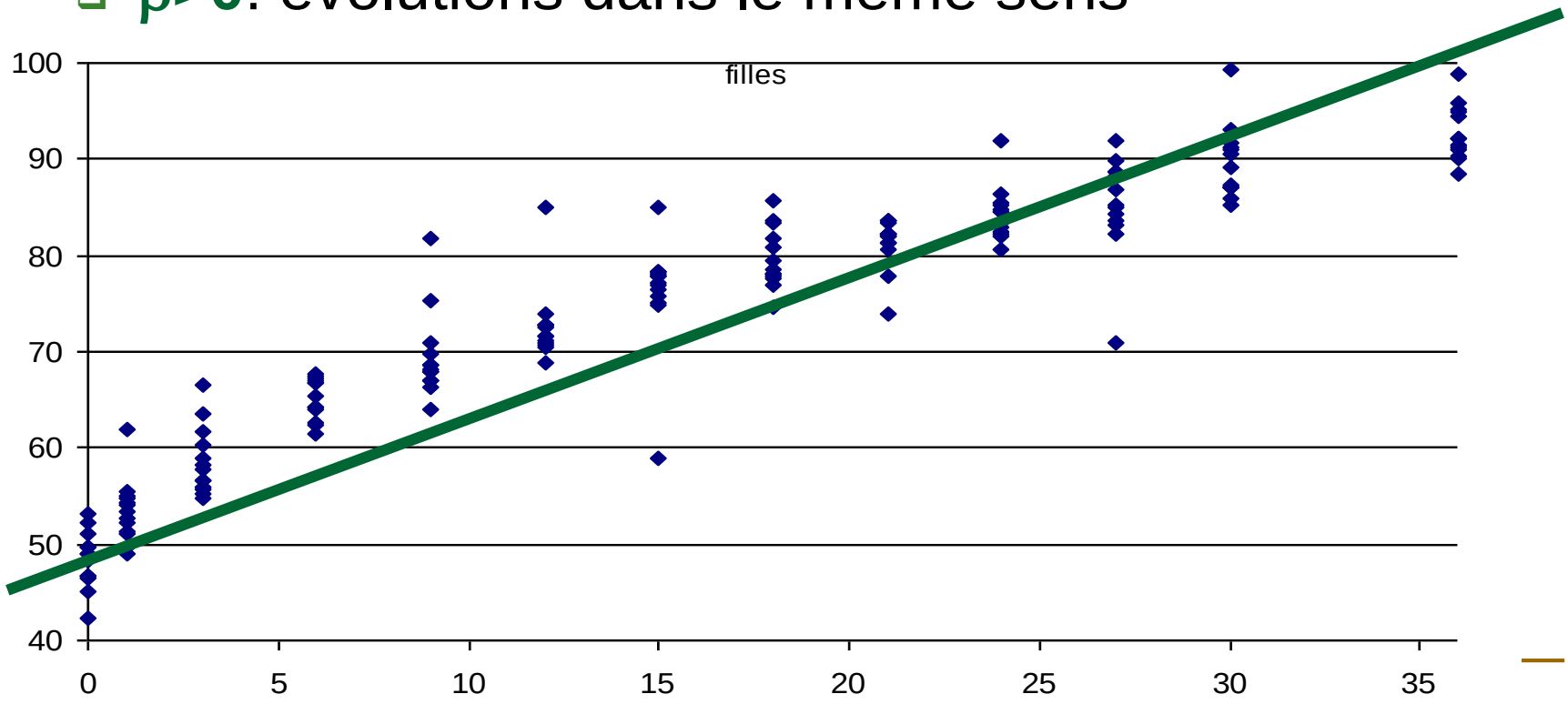
Interprétation

- Pente β:
  - **β=0**: pas de lien, évolutions indépendantes
  - **β<0**: évolutions en sens contraire

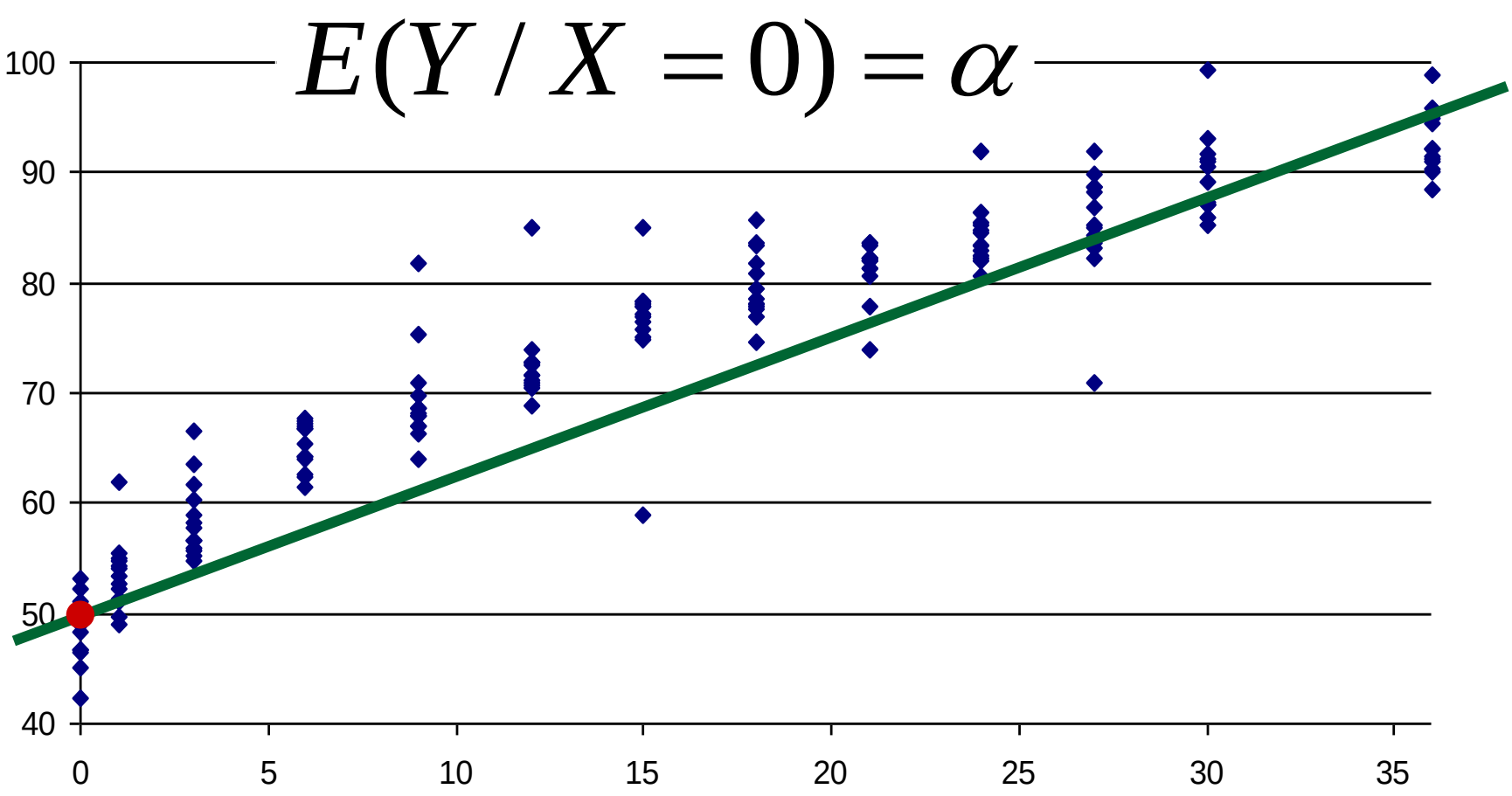


Interprétation

- Pente β:
  - **β=0**: pas de lien, évolutions indépendantes
  - **β<0**: évolutions en sens contraire
  - **β>0**: évolutions dans le même sens



- Ordonnée à l'origine α



IV. Test de la pente

- Si $\beta=0$ => pas de lien entre Y et X
- Lien entre Y et X est-il significatif?

  => $\beta \neq 0$?
- ***b*** estimation de β

Hasard=> fluctuation de ***b*** observé

=> Test statistique

- Hypothèses:

H0: $\beta = 0$, il n'y a pas de lien entre X et Y

H1: $\beta \neq 0$, il y a un lien entre X et Y

- Sous H0

$$t_0 = \frac{b - \beta}{\sqrt{{s_b}^2}}$$ ~> Student à n-2 ddl

Avec

$${s_b}^2 = \frac{\frac{{s_Y}^2}{{s_X}^2} - b^2}{n-2}$$

Exercice

- Modèle linéaire: utilisation du logiciel R

=> fonction ***lm*** linear model

```
mod1<-lm(TAIL~1+AGE)
mod1
```

Exercice

- Modèle linéaire: utilisation du logiciel R

=> fonction ***lm*** linear model

mod1<-lm(TAIL ~ AGE)
mod1

```
Call:  lm(formula = TAIL ~  AGE)

Coefficients:
(Intercept)          AGE
    73.7290       0.4377
```

Exercice

- Modèle linéaire: utilisation du logiciel R

=> fonction ***lm*** linear model

mod1<-lm(TAIL~ AGE)
mod1

```
Call:  lm(formula = TAIL ~  AGE)

Coefficients:
(Intercept)         AGE
    73.7290       0.4377
```

Exercice

- Modèle linéaire: utilisation du logiciel R

=> fonction ***lm*** linear model

mod1<-lm(TAIL~ AGE)
mod1

```
Call:  lm(formula = TAIL ~  AGE)

Coefficients:
(Intercept)          AGE
    73.7290       0.4377
```

a **b**

Exercice

summary(mod1)

Exercice

summary(mod1)

```
Call: lm(formula = TAIL ~ AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-40.030  -6.899   2.999   8.120  24.999

Coefficients:
             Estimate        Std. Error      t        value Pr(>|t|)
(Intercept) 73.729005   0.744041   99.09   <2e-16 ***
AGE          0.437703   0.005423   80.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 635 degrees of freedom
Multiple R-squared: 0.9112,     Adjusted R-squared: 0.9111
F-statistic:  6516 on 1 and 635 DF,  p-value: < 2.2e-16
```

Exercice

summary(mod1)

```
Call: lm(formula = TAIL ~ AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-40.030  -6.899   2.999   8.120  24.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.729005   0.744041   99.09   <2e-16 ***
AGE          0.437703   0.005423   80.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 635 degrees of freedom
Multiple R-squared: 0.9112,     Adjusted R-squared: 0.9111
F-statistic:  6516 on 1 and 635 DF,  p-value: < 2.2e-16
```

a

b

Exercice

summary(mod1)

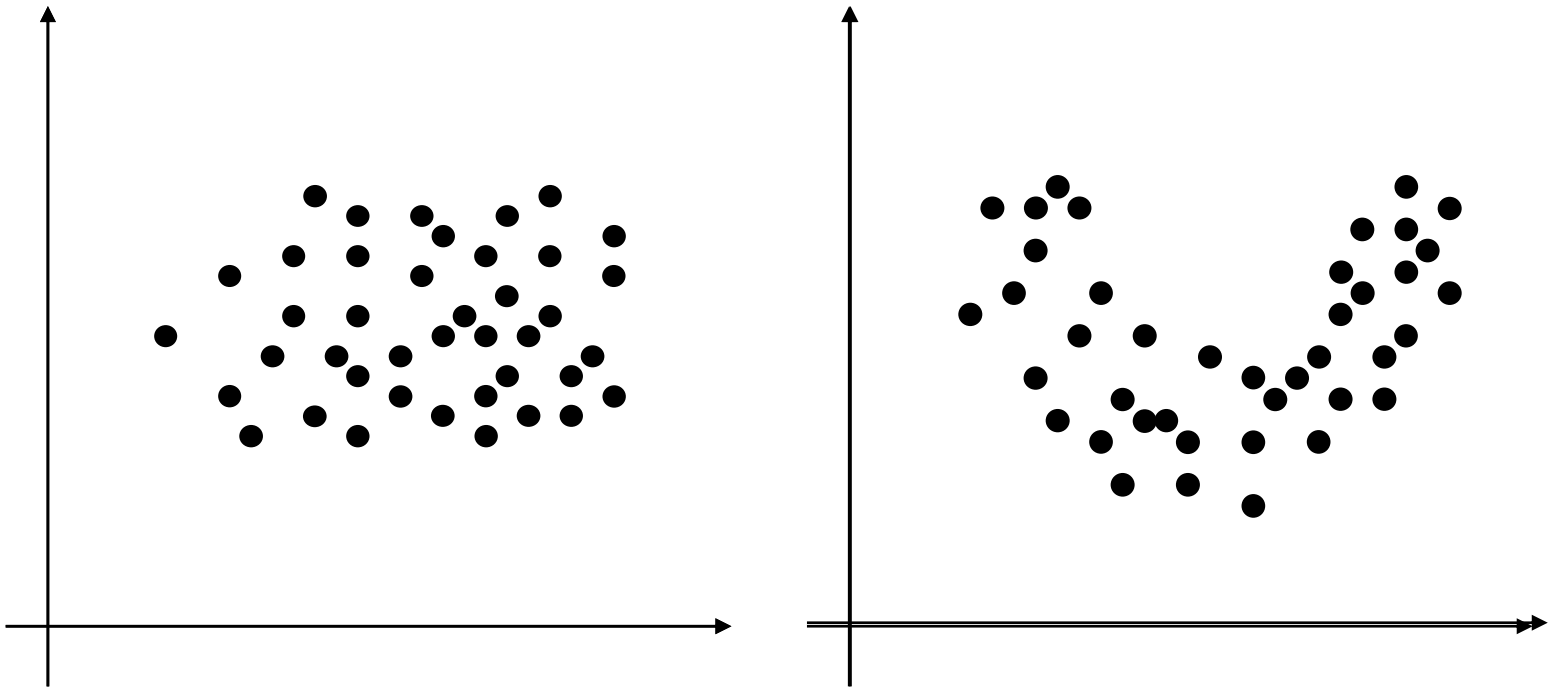
```
Call: lm(formula = TAIL ~ AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-40.030  -6.899   2.999   8.120  24.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.729005   0.744041   99.09   <2e-16 ***
AGE          0.437703   0.005423   80.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 635 degrees of freedom
Multiple R-squared: 0.9112,     Adjusted R-squared: 0.9111
F-statistic:  6516 on 1 and 635 DF,  p-value: < 2.2e-16
```

test $\beta=0$

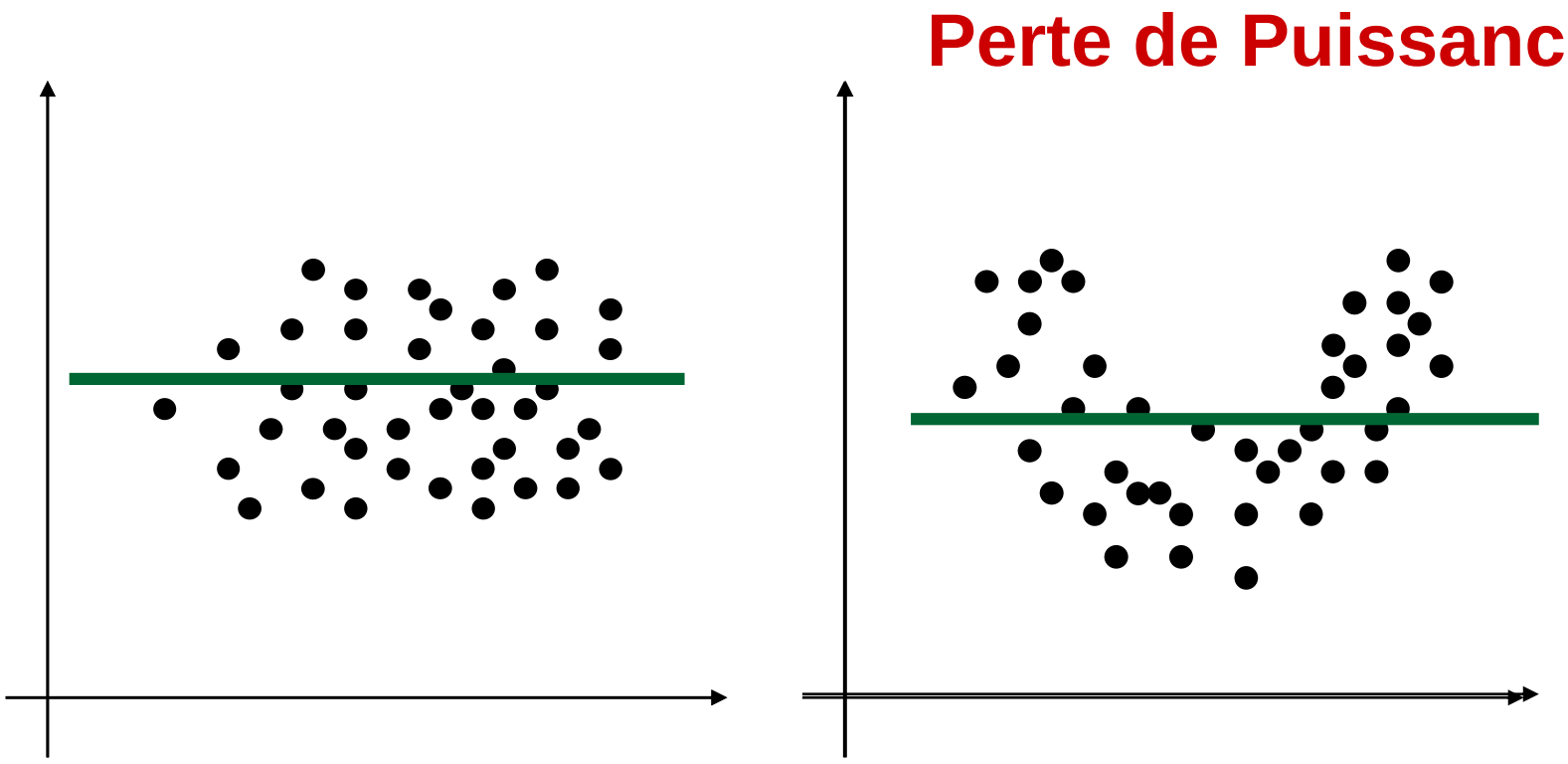
Conditions d'applications

- ***L(Y/X)~>$\mathcal{N}$***
- ***V(Y/X)*** constantes pour tout X
- à X donné, Y_i indépendants
- La régression est linéaire

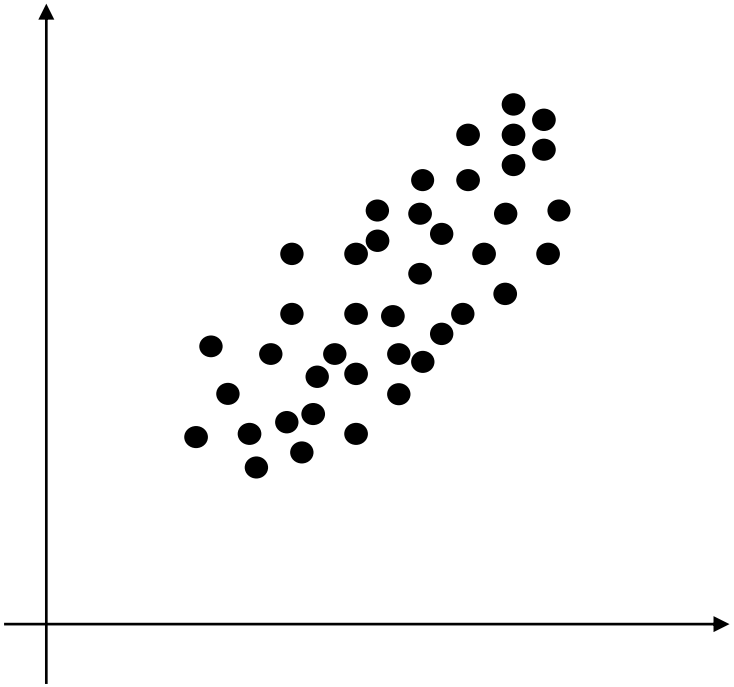
Linéarité

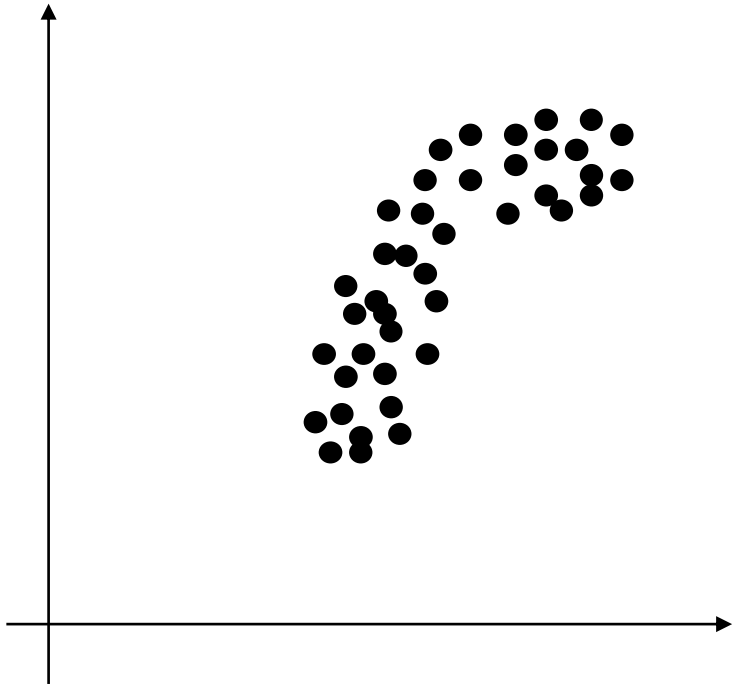


- Linéarité



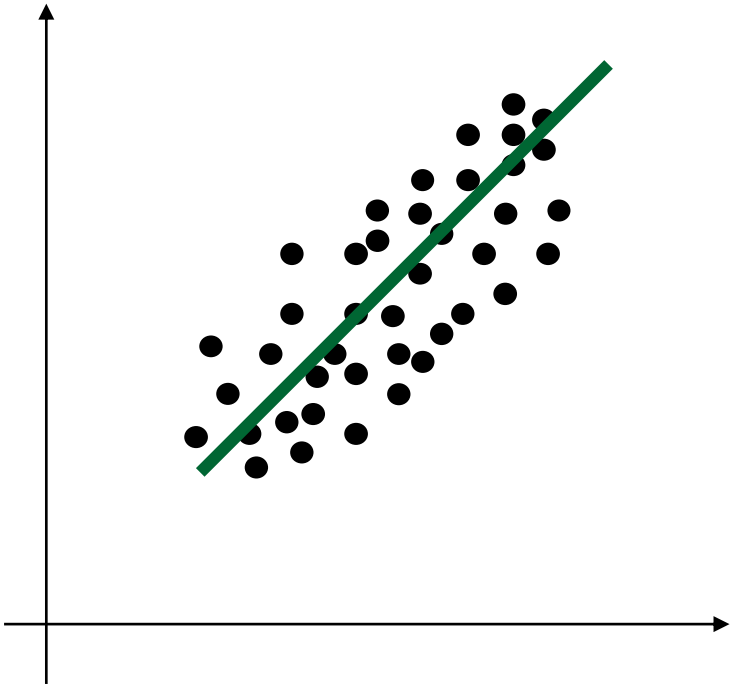
- Linéarité

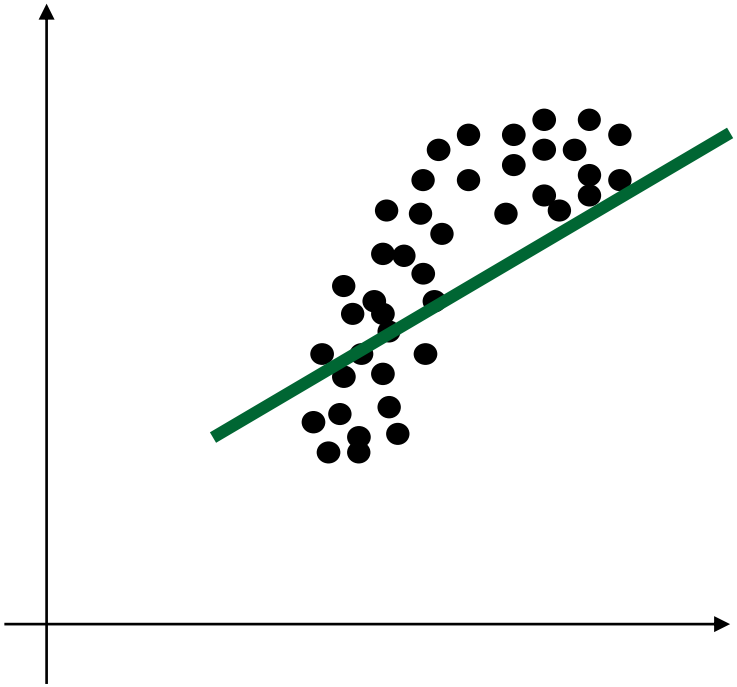




- Linéarité

Perte de Puissance





- Conditions d'applications
  - ***L(Y/X)*~>$\mathcal{N}$**
  - ***V(Y/X)*** constantes pour tout X
    homoscédasticité
  - à X donné, Y_i indépendants
  - La régression est linéaire

- Conditions d'applications
  - ***L(Y/X)*~>$\mathcal{N}$** → qqnorm
  - ***V(Y/X)*** constantes pour tout X
    homoscédasticité
  - à X donné, Y_i indépendants
  - La régression est linéaire

- Conditions d'applications
  - ***L(Y/X)*~>$\mathcal{N}$** → qqnorm
  - ***V(Y/X)*** constantes pour tout X → plot(Tailles estimée, Résidus)
    homoscédasticité
  - à X donné, Y_i indépendants
  - La régression est linéaire

- Conditions d'applications
  - ***L(Y/X)*~>$\mathcal{N}$** → qqnorm
  - ***V(Y/X)*** constantes pour tout X → plot(Tailles estimée, Résidus)
    homoscédasticité
  - à X donné, Y_i indépendants → protocole
  - La régression est linéaire

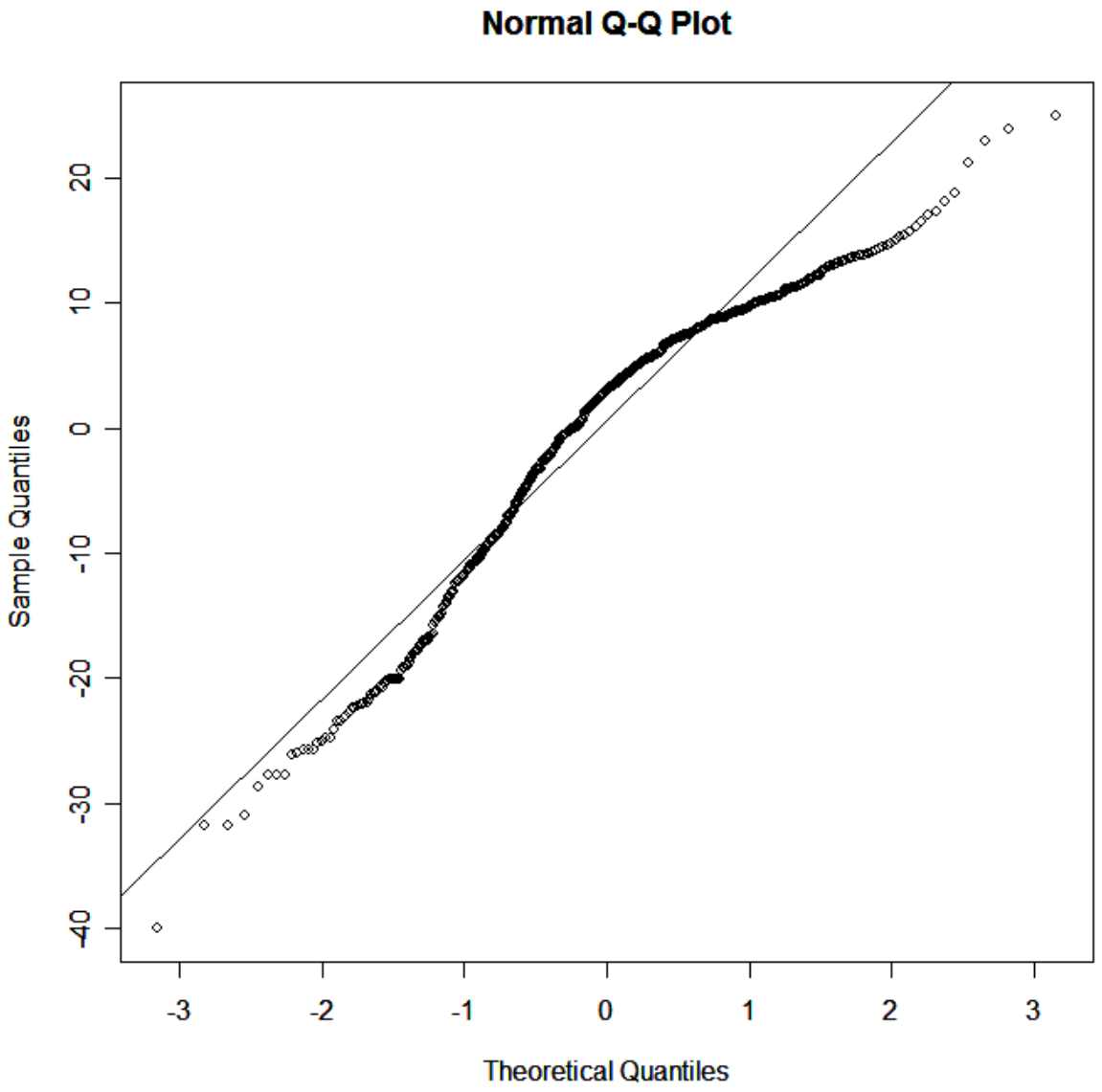
Conditions d'applications

- ***L(Y/X)*~>$\mathcal{N}$** → qqnorm
- ***V(Y/X)*** constantes pour tout X homoscédasticité → plot(Tailles estimée, Résidus)
- à X donné, Y_i indépendants → protocole
- La régression est linéaire → plot(Taille,Age)

Exercice

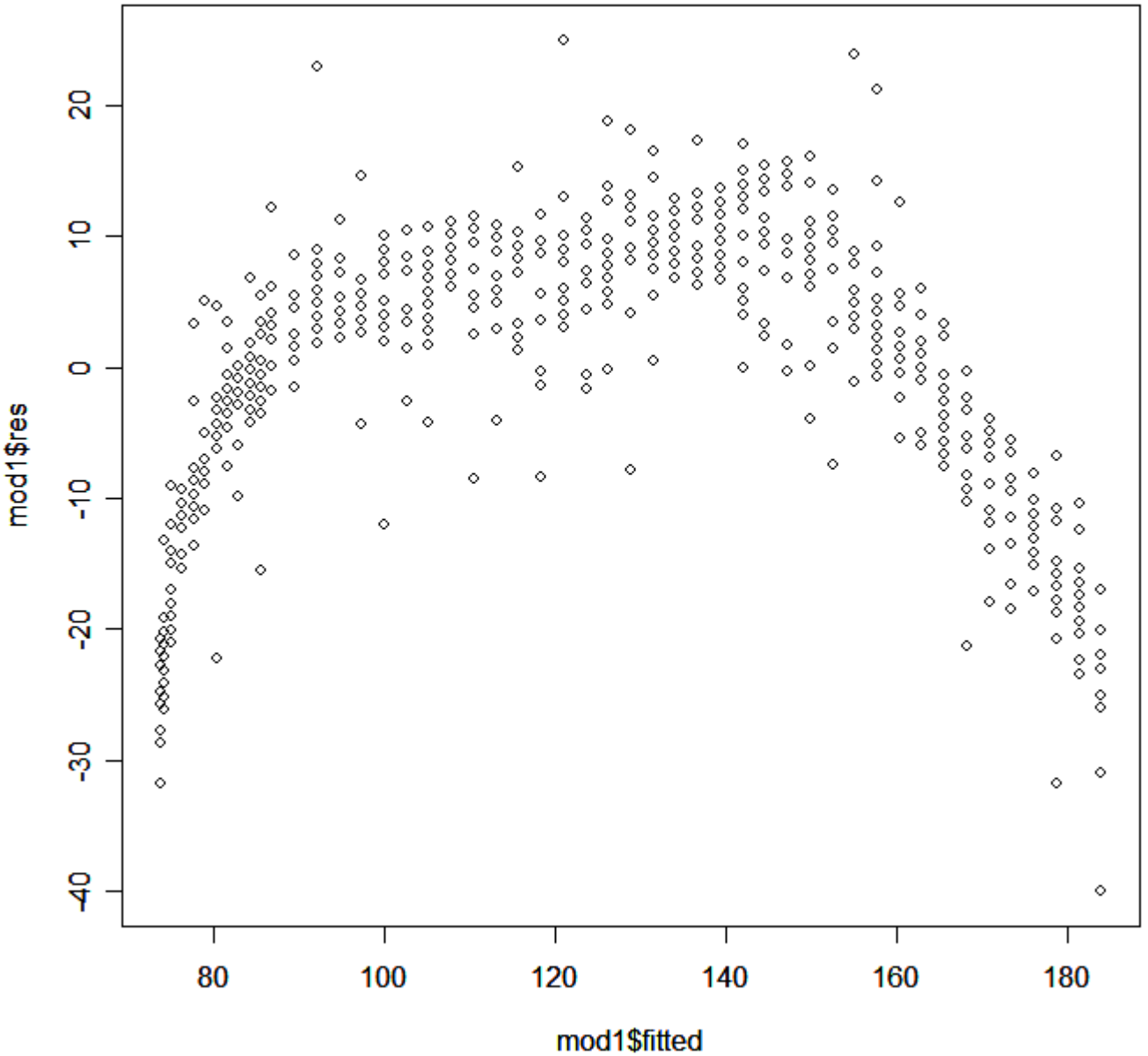
- ***L(Y/X)~>$\mathcal{N}$***

qqnorm(mod1$res)
qqline(mod1$res)

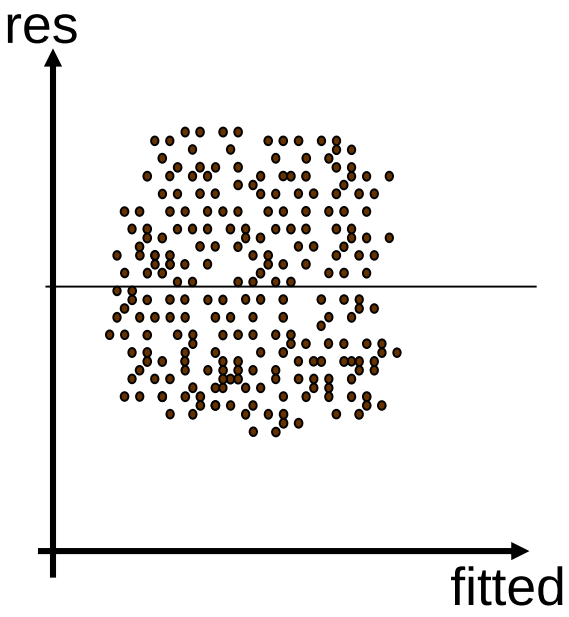


Exercice

plot(mod1$fitted,mod1$res)

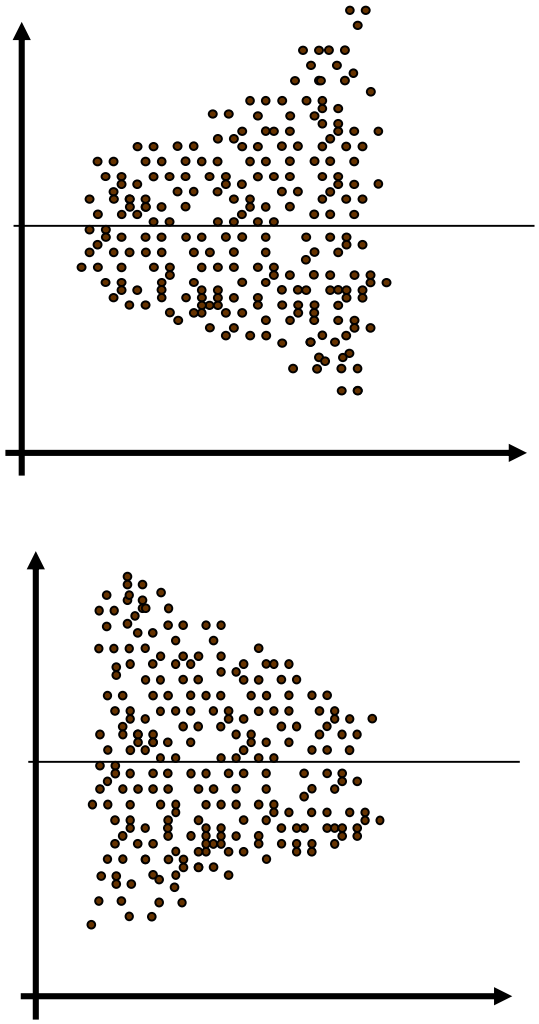


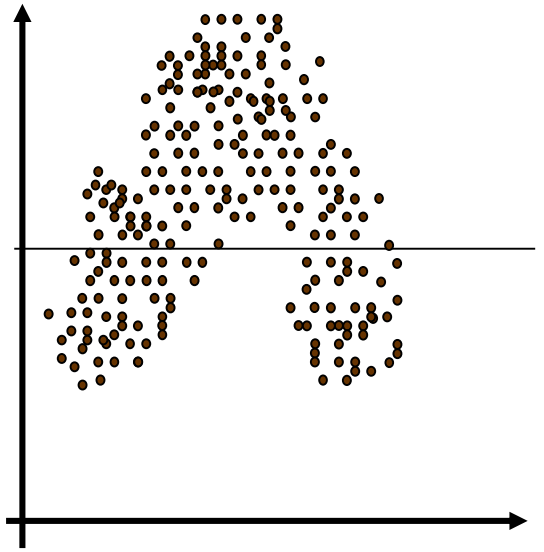
homoscédasticité



Var(Y/X)=cst

hétéroscédasticité

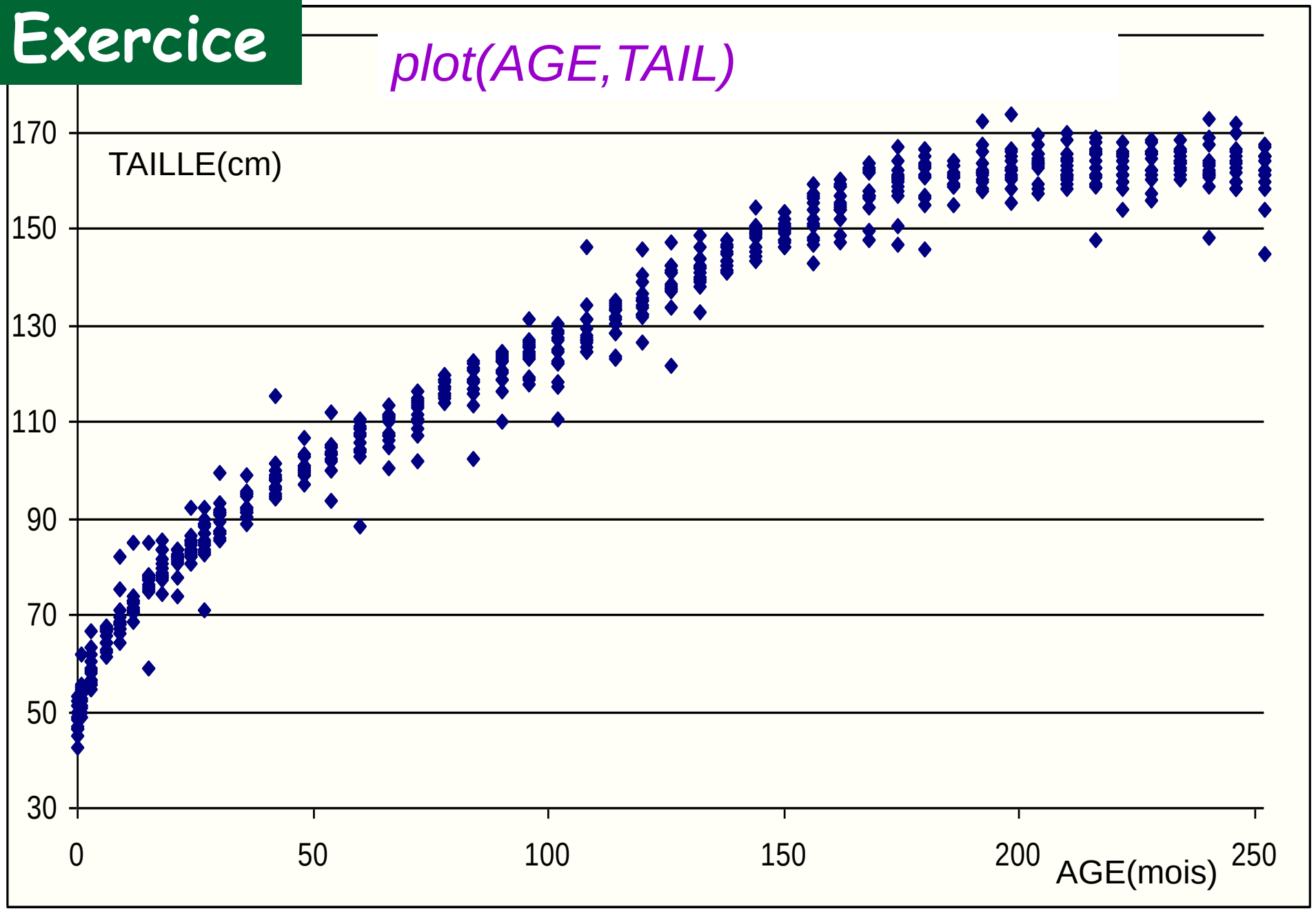




non-linéarité

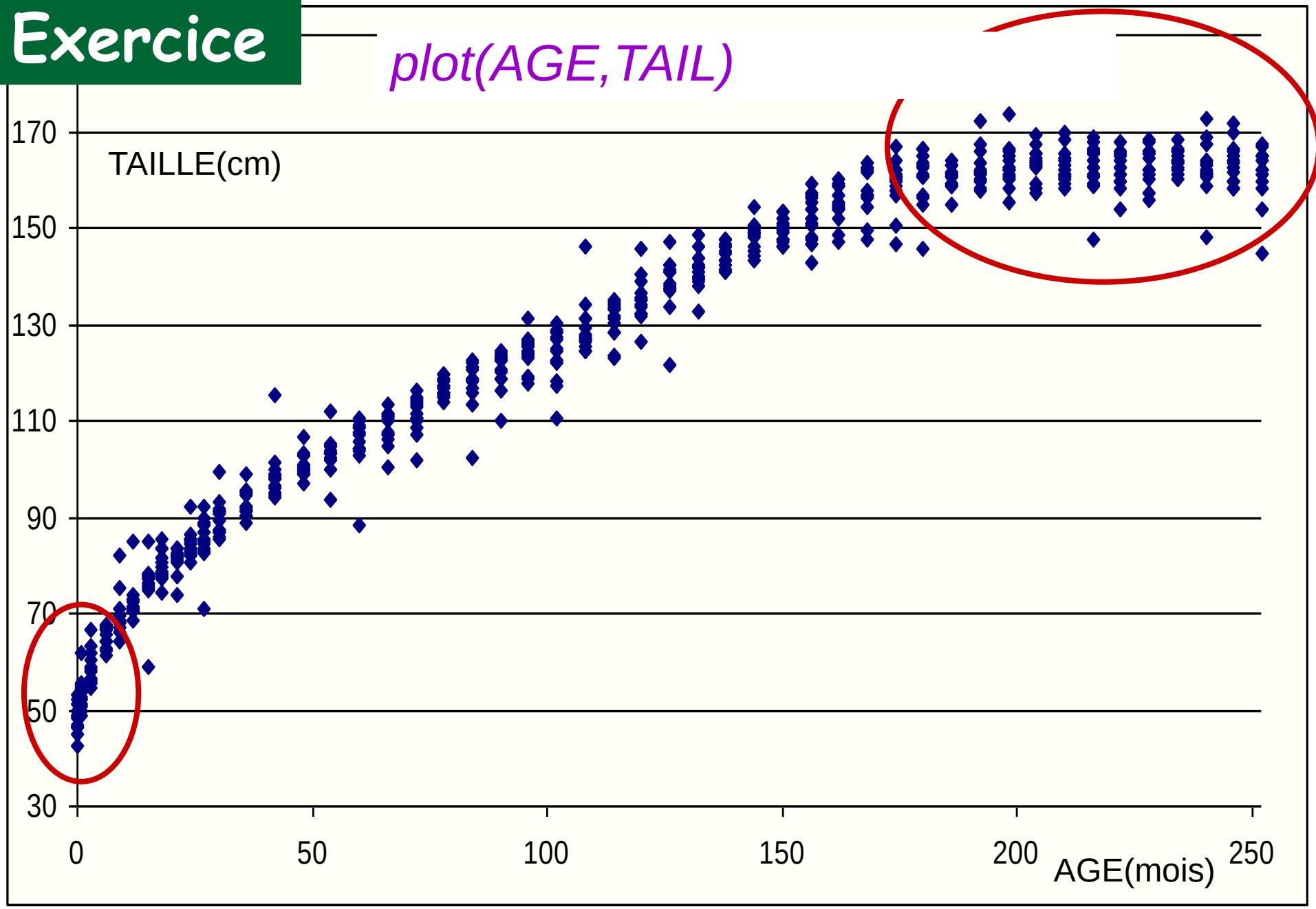
Exercice

plot(AGE,TAIL)

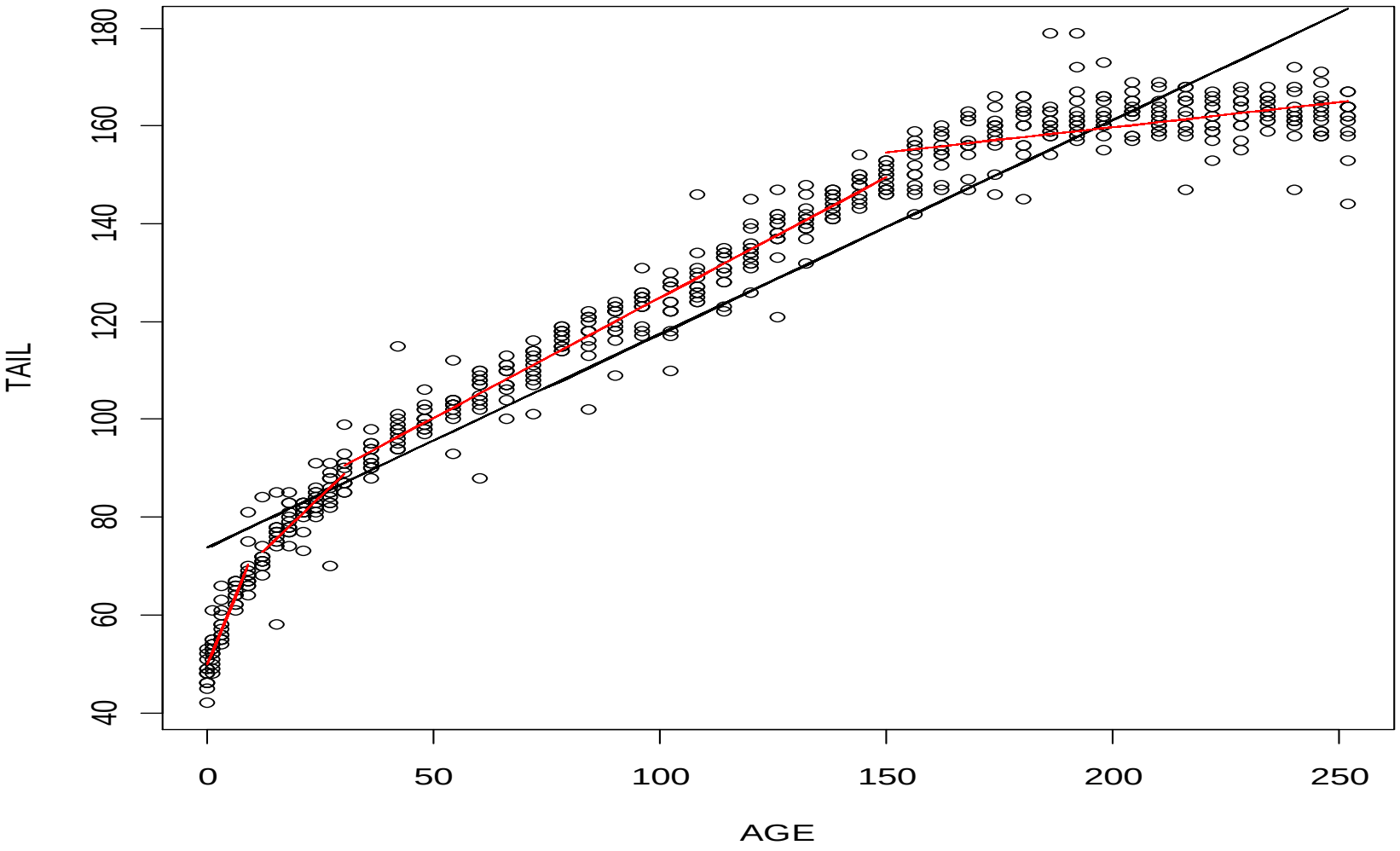


Exercice

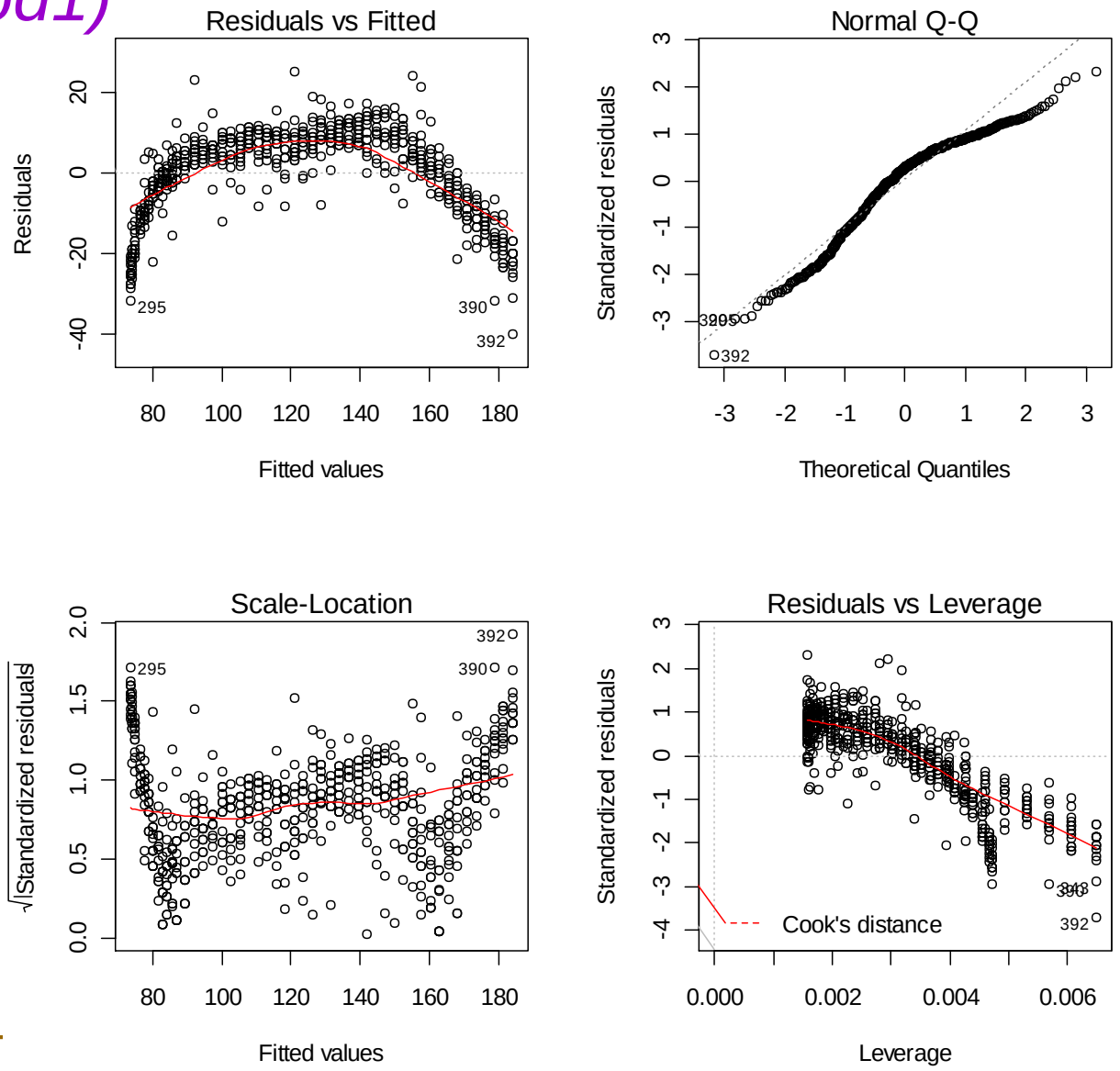
plot(AGE,TAIL)



Exercice



```
par(mfrow=c(2,2))
plot(mod1)
```



V. Précision

- Hasard=> fluctuation de ***b***
- Intervalle de confiance de la pente
  - b~>t_{n-2}

$$b \pm t_{n-2,\alpha/2} \times \sqrt{s_b^2}$$

V. Précision

- Hasard=> fluctuation de ***b***
- Intervalle de confiance de la pente
  - b~>t_{n-2}

$$b \pm t_{n-2,\alpha/2} \times \sqrt{s_b^2}$$

- Conditions d'applications
  - ***L(Y/X)***~>$\mathcal{N}$
  - ***V(Y/X)*** constantes pour tout X
  - à X donné, Y_i indépendants
  - La régression est linéaire

Exercice

- Intervalles de confiance des paramètres

confint(mod1)

Exercice

- Intervalles de confiance des paramètres

confint(mod1)

	2.5 %	97.5 %
(Intercept)	72.2707108	75.1872989
AGE	0.4270751	0.4483309

- Intervalle de confiance de la droite

$$E(Y / X) = \alpha + \beta X$$

Estimé par $m_{Y/X} = a + bX$

$$m_{Y/X} \pm t_{n-2,\alpha/2} \times \sqrt{s_{m_{Y/X}}^{2}}$$

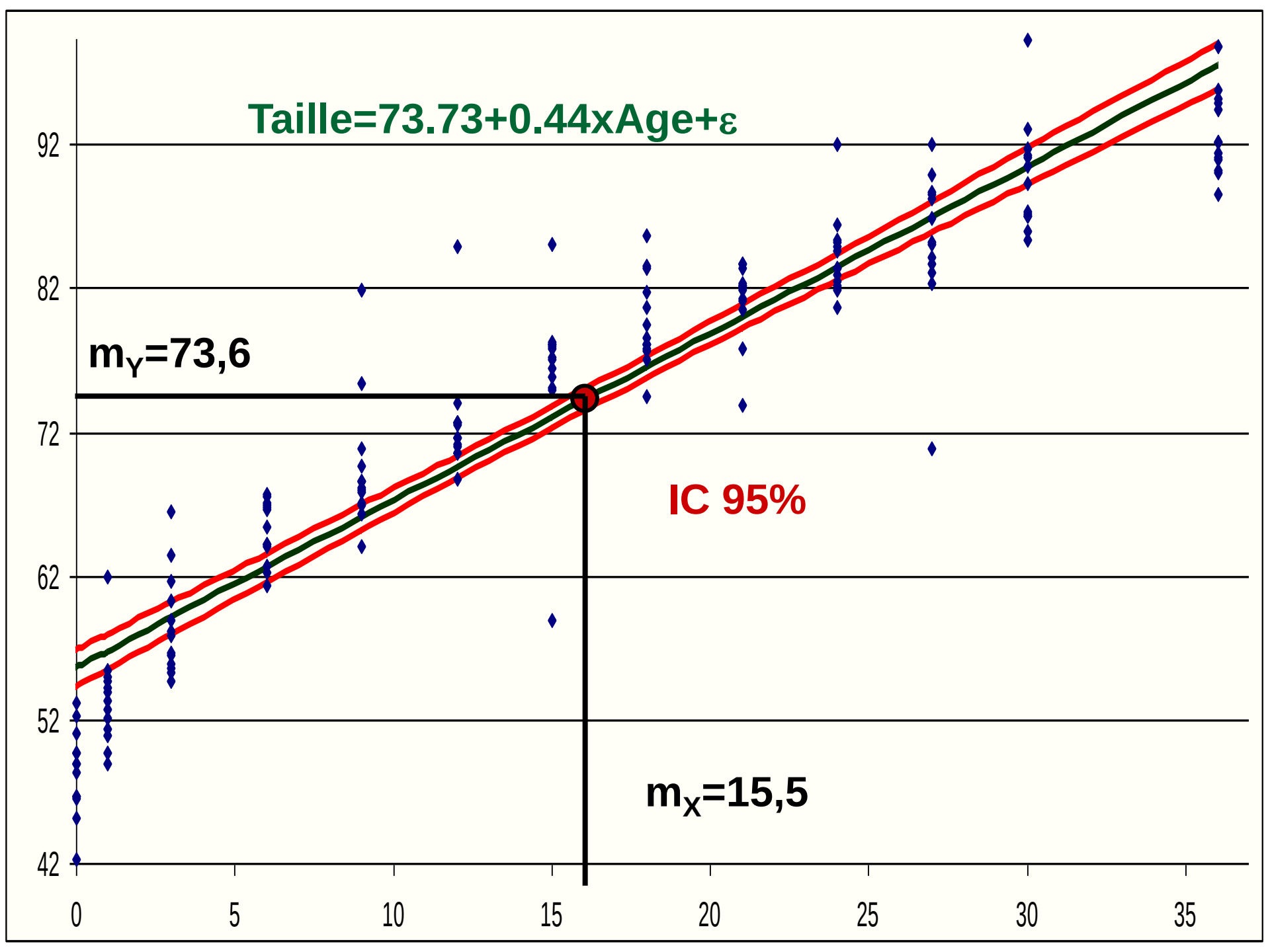
- Intervalle de confiance de la droite

$$E(Y/X) = \alpha + \beta X$$

Estimé par $m_{Y/X} = a + bX$

$$m_{Y/X} \pm t_{n-2,\alpha/2} \times \sqrt{s_{m_{Y/X}}^2}$$

- Conditions d'applications
  - ***L(Y/X)~>$\mathcal{N}$***
  - ***V(Y/X)*** constantes pour tout X
  - à X donné, Y_i indépendants
  - La régression est linéaire

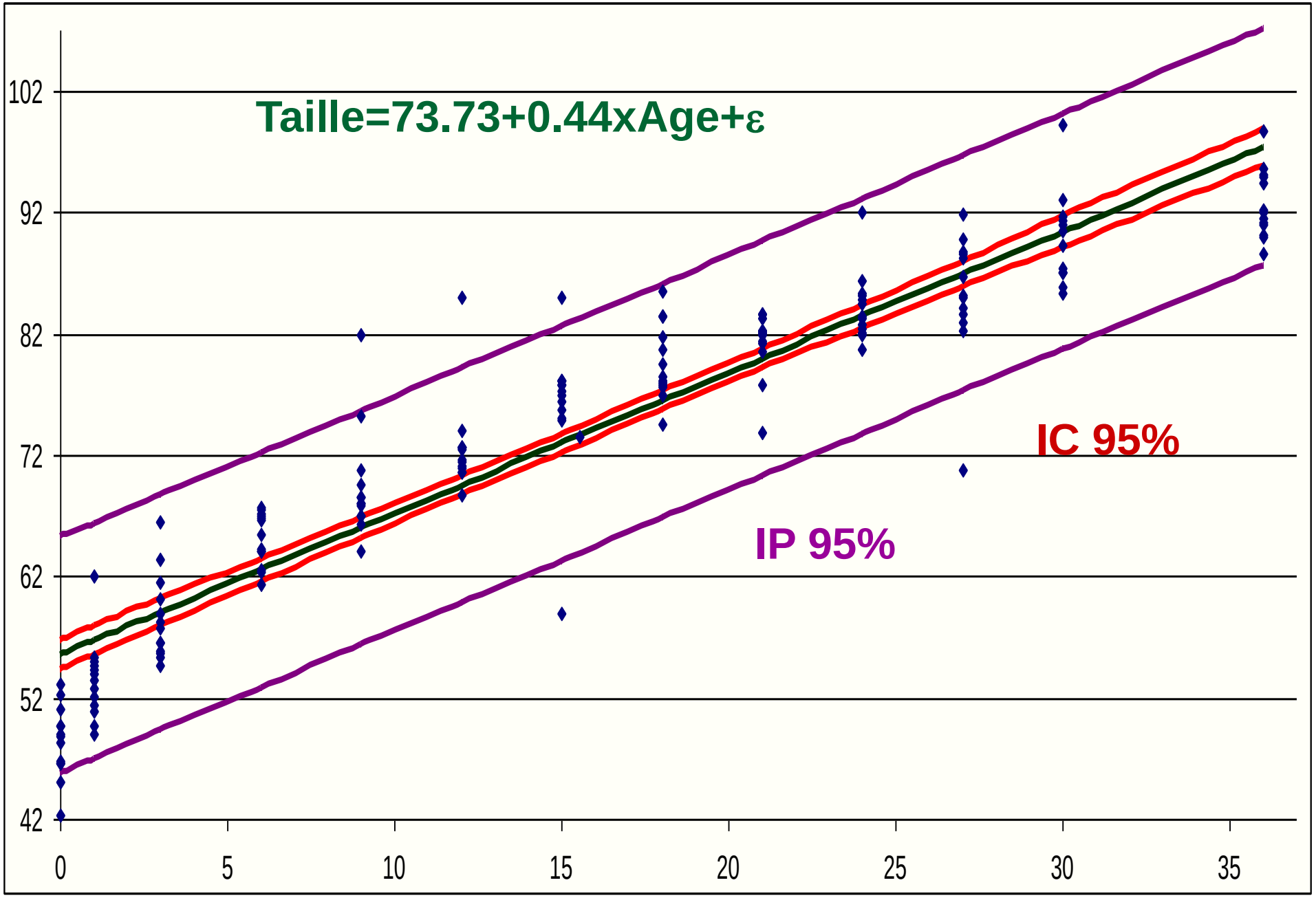

Taille=73.73+0.44xAge+ε
92
82
72
62
52
42
0
5
10
15
20
25
30
35
m_Y=73,6
IC 95%
m_X=15,5

- Intervalle de prédiction
  - Pour un Age (X) fixé, **prédiction** de la Taille (Y)

$$\mathbf{Y_p = a + b\ X}$$

$$\mathbf{Taille_p = 73{,}73 + 0{,}44\ Age}$$

  - Précision: $y_p \pm t_{n-2,\alpha/2}\sqrt{s^2_{y_p}}$



new.x = data.frame(AGE = c(18.2,2,50,108))

- Intervalle de **confiance** de la taille **estimée**:

IntConf<-predict(mod1,newdata=new.x,interval="confidence")

- Intervalle de **prédiction** de la taille **prédite**:

IntPred<-predict(mod1,newdata=new.x,interval="prediction")

Exercice

- Intervalle de **confiance** de la taille **estimée**:

IntConf

	fit	lwr	upr
1	81.69520	80.38770	83.00270
2	74.60441	73.16068	76.04814
3	95.61416	94.54321	96.68510
4	121.00093	120.15756	121.84430

- Intervalle de **prédiction** de la taille **prédite**:

IntPred

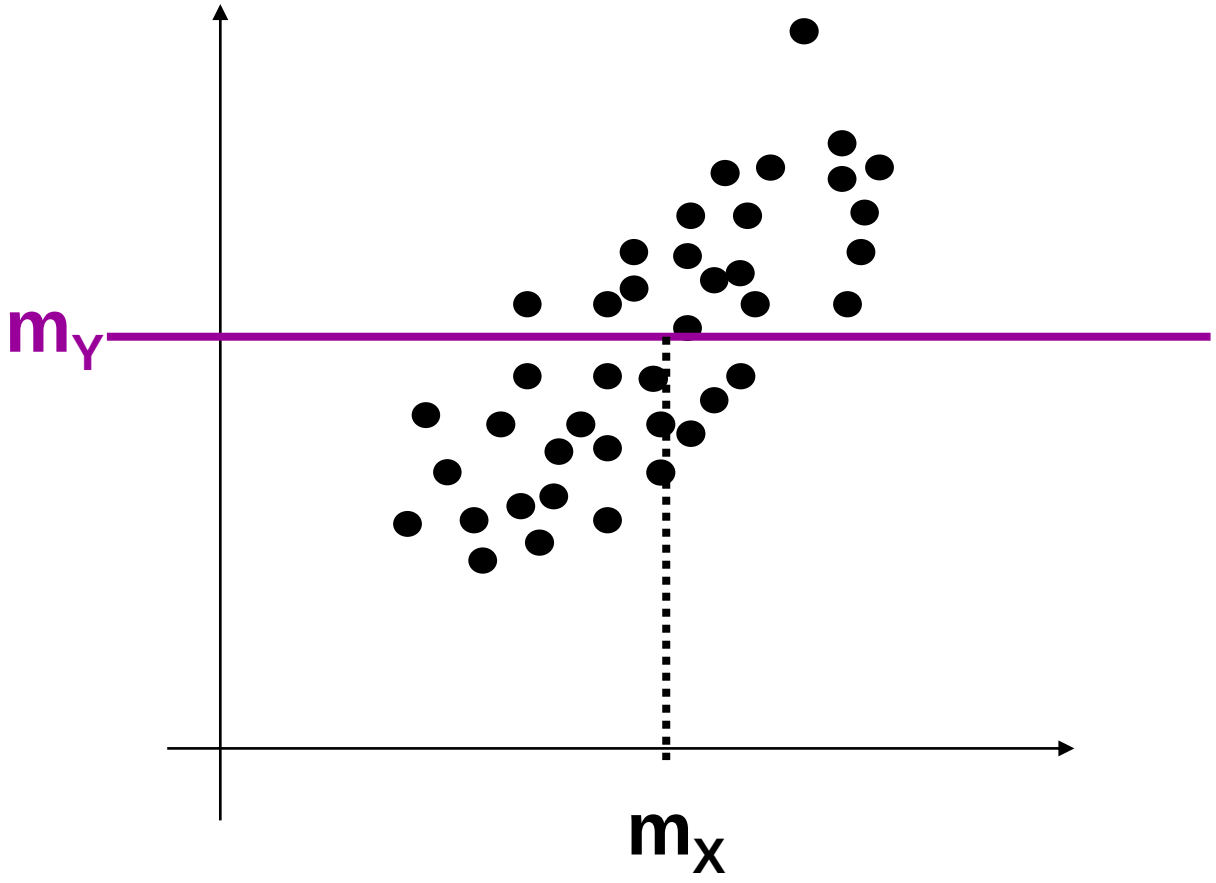
	fit	lwr	upr
1	81.69520	60.39828	102.99211
2	74.60441	53.29870	95.91012
3	95.61416	74.33045	116.89786
4	121.00093	99.72746	142.27439

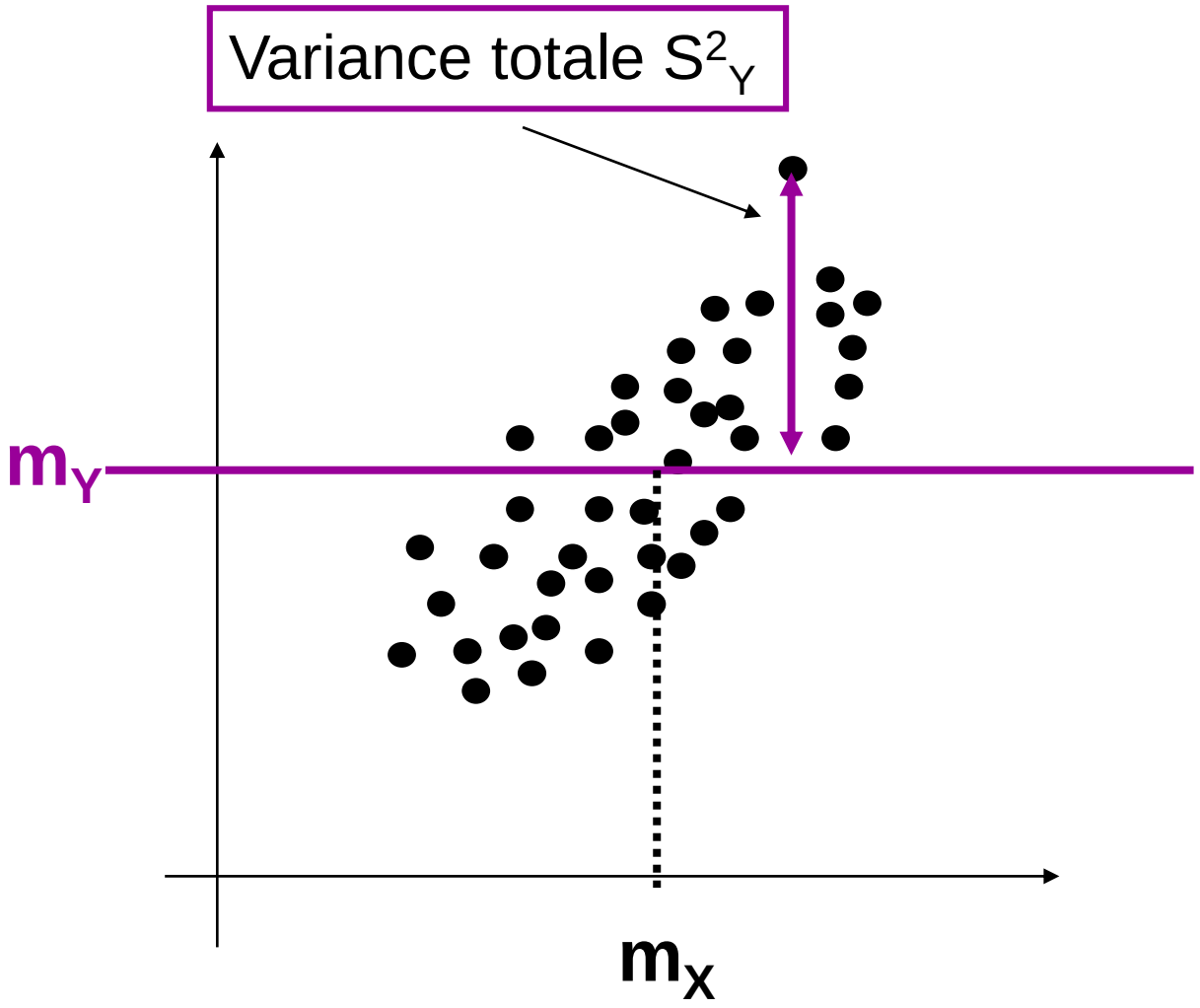
VI. Adéquation

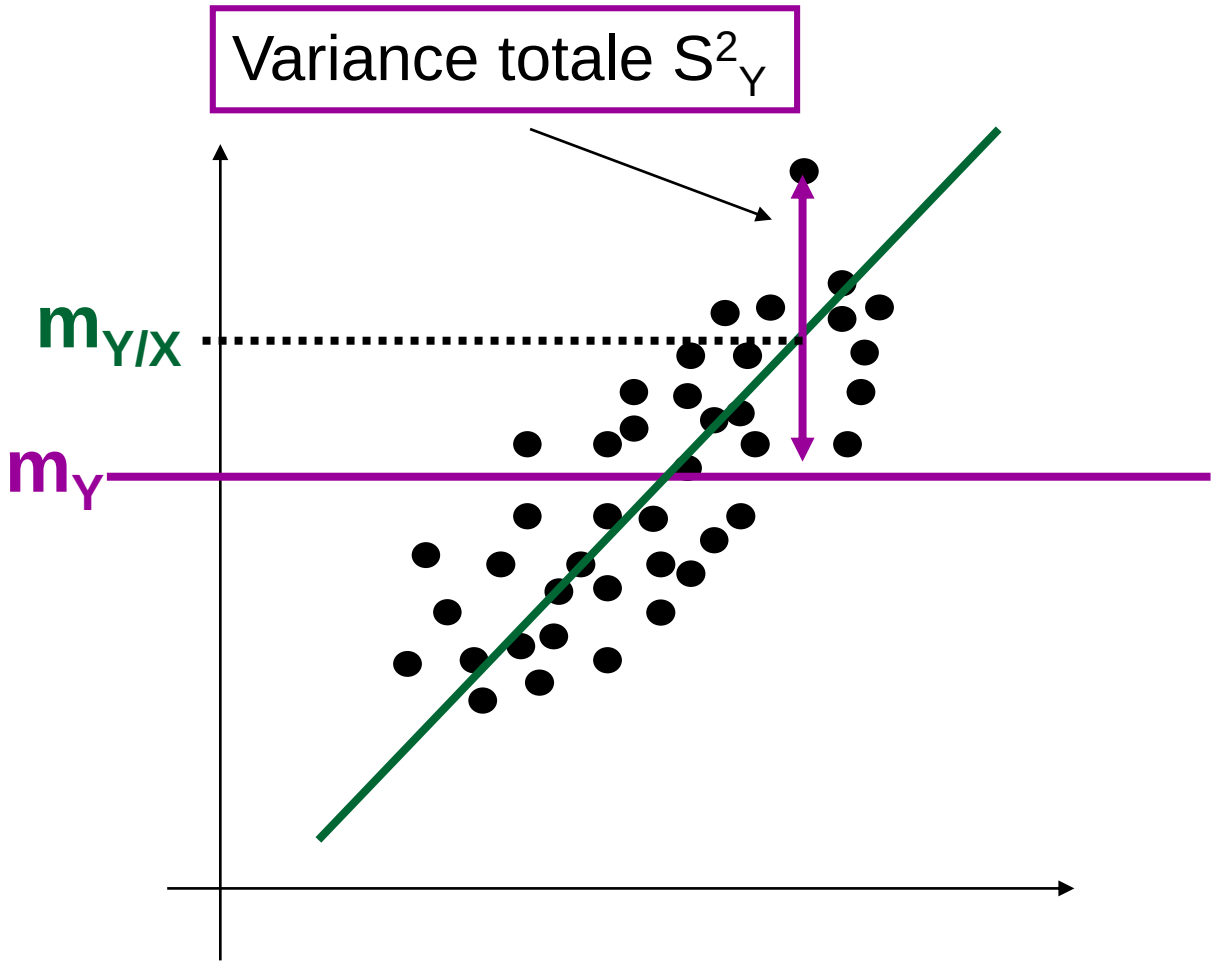
- Le modèle est-il un bon résumé des observations?

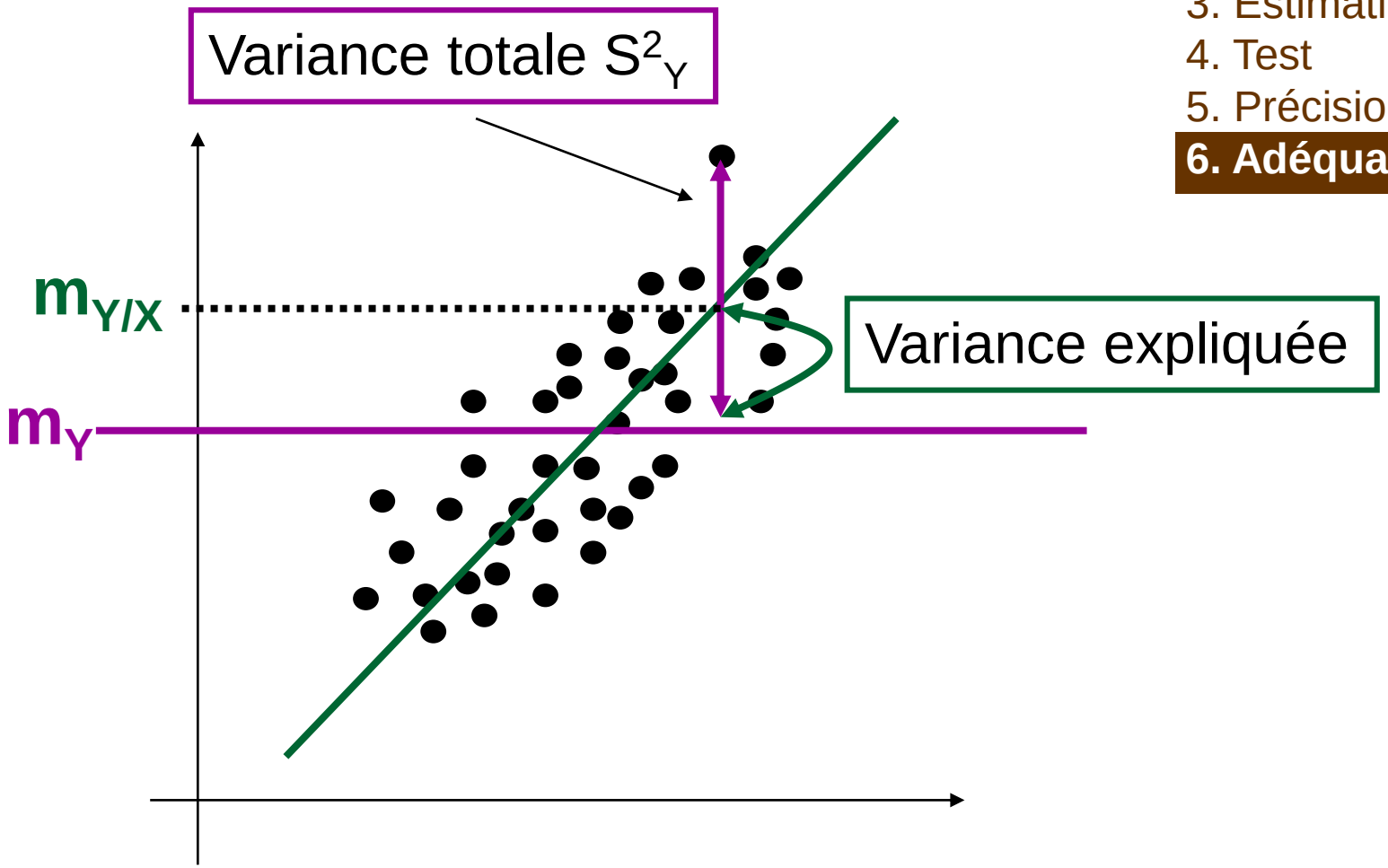
- Pourcentage de variance expliquée:

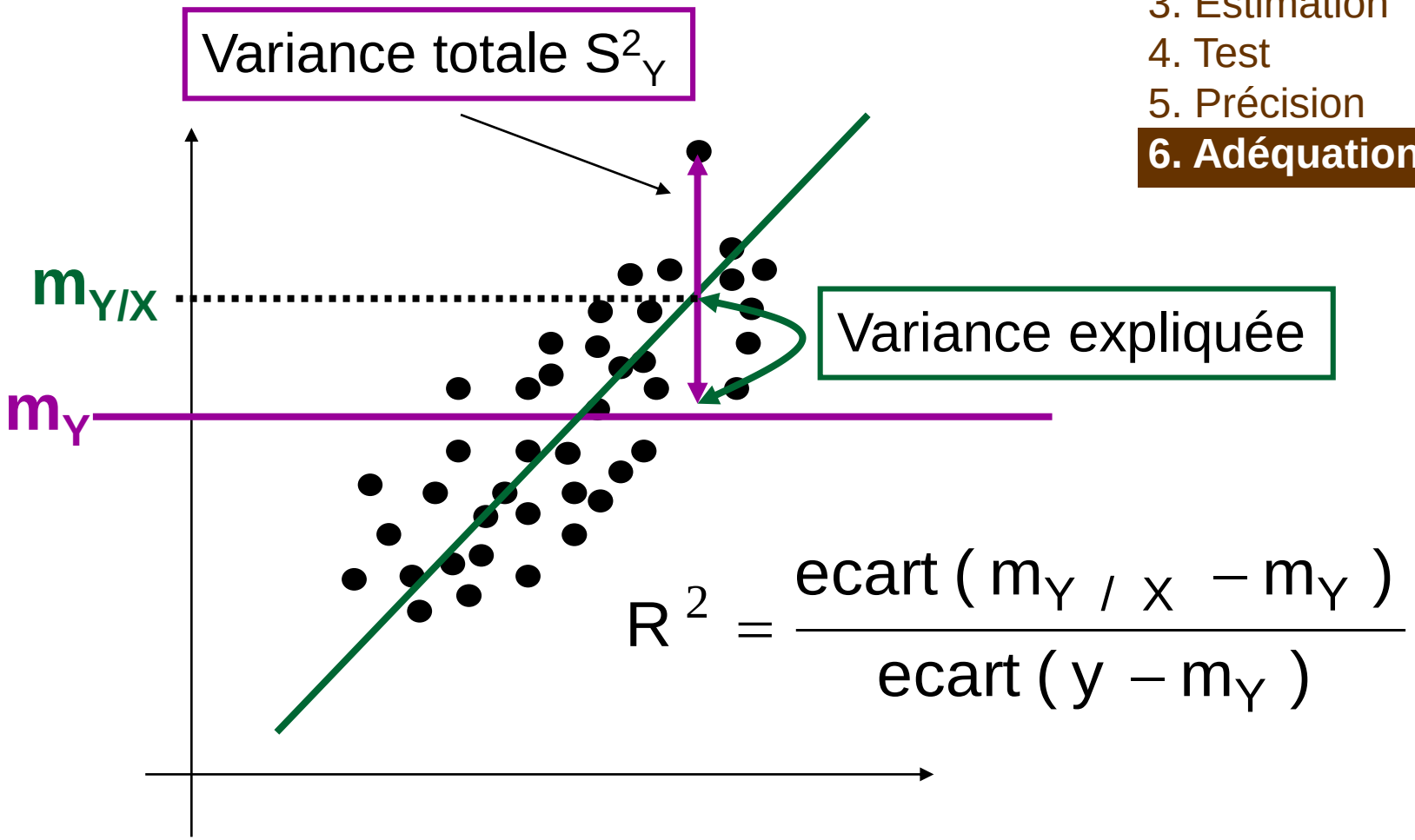
$$R^2 = \frac{\text{Part de variance expliquée par la régression}}{\text{Variance totale}}$$











- Pourcentage de variance expliquée:

$$R^2 = \frac{\sum \left(m_{Y/X_i} - m_Y \right)^2}{\sum \left(y_i - m_Y \right)^2}$$

- Exemple: **R^2=88%**

- Remarque:

  R: estimation du **coefficient de corrélation** entre X et Y

Exercice

- Estimation du **coefficient de corrélation** entre X et Y

r<-cor(TAIL,AGE)

- Estimation de R^2

*r*r*

ou

var(mod1$fitted.value)/var(TAIL)

Exercice

- Estimation du **coefficient de corrélation** entre X et Y

r<-cor(TAIL,AGE)

0.9545663

- Estimation de R^2

*r*r*

ou

0.9111967

var(mod1$fitted.value)/var(TAIL)

summary(mod1)

```
Call: lm(formula = TAIL ~ AGE)

Residuals:
    Min      1Q  Median      3Q     Max
-40.030  -6.899   2.999   8.120  24.999

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 73.729005   0.744041   99.09   <2e-16 ***
AGE          0.437703   0.005423   80.72   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 10.82 on 635 degrees of freedom
Multiple R-squared: 0.9112,     Adjusted R-squared: 0.9111
F-statistic:  6516 on 1 and 635 DF,  p-value: < 2.2e-16
```

VII. Régression multiple

- Plusieurs causes dans l'évolution de la taille:
  - Age (X_1)
  - Facteur socio-économiques (X_2)
  - Taux d'hormones de croissance (X_3)

$$E(Y / X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

- Estimation:
  - α, β_1, β_2, β_3 estimés en tenant compte des 3 VA

    => Ajustement

- Interactions

$$E(Y / X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$$

- Estimation:
  - α, β_1, β_2, β_3 estimés en tenant compte des 3 VA

    => Ajustement

- Interactions

$$E(Y / X_1, X_2, X_3) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_2 X_3$$

- Tests des β_1, β_2, β_3 à 0
- Interprétation identique
- Adéquation identique
- Approche pas à pas
- Choix des variables: notion de modèle
- Variables très corrélées

Exercice

- Prédire l'âge en fonction de 8 mesures
  - Crâne (BIP)
  - Tronc (LATHO)
  - Membres supérieurs et inférieurs (LOMAIN, PERPOIGN, PERCHEV, PIEDS)
  - Globales (STAT, POIDS)
- Echantillon de 1000 enfants de 2 à 16 ans

Exercice

- En moyenne:

$$AGE = \alpha + \beta 1 \times BIP + \beta 2 \times LATHO + \beta 3 \times LOMAIN + \beta 4 \times PERPOIGN + \beta 5 \times PERCHEV + \beta 6 \times PIEDS + \beta 7 \times STAT + \beta 8 \times POIDS$$

Exercice

- En moyenne:

AGE= α+β1xBIP+β2xLATHO+β3xLOMAIN+
β4xPERPOIGN+β5xPERCHEV+β6xPIEDS+
β7xSTAT+β8xPOIDS

TP<-read.csv2("C:\\BIOSTAT\\AGE.csv", header=TRUE)

Exercice

- En moyenne:

AGE= $\alpha+\beta 1 x$BIP$+\beta 2 x$LATHO$+\beta 3 x$LOMAIN$+$
$\beta 4 x$PERPOIGN$+\beta 5 x$PERCHEV$+\beta 6 x$PIEDS$+$
$\beta 7 x$STAT$+\beta 8 x$POIDS

TP<-read.csv2("C:\\BIOSTAT\\AGE.csv", header=TRUE)

- Statistiques descriptives *attach(TP)*

mean(AGE)

var(AGE)

hist(AGE)

Exercice

- En moyenne:

AGE= α+β1xBIP+β2xLATHO+β3xLOMAIN+
β4xPERPOIGN+β5xPERCHEV+β6xPIEDS+
β7xSTAT+β8xPOIDS

TP<-read.csv2("C:\\BIOSTAT\\AGE.csv", header=TRUE)

- Statistiques descriptives *attach(TP)*

mean(AGE) =10.373

var(AGE) =11.53541

Exercice

- Graphique:

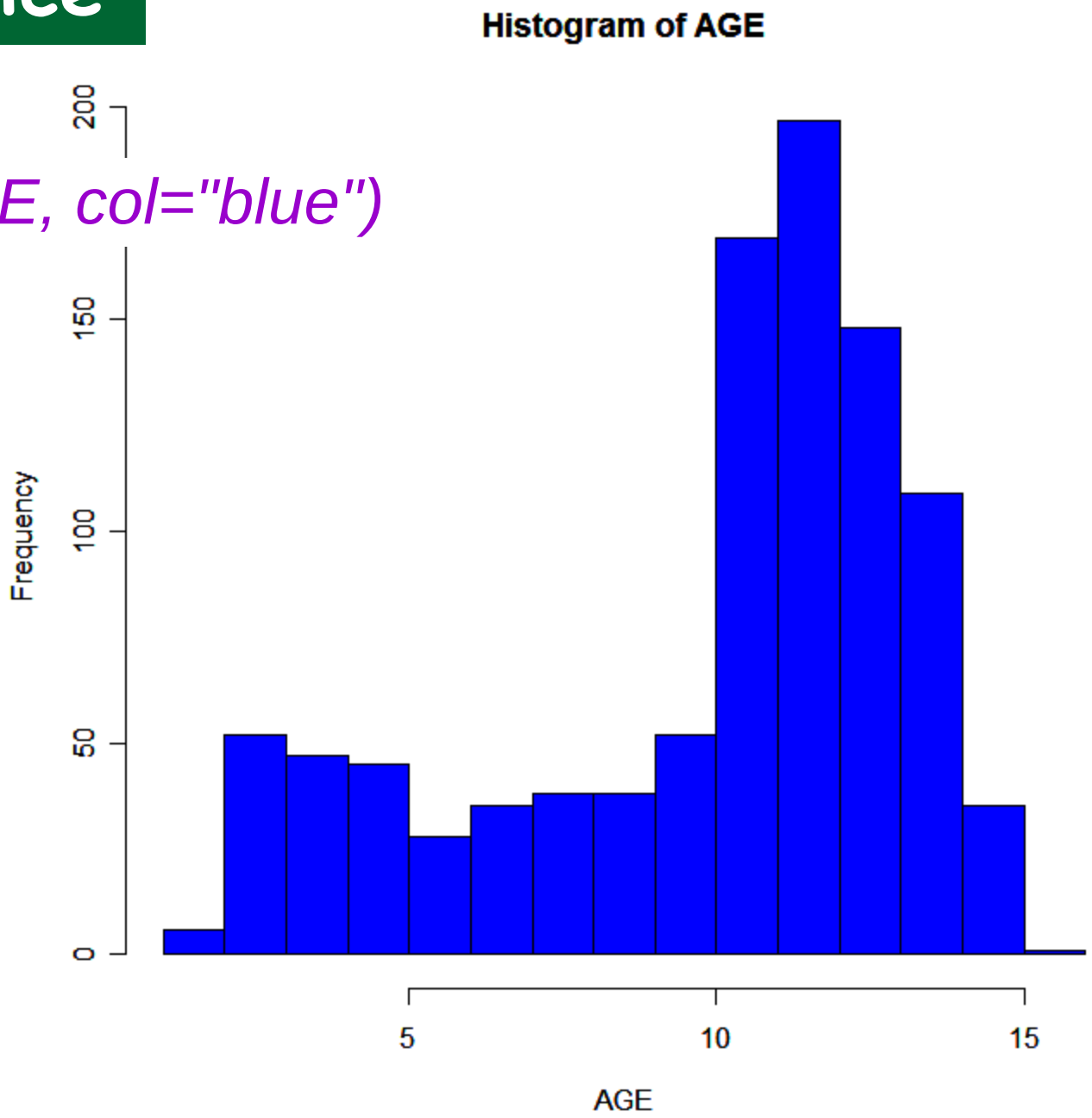
hist(AGE, col="blue")

DATA=data.frame(AGE,BIP, LOMAIN,LATHO, PIEDS, POIDS, PERPOIGN, PERCHEV, STAT)

splom (DATA)

Exercice

hist(AGE, col="blue")



Exercice

- Graphique:

hist(AGE, col="blue")

DATA=data.frame(AGE,BIP, LOMAIN,LATHO, PIEDS, POIDS, PERPOIGN, PERCHEV, STAT)

splom (DATA)

Exercice

- Graphique:

hist(AGE, col="blue")

DATA=data.frame(AGE,BIP, LOMAIN,LATHO, PIEDS, POIDS, PERPOIGN, PERCHEV, STAT)

splom (DATA)

Erreur : impossible de trouver la fonction "splom"

Exercice

- Graphique:

hist(AGE, col="blue")

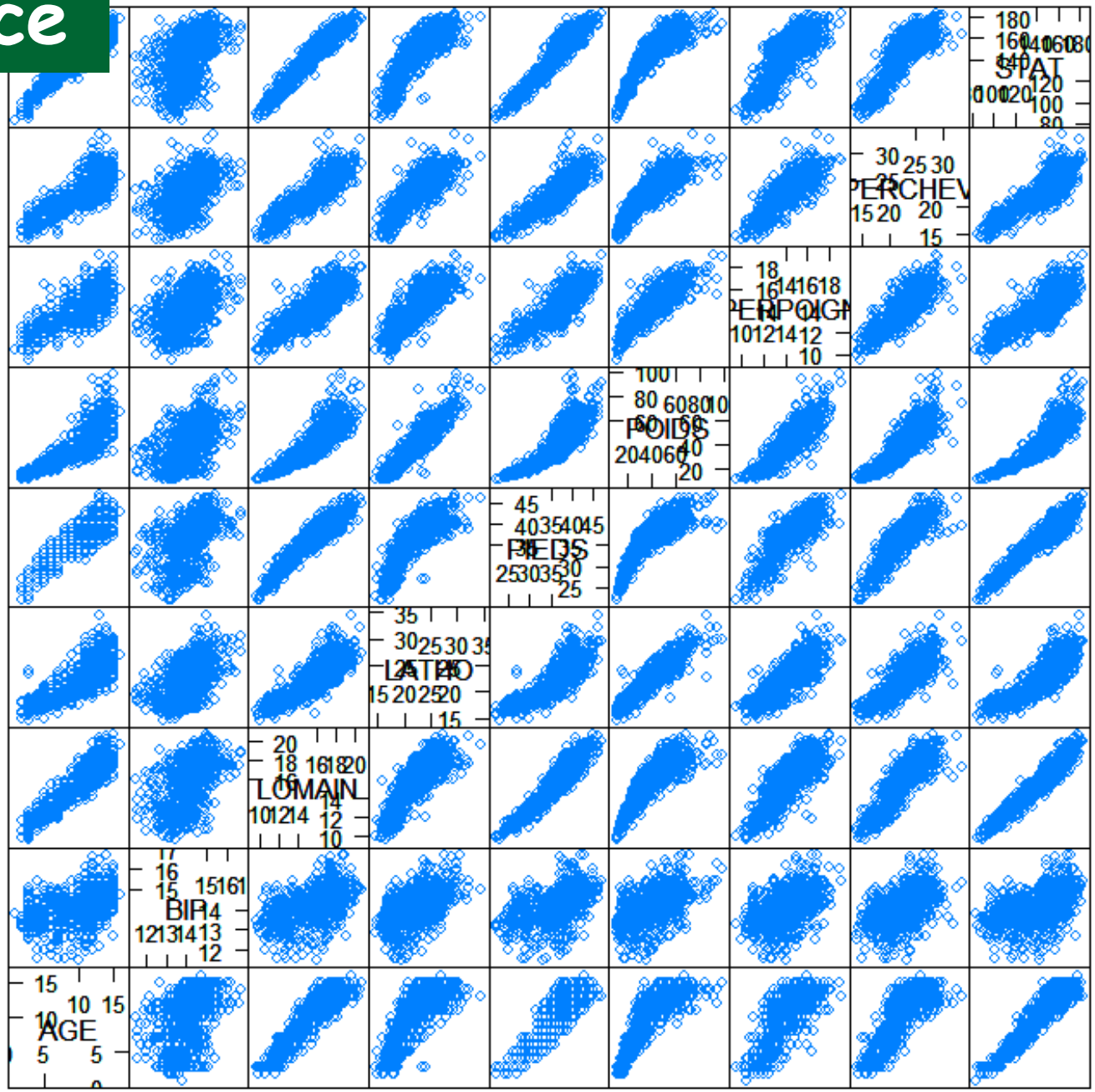
DATA=data.frame(AGE,BIP, LOMAIN,LATHO, PIEDS, POIDS, PERPOIGN, PERCHEV, STAT)

splom (DATA)

Erreur : impossible de trouver la fonction "splom"

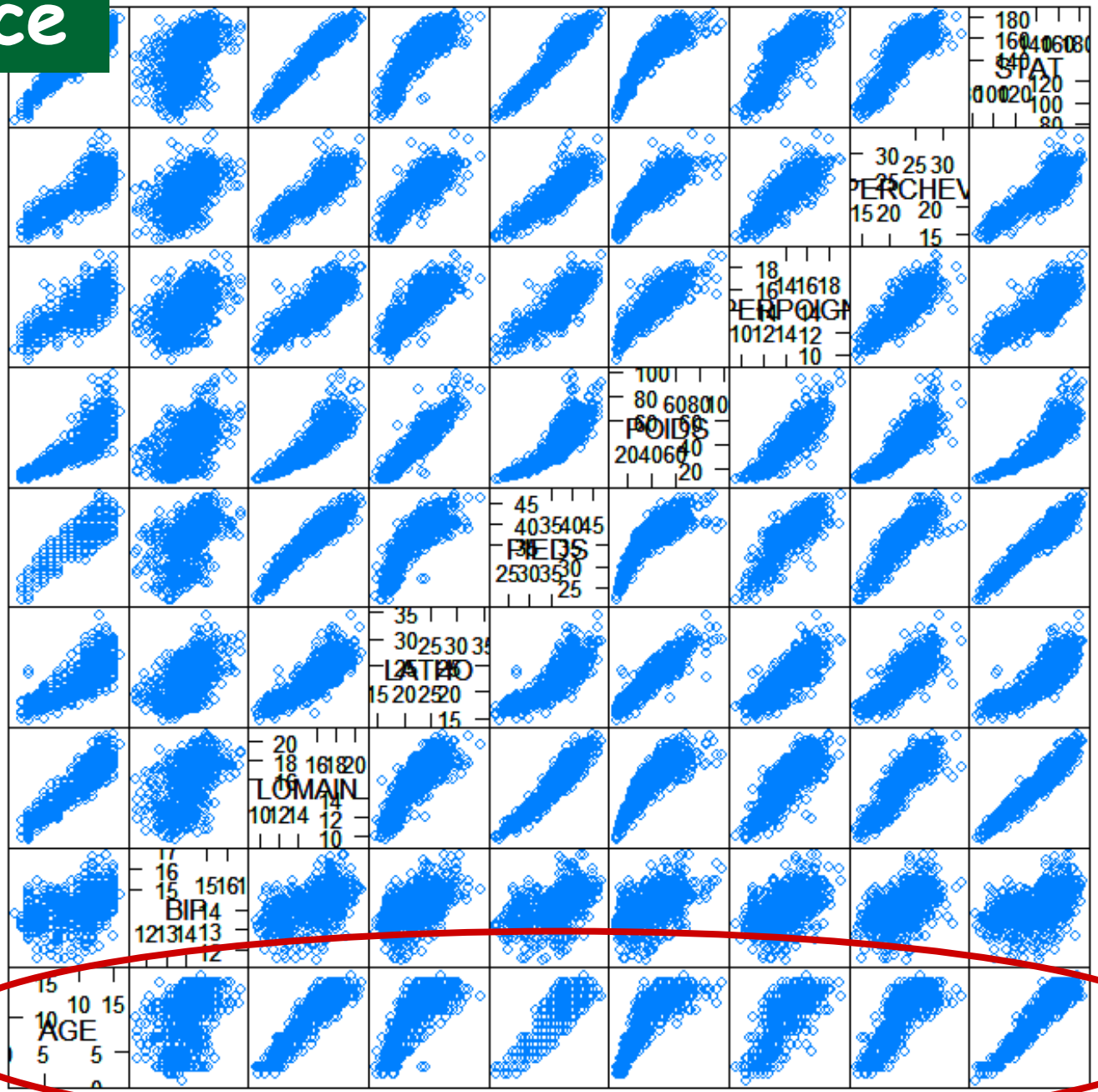
⇒ package lattice

Exercice



Scatter Plot Matrix

Exercice



Scatter Plot Matrix

Exercice

- Estimation:

AGE= α+β1xBIP+β2xLATHO+β3xLOMAIN+β4xPERPOIGN
+β5xPERCHEV+β6xPIEDS+β7xSTAT+β8xPOIDS

reg1<-lm(AGE ~ BIP+LATHO+LOMAIN+PERPOIGN +PERCHEV+PIEDS+STAT+POIDS)

Exercice

summary(reg1)

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:
    Min      1Q  Median      3Q     Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02    0.611    0.54156
LATHO        1.219e-01  2.659e-02    4.583    5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02    1.704    0.08877 .
PERPOIGN -1.370e-01  4.695e-02   -2.917    0.00361 **
PERCHEV  -4.654e-02  2.597e-02   -1.792    0.07341 .
PIEDS        7.823e-04  2.612e-02    0.030    0.97611
STAT         1.546e-01  7.263e-03   21.289   < 2e-16 ***
POIDS       -2.047e-02  7.153e-03   -2.861    0.00431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 991 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:
    Min      1Q  Median      3Q     Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02    0.611    0.54156
LATHO        1.219e-01  2.659e-02    4.583    5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02    1.704    0.08877 .
PERPOIGN -1.370e-01  4.695e-02   -2.917    0.00361 **
PERCHEV  -4.654e-02  2.597e-02   -1.792   0.07341 .
PIEDS         7.823e-04  2.612e-02    0.030   0.97611
STAT          1.546e-01  7.263e-03  21.289   < 2e-16 ***
POIDS       -2.047e-02  7.153e-03   -2.861   0.00431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 991 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

régression

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS
Residuals:
    Min      1Q  Median      3Q     Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate    Std. Error   t value    Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02    0.611    0.54156
LATHO        1.219e-01  2.659e-02    4.583    5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02    1.704    0.08877 .
PERPOIGN -1.370e-01  4.695e-02   -2.917    0.00361 **
PERCHEV  -4.654e-02  2.597e-02   -1.792    0.07341 .
PIEDS        7.823e-04  2.612e-02    0.030    0.97611
STAT         1.546e-01  7.263e-03   21.289   < 2e-16 ***
POIDS       -2.047e-02  7.153e-03   -2.861    0.00431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 991 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

Prédicteurs,
Variables explicatives

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:
    Min      1Q  Median      3Q     Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02    0.611    0.54156
LATHO        1.219e-01  2.659e-02    4.583    5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02    1.704    0.08877 .
PERPOIGN -1.370e-01  4.695e-02   -2.917    0.00361 **
PERCHEV  -4.654e-02  2.597e-02   -1.792    0.07341 .
PIEDS        7.823e-04  2.612e-02    0.030    0.97611
STAT         1.546e-01  7.263e-03   21.289   < 2e-16 ***
POIDS       -2.047e-02  7.153e-03   -2.861    0.00431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 991 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

estimations des paramètres, ajustées

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:
     Min       1Q   Median       3Q      Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02    0.611    0.54156
LATHO        1.219e-01  2.659e-02    4.583    5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02    1.704    0.08877 .
PERPOIGN    -1.370e-01  4.695e-02   -2.917    0.00361 **
PERCHEV     -4.654e-02  2.597e-02   -1.792    0.07341 .
PIEDS        7.823e-04  2.612e-02    0.030    0.97611
STAT         1.546e-01  7.263e-03   21.289    < 2e-16 ***
POIDS       -2.047e-02  7.153e-03   -2.861    0.00431 **
---
Si
Re
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

estimations des paramètres, ajustées

AGE=-13+0,03BIP+0,1LATHO+0,01LOMAIN-0,14PERPOIGN-0,05PERCHEV +0,001PIEDS+0,2STAT-0,02POIDS

Exercice

summary(reg1)

Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:

Min	1Q	Median	3Q	Max
-3.12658	-0.72416	-0.04954	0.67239	4.36643

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.300e+01	8.684e-01	-14.966	< 2e-16 ***
BIP	3.312e-02	5.423e-02	0.611	0.54156
LATHO	1.219e-01	2.659e-02	4.583	5.17e-06 ***
LOMAIN	1.013e-01	5.947e-02	1.704	0.08877 .
PERPOIGN	-1.370e-01	4.695e-02	-2.917	0.00361 **
PERCHEV	-4.654e-02	2.597e-02	-1.792	0.07341 .
PIEDS	7.823e-04	2.612e-02	0.030	0.97611
STAT	1.546e-01	7.263e-03	21.289	< 2e-16 ***
POIDS	-2.047e-02	7.153e-03	-2.861	0.00431 **

Si

signification

AGE=-13+0,03BIP+0,1LATHO+0,01LOMAIN-0,14PERPOIGN-0,05PERCHEV
+0,001PIEDS+0,2STAT-0,02POIDS

Re
Multiple R-squared: 0.8989, Adjusted R-squared: 0.8981
F-statistic: 1102 on 8 and 991 DF, p-value: < 2.2e-16

Exercice

summary(reg1)

```
Call:
lm(formula = AGE ~ BIP+LATHO+LOMAIN+PERPOIGN+PERCHEV+PIEDS+STAT+POIDS)
Residuals:
     Min       1Q   Median       3Q      Max
-3.12658 -0.72416 -0.04954  0.67239  4.36643

Coefficients:
             Estimate     Std. Error    t value     Pr(>|t|)
(Intercept) -1.300e+01  8.684e-01  -14.966   < 2e-16 ***
BIP          3.312e-02  5.423e-02     0.611     0.54156
LATHO        1.219e-01  2.659e-02     4.583     5.17e-06 ***
LOMAIN       1.013e-01  5.947e-02     1.704     0.08877 .
PERPOIGN -1.370e-01  4.695e-02    -2.917     0.00361 **
PERCHEV  -4.654e-02  2.597e-02    -1.792    0.07341 .
PIEDS        7.823e-04  2.612e-02     0.030    0.97611
STAT         1.546e-01  7.263e-03   21.289    < 2e-16 ***
POIDS       -2.047e-02  7.153e-03    -2.861    0.00431 **
---
Si
Re
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

signification

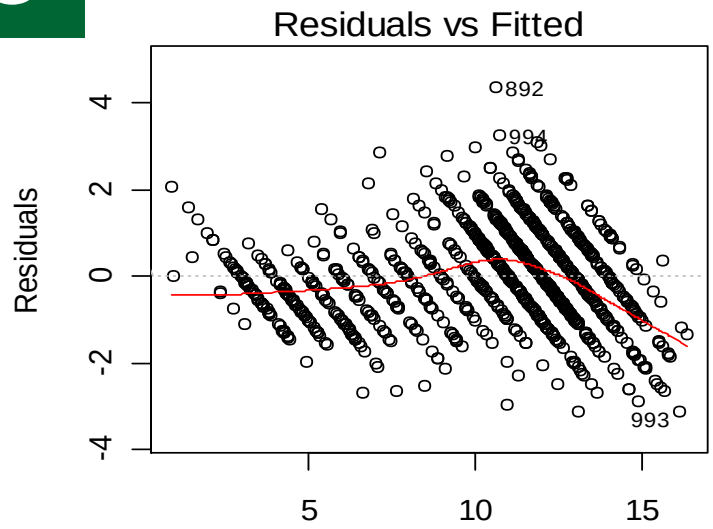
AGE=-13+0,03BIP+0,1LATHO+0,01LOMAIN-0,14PERPOIGN-0,05PERCHEV +0,001PIEDS+0,2STAT-0,02POIDS

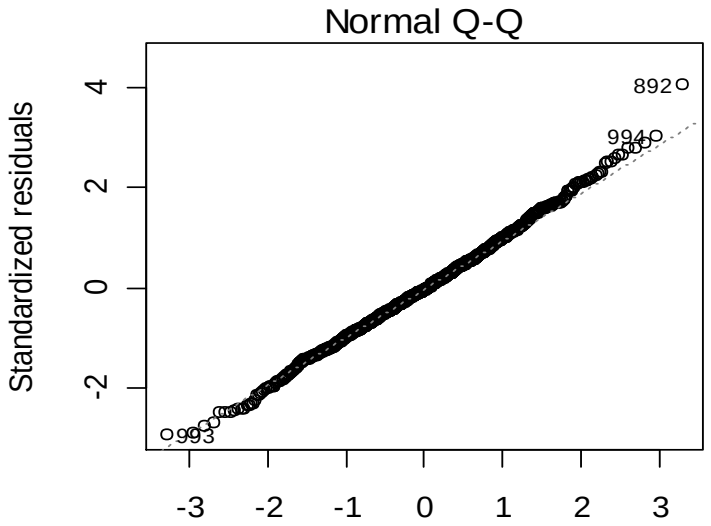
- Que faut-il regarder ensuite ?
  - conditions d'application
  - intervalles de confiance des paramètres
  - adéquation: R^2

Conditions d'applications

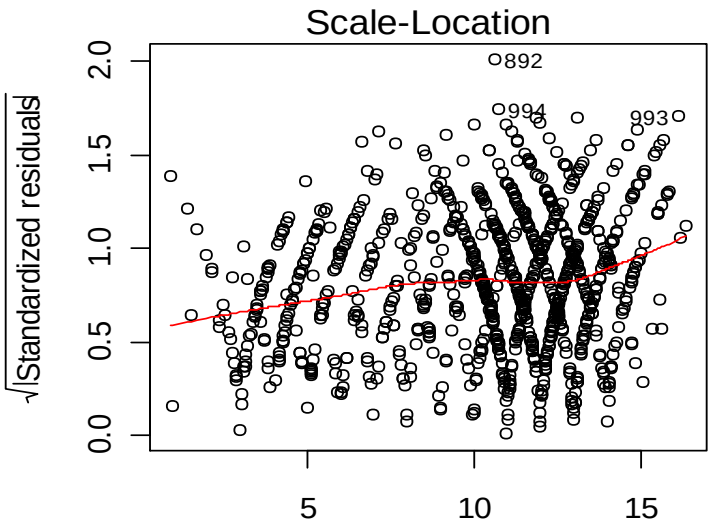
- ***L(Y/X)*~>$\mathcal{N}$** → qqnorm
- ***V(Y/X)*** constantes pour tout X → plot(age estimé, résidus)
  homoscédasticité
- à X donné, Y_i indépendants → protocole
- La régression est linéaire → plot(AGE,prédicteurs)

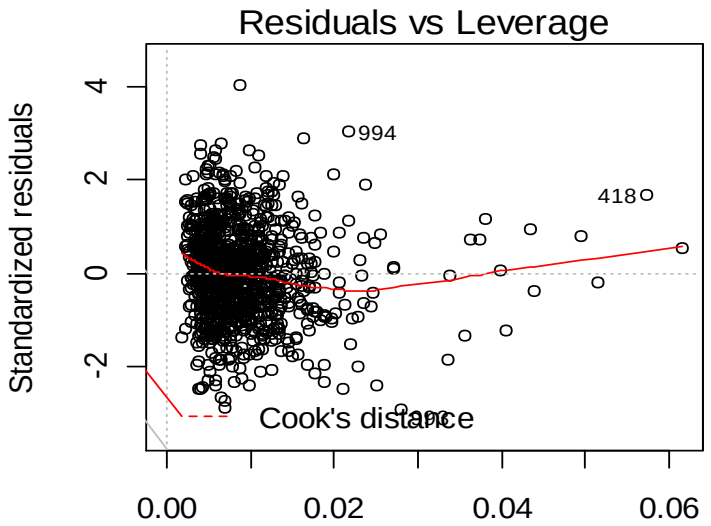
Exercice





par(mfrow=c(2,2)
plot(reg1)





Exercice

- Intervalles de confiance des paramètres

confint(reg1)

Exercice

- Intervalles de confiance des paramètres

confint(reg1)

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Exercice

- Intervalles de confiance des paramètres

confint(reg1)

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Exercice

- Intervalles de confiance des paramètres

confint(reg1)

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Exercice

- Intervalles de confiance des paramètres

confint(reg1)

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Exercice

- Intervalles de confiance des paramètres

confint(reg1)

	2.5 %	97.5 %
(Intercept)	-14.70058351	-11.292408725
BIP	-0.07330209	0.139535454
LATHO	0.06968244	0.174040764
LOMAIN	-0.01538828	0.218001320
PERPOIGN	-0.22908876	-0.044831392
PERCHEV	-0.09750881	0.004420695
PIEDS	-0.05047023	0.052034764
STAT	0.14037312	0.168879663
POIDS	-0.03450573	-0.006430739

Exercice

- Adéquation: R^2

var(reg1$fitted.value)/var(AGE)

Exercice

- Adéquation: R^2

var(reg1$fitted.value)/var(AGE)

0.8989102

```
POIDS         -2.047e-02  7.153e-03    -2.861    0.00431 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.084 on 991 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8981
F-statistic:  1102 on 8 and 991 DF,  p-value: < 2.2e-16
```

Exercice

- **Prédiction**
  - nouvelles valeurs des prédicteurs
  - ex: AGE?

BIP	LA THO	LO MAIN	STAT	PER POiGN	PER CHEV	POIDS	PIEDS
14,2	23,5	15,9	148,2	15,5	23	36	38

new.x<-data.frame(BIP=14.2, LATHO=23.5, LOMAIN=15.9, STAT=148.2, PERPOIGN=15.5, PERCHEV=23, POIDS=36, PIEDS=38)

new.x

Exercice

- **Prédiction**
  - nouvelles valeurs des prédicteurs
  - ex: AGE?

BIP	LA THO	LO MAIN	STAT	PER POiGN	PER CHEV	POIDS	PIEDS
14,2	23,5	15,9	148,2	15,5	23	36	38

new.x<-data.frame(BIP=14.2, LATHO=23.5, LOMAIN=15.9, STAT=148.2, PERPOIGN=15.5, PERCHEV=23, POIDS=36, PIEDS=38)

new.x

```
  BIP LATHO LOMAIN  STAT PERPOIGN PERCHEV POIDS PIEDS
1 14.2  23.5   15.9 148.2    15.5      23    36    38
```

Exercice

age réel = 11 ans

- **Intervalle de confiance**

predict(reg1,newdata=new.x,se.fit=TRUE,interval="confidence")

```
$fit
       fit      lwr      upr
1 10.96339 10.77563 11.15115
```

- **Intervalle de prédiction**

predict(reg1,newdata=new.x,se.fit=TRUE,interval="prediction")

```
$fit
       fit      lwr      upr
1 10.96339 8.827496 13.09928
```

Sélection de variables

- Guillaume d'Ockham, 1285-1349

  « *Les multiples ne doivent*

  pas êtres utilisés sans nécessité »

  = principe de **parcimonie**

  => ne pas ajouter de nouvelles variables tant que celles présentes suffisent

  => balance entre explication / prédiction

  trop de variables: explication + / prédiction –

overfitting~hyperadéquation

Sélection de variables

- **Critère de sélection**

Akaike Information Criterion AIC

$$AIC=2p-2\ln(L)$$

nombre de paramètres

vraisemblance

=>AIC le plus petit possible

Sélection de variables

- **Critère de sélection**

Akaike Information Criterion AIC

$$AIC=2p-2\ln(L)$$

nombre de paramètres

vraisemblance

=>AIC le plus petit possible

Exercice

- **Sélection de variables: pas à pas**

reglow<-lm(AGE ~ STAT)
summary(reglow)

Exercice

- **Sélection de variables: pas à pas**

reglow<-lm(AGE ~ STAT)
summary(reglow)

```
Call: lm(formula = AGE ~  STAT)
Residuals:
     Min       1Q   Median       3Q      Max
-3.22224 -0.74277 -0.02807  0.73413  4.29016

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.909459   0.244761  -48.66   <2e-16 ***
STAT          0.153978   0.001674   91.98   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.104 on 998 degrees of freedom
Multiple R-squared: 0.8945,     Adjusted R-squared: 0.8944
F-statistic:  8460 on 1 and 998 DF,  p-value: < 2.2e-16
```

Exercice

- **Sélection de variables: pas à pas**

aicreg<-stepAIC(reg1,scope=list(upper=reg1,lower=reglow),direction=c("both"))

```
Start:  AIC=170.67
AGE ~  BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + PIEDS +
    STAT + POIDS

             Df  Sum of Sq    RSS    AIC
- PIEDS       1    0.0011   1165.0  168.68
- BIP         1    0.4384   1165.4  169.05
<none>                     1165.0  170.68
- LOMAIN      1    3.4116   1168.4  171.60
- PERCHEV     1    3.7755   1168.7  171.91
- POIDS       1    9.6243   1174.6  176.90
- PERPOIGN    1   10.0043   1175.0  177.23
- LATHO       1   24.6905   1189.6  189.65

Step:  AIC=168.68
AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + PERCHEV + STAT + POIDS

             Df  Sum of Sq    RSS    AIC
- BIP         1    0.4433   1165.4  167.06
<none>                     1165.0  168.68
- LOMAIN      1    3.6758   1168.6  169.83
- PERCHEV     1    4.0625   1169.0  170.16
+ PIEDS       1    0.0011   1165.0  170.68
- POIDS       1    9.9216   1174.9  175.16
- PERPOIGN    1   10.4397   1175.4  175.60
- LATHO       1   24.7051   1189.7  187.66
```

Step: AIC=167.06
AGE ~ LATHO + LOMAIN + PERPOIGN + PERCHEV + STAT + POIDS

	Df	Sum of Sq	RSS	AIC
<none>			1165.4	167.06
- PERCHEV	1	3.7648	1169.2	168.28
- LOMAIN	1	3.8633	1169.2	168.37
+ BIP	1	0.4433	1165.0	168.68
+ PIEDS	1	0.0060	1165.4	169.05
- POIDS	1	9.7153	1175.1	173.36
- PERPOIGN	1	10.6173	1176.0	174.12
- LATHO	1	26.2754	1191.7	187.35

Exercice

Sélection de variables: modèle final

regfin<-lm(AGE ~ LATHO+LOMAIN+PERPOIGN+PERCHEV+STAT+POIDS)

```
Call:
lm(formula=AGE ~ LATHO+LOMAIN+PERPOIGN+PERCHEV+STAT+POIDS)
Residuals:
     Min       1Q   Median       3Q      Max
-3.14469 -0.73537 -0.04168  0.68040  4.37259
Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -12.611320    0.592893  -21.271   < 2e-16 ***
LATHO           0.124299    0.026270    4.732    2.55e-06 ***
LOMAIN          0.104090    0.057371    1.814    0.0699 .
PERPOIGN -0.137719    0.045788   -3.008    0.0027 **
PERCHEV   -0.044138   0.024643   -1.791    0.0736 .
STAT            0.154353   0.006506   23.724   < 2e-16 ***
POIDS          -0.020256   0.007040   -2.877    0.0041 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.083 on 993 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8983
F-statistic:  1471 on 6 and 993 DF,  p-value: < 2.2e-16
```

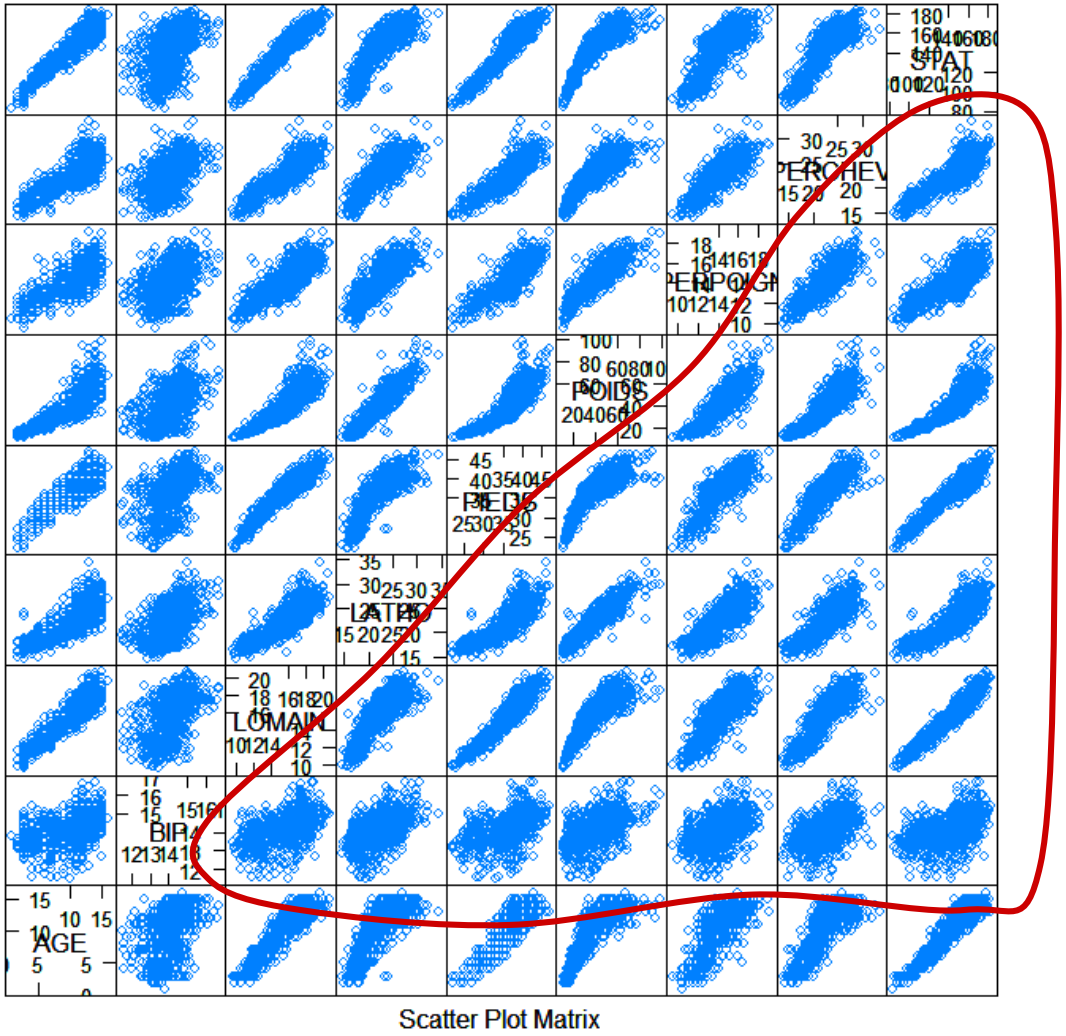
Exercice

Sélection de variables: modèle final

regfin<-lm(AGE ~ LATHO+LOMAIN+PERPOIGN+PERCHEV+STAT+POIDS)

```
Call:
lm(formula=AGE ~ LATHO+LOMAIN+PERPOIGN+PERCHEV+STAT+POIDS)
Residuals:
     Min       1Q   Median       3Q      Max
-3.14469 -0.73537 -0.04168  0.68040  4.37259
Coefficients:
                Estimate   Std. Error   t value   Pr(>|t|)
(Intercept)   -12.611320    0.592893  -21.271   < 2e-16 ***
LATHO           0.124299    0.026270    4.732    2.55e-06 ***
LOMAIN          0.104090    0.057371    1.814    0.0699 .
PERPOIGN -0.137719    0.045788   -3.008    0.0027 **
PERCHEV   -0.044138   0.024643   -1.791    0.0736 .
STAT            0.154353   0.006506   23.724   < 2e-16 ***
POIDS          -0.020256   0.007040   -2.877   0.0041 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.083 on 993 degrees of freedom
Multiple R-squared: 0.8989,     Adjusted R-squared: 0.8983
F-statistic:  1471 on 6 and 993 DF,  p-value: < 2.2e-16
```

Interaction entre les variables



Scatter Plot Matrix

- **Interaction**

En moyenne:

AGE= α+β1.BIP+β2.PIEDS+β3.LATHO+β4.LOMAIN
+β5.PERPOIGN +β6.PERCHEV+β7.STAT+β8.POIDS
+β9.LOMAIN.STAT+β10.LATHO.POIDS+β11.POIDS.STAT

regint<-lm(AGE ~ BIP + PIEDS + LATHO+LOMAIN+PERPOIGN+PERCHEV+STAT +POIDS+LOMAIN:STAT+LATHO:POIDS+POIDS:STAT)

```
Call:
lm(formula = AGE ~ BIP + PIEDS + LATHO + LOMAIN + PERPOIGN +
    PERCHEV + STAT + POIDS + LOMAIN:STAT + LATHO:POIDS + POIDS:STAT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2000 -0.7007 -0.0105  0.6227  3.6067

Coefficients:
                  Estimate     Std. Error      t value       Pr(>|t|)
(Intercept)     -2.013e+01    2.810e+00    -7.163      1.54e-12 ***
BIP              8.447e-02     5.141e-02     1.643       0.100703
PIEDS           -4.617e-02     2.505e-02    -1.843       0.065579 .
LATHO            1.696e-01    5.859e-02     2.894       0.003884 **
LOMAIN           4.814e-01    2.871e-01     1.677       0.093900 .
PERPOIGN        -1.999e-01    4.497e-02    -4.446       9.76e-06 ***
PERCHEV         -6.689e-02     2.460e-02    -2.719       0.006654 **
STAT             1.966e-01    1.769e-02    11.111       < 2e-16 ***
POIDS            1.825e-01     5.289e-02     3.451       0.000582 ***
LOMAIN:STAT -2.178e-03    1.891e-03    -1.152       0.249747
LATHO:POIDS -1.262e-03    1.165e-03    -1.083       0.278864
STAT:POIDS      -9.156e-04  4.195e-04    -2.183       0.029305 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 988 degrees of freedom
Multiple R-squared:  0.9104,          Adjusted R-squared:  0.9094
F-statistic: 912.7 on 11 and 988 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = AGE ~ BIP + PIEDS + LATHO + LOMAIN + PERPOIGN +
    PERCHEV + STAT + POIDS + LOMAIN:STAT + LATHO:POIDS + POIDS:STAT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2000 -0.7007 -0.0105  0.6227  3.6067

Coefficients:
                  Estimate     Std. Error      t value       Pr(>|t|)
(Intercept)     -2.013e+01    2.810e+00    -7.163      1.54e-12 ***
BIP              8.447e-02     5.141e-02     1.643       0.100703
PIEDS           -4.617e-02     2.505e-02    -1.843       0.065579 .
LATHO            1.696e-01    5.859e-02     2.894       0.003884 **
LOMAIN           4.814e-01    2.871e-01     1.677       0.093900 .
PERPOIGN        -1.999e-01    4.497e-02    -4.446       9.76e-06 ***
PERCHEV         -6.689e-02    2.460e-02    -2.719       0.006654 **
STAT             1.966e-01    1.769e-02    11.111       < 2e-16 ***
POIDS            1.825e-01    5.289e-02     3.451       0.000582 ***
LOMAIN:STAT -2.178e-03    1.891e-03    -1.152       0.249747
LATHO:POIDS -1.262e-03    1.165e-03    -1.083       0.278864
STAT:POIDS      -9.156e-04  4.195e-04    -2.183       0.029305 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 988 degrees of freedom
Multiple R-squared:  0.9104,          Adjusted R-squared:  0.9094
F-statistic: 912.7 on 11 and 988 DF,  p-value: < 2.2e-16
```

Exercice

aicreg<-stepAIC(reg1,scope=list(upper=regint,lower=reglow), direction=c("both"))

```
Start:  AIC=170.67
AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + STAT + POIDS + PIEDS + PERCHEV

                Df  Sum of Sq    RSS     AIC
+ STAT:POIDS     1   130.617     1034.3  53.753
+ LOMAIN:STAT    1   115.456     1049.5  68.305
+ LATHO:POIDS    1    75.661     1089.3  105.521
- PIEDS          1     0.001     1165.0  168.676
- BIP            1     0.438     1165.4  169.051
<none>                           1165.0  170.675
- LOMAIN         1     3.412     1168.4  171.599
- PERCHEV        1     3.776     1168.7  171.911
- POIDS          1     9.624     1174.6  176.902
- PERPOIGN       1    10.004     1175.0  177.226
- LATHO          1    24.690     1189.6  189.648
```

(...)

Exercice

```
Step:  AIC=53.75
AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + STAT + POIDS + PIEDS +
    PERCHEV + STAT:POIDS

                Df Sum of Sq    RSS     AIC
<none>                1034.3  53.753
- BIP            1     2.998   1037.3   54.648
+ LOMAIN:STAT    1     0.617   1033.7   55.157
- PIEDS          1     3.527   1037.9   55.157
+ LATHO:POIDS    1     0.458   1033.9   55.311
- PERCHEV        1     7.993   1042.3   59.451
- LOMAIN         1     8.365   1042.7   59.808
- LATHO          1    20.198   1054.5   71.092
- PERPOIGN       1    20.971   1055.3   71.825
- STAT:POIDS     1   130.617   1165.0  170.675
```

Exercice

Summary(aicreg)

```
Call:
lm(formula = AGE ~ BIP + LATHO + LOMAIN + PERPOIGN + STAT + POIDS + PIEDS + PERCHEV + STAT:POIDS)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1695 -0.6894 -0.0093  0.6243  3.5885

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.702e+01  8.941e-01 -19.031  < 2e-16 ***
BIP            8.700e-02  5.135e-02   1.694  0.09056 .
LATHO          1.103e-01  2.509e-02   4.397 1.22e-05 ***
LOMAIN         1.593e-01  5.630e-02   2.830  0.00475 **
PERPOIGN      -1.999e-01  4.462e-02  -4.480 8.33e-06 ***
STAT           1.842e-01  7.340e-03  25.093  < 2e-16 ***
POIDS          2.318e-01  2.355e-02   9.844  < 2e-16 ***
PIEDS         -4.588e-02  2.497e-02  -1.837  0.06646 .
PERCHEV       -6.793e-02  2.456e-02  -2.766  0.00578 **
STAT:POIDS    -1.437e-03  1.285e-04 -11.181  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.022 on 990 degrees of freedom
Multiple R-squared:  0.9102,                    Adjusted R-squared:  0.9094
F-statistic:  1116 on 9 and 990 DF,  p-value: < 2.2e-16
```

Exercice

- Vérifier l'adéquation: $R^2 = 0.9102$

- Donner les intervalles de confiance des paramètres

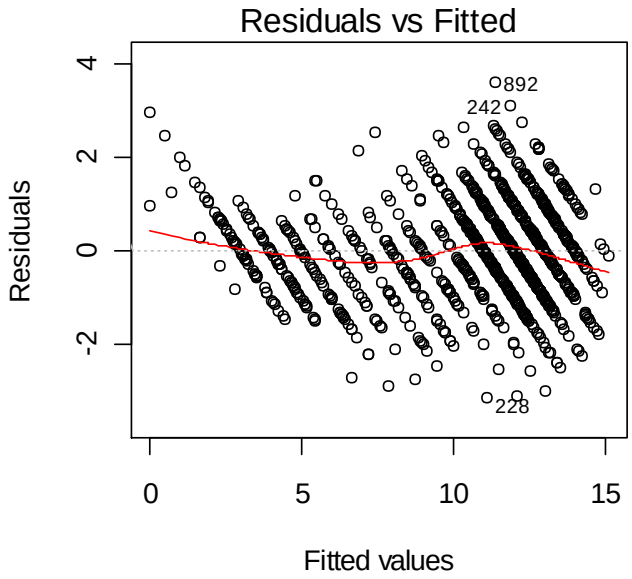
confint(regfin)

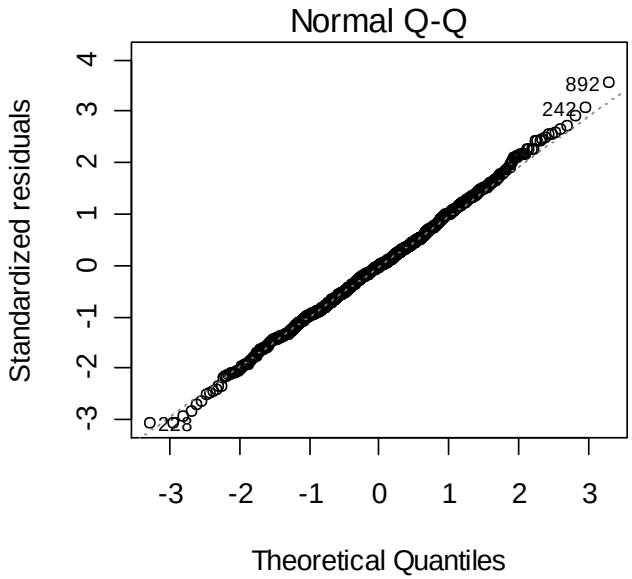
	2.5 %	97.5 %
(Intercept)	-18.770033112	-15.260926614
BIP	-0.013775579	0.187766190
LATHO	0.061077798	0.159545050
LOMAIN	0.048827123	0.269795052
PERPOIGN	-0.287443872	-0.112336511
STAT	0.169774355	0.198581248
POIDS	0.185581132	0.277997783
PIEDS	-0.094891805	0.003123193
PERCHEV	-0.116121184	-0.019734497
STAT:POIDS	-0.001689541	-0.001185036

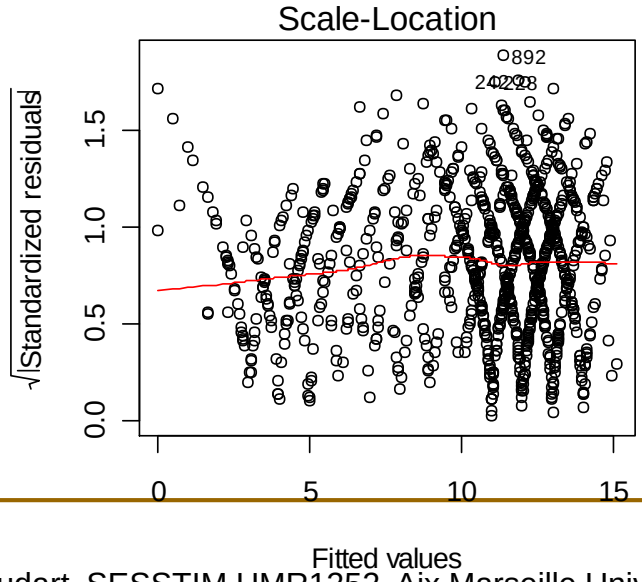
- Vérifier les conditions d'application

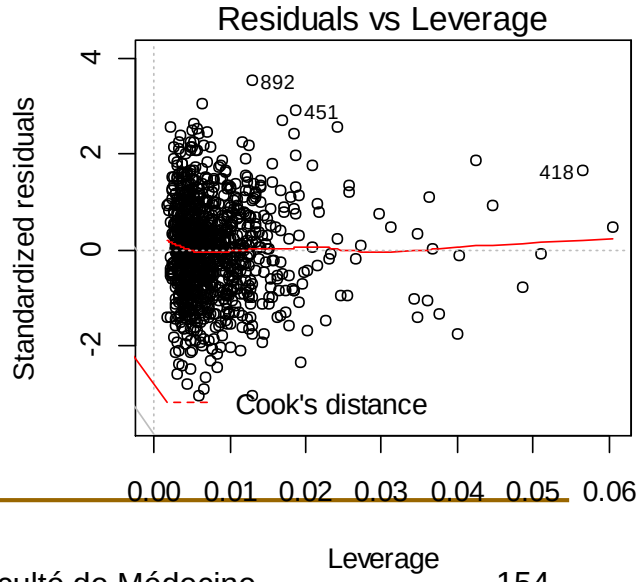
Exercice

par(mfrow=c(2,2))
plot(aicreg)









Références

- J. Bouyer: *Méthodes statistiques, Médecine-Biologie*, ed INSERM
- J. Bouyer: *Epidémiologie quantitative,* ed INSERM
- CIMES: *Biostatistiques*, ed Omnisciences
- JJ. Faraway: Linear Models with R, ed Chapman&Hall

Contact

jean.gaudart@univ-amu.fr

http://sesstim.univ-amu.fr