Statistique descriptive

Ajustement linéaire et méthode des moindres carrés

Objectif

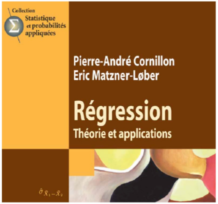
- Pour des variables X et Y quantitatives discrètes ou continues.
- Données bidimensionnelles $(x_i, y_i)_{i=1,\ldots,n}$.
- « Expliquer » Y en fonction de X : exprimer une dépendance fonctionnelle de Y comme fonction de X du type $y = f(x)$.
- f est choisie à partir de la forme du nuage de points.
  - Dépendance linéaire/affine : $y = ax + b$.
  - Dépendance polynômiale : $y = ax^2 + bx + c$.
  - Dépendance exponentielle, logarithmique, puissance non entière ...
- Relation non exacte.
- « Tendance » indiquée par le nuage de points.
- Prévision à partir de cette relation.

Un peu de vocabulaire

Ici : simple

- Expliquer Y en fonction de X.
- Y variable à expliquer : celle dont on cherche à expliquer les variations.
- X variable explicative.
- Régression linéaire : ajustement linéaire.

Bibliographie : Pierre-André Cornillon, Eric Matzner-Lober



Premiers indicateurs d'une relation linéaire

- Nuage de points avec une forme allongée.
- Coefficient de corrélation de Pearson proche de 1 ou -1. On rappelle la définition du coefficient de corrélation

$$r(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \sigma_y}.$$

Exemple: FCM

- La fréquence cardiaque maximale (FCM) est le rythme que le cœur humain d'une personne donnée atteint lors des plus fortes sollicitations (exprimée en battements par minute).
- Etude de la relation éventuelle entre l'âge d'un individu X et sa fréquence cardiaque maximale Y à partir des mesures de X et Y pour 13 individus (test d'effort).

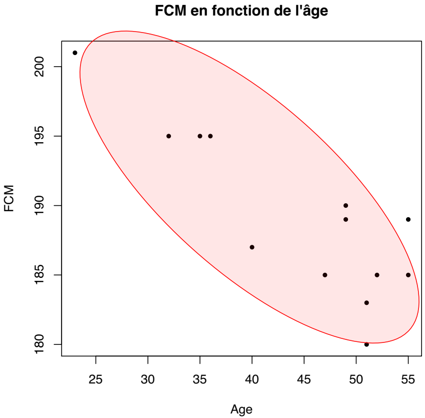
√ $\bar{x} = \frac{1}{n} \sum_{i=1}^{13} x_i$

√ $\bar{y} = \frac{1}{n} \sum_{i=1}^{13} y_i$

Individu i	Age x_i	FCM y_i
1	40	187
2	36	195
3	51	180
4	49	190
5	47	185
6	51	183
7	32	195
8	55	185
9	55	189
10	23	201
11	49	189
12	52	185
13	35	195

Exemple: FCM

- $\overline{x} = 44.23, \sigma_x = 10.03, \overline{y} = 189.15, \sigma_y = 5.93.$
- $\mathrm{Cov}(X, Y) = -50.54, r(X, Y) = -0.85.$



Méthode des moindres carrés: droite de régression

- Chercher la droite d'équation $y = ax + b$ la plus « proche » des n points (x_i, y_i).
- a et b constantes inconnues à préciser à partir des données
- Relation non exacte, points du nuage non alignés sur la droite: $y_i \neq ax_i + b$
- Erreur $e_i = y_i - (ax_i + b)$ pour tout $i = 1, \ldots, n$
- Erreur : tout ce qui n'est pas pris en compte dans la relation pour expliquer Y en fonction de X
- Modèle : $y_i = ax_i + b + e_i$

Méthode des moindres carrés

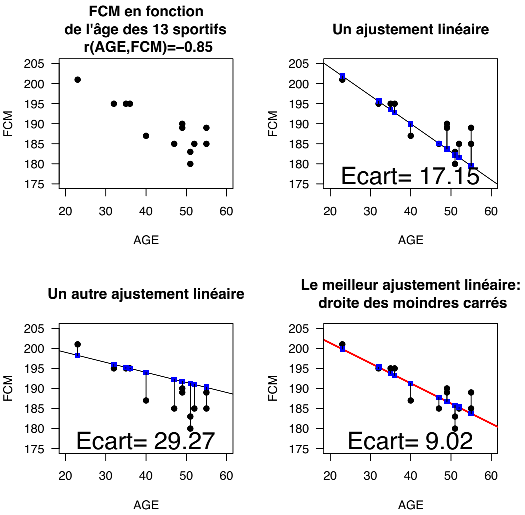
Trouver les valeurs $\widehat{a}$ et $\widehat{b}$ qui minimisent la fonction $\varphi(a,b)$ représentant la sommes des carrés des erreurs :

$$\varphi(a,b) = \sum_{i=1}^{n} (y_i - a x_i - b)^2 = \sum_{i=1}^{n} e_i^2 .$$

sont connus

inconnus

Exemple



FCM en fonction
de l'âge des 13 sportifs
r(AGE,FCM)=−0.85
Un ajustement linéaire
Ecart= 17.15
Un autre ajustement linéaire
Ecart= 29.27
Le meilleur ajustement linéaire:
droite des moindres carrés
Ecart= 9.02
FCM
AGE
205
200
195
190
185
180
175
20
30
40
50
60

Calcul des coefficients de la droite de régression

- $\varphi(a, b)$: fonction des deux variables a et b.
- Les valeurs $\widehat{a}$ et $\widehat{b}$ de a et b minimisant $\varphi(a, b)$ annulent les dérivées partielles de $\varphi(a, b)$ par rapport à a et à b.
- $\widehat{a}$ et $\widehat{b}$ solutions du système : $\left\{ \begin{array}{rcl} \dfrac{\partial\varphi(a,b)}{\partial a} & = & 0 \\ \dfrac{\partial\varphi(a,b)}{\partial b} & = & 0 \end{array} \right.$.

Calcul des coefficients de la droite de régression

$\varphi(a,b) = \frac{1}{n}\sum_{i=1}^{n}(y_i - ax_i - b)^2$

- On calcule les dérivées partielles

$$\frac{\partial\varphi(a,b)}{\partial a} = \sum_{i=1}^{n} -2x_i(y_i - ax_i - b)$$
$$\frac{\partial\varphi(a,b)}{\partial b} = \sum_{i=1}^{n} -2(y_i - ax_i - b)$$

- On doit donc résoudre le système

$$\sum_{i=1}^{n} -x_i(y_i - ax_i - b) = 0$$
$$\sum_{i=1}^{n} -(y_i - ax_i - b) = 0$$

Calcul des coefficients de la droite de régression

- Soit encore

$$\begin{aligned} \frac{a}{n}\sum_{i=1}^{n} x_i^2 + b\overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i \\ a\overline{x} + b &= \overline{y} \end{aligned}$$

- En multipliant la seconde égalité par $\overline{x}$, et en soustrayant, on trouve que la solution $\widehat{a}$ vérifie

$$\widehat{a}\,\mathrm{Var}(X) = \mathrm{Cov}(X, Y)$$

et finalement

$$\boxed{\widehat{a} = \frac{\mathrm{Cov}(X,Y)}{Var(X)},\ \widehat{b} = \overline{y} - \widehat{a}\overline{x}.}$$

Droite de régression

$$\boxed{\widehat{a} = \frac{\mathrm{Cov}(X,Y)}{Var(X)}, \;\; \widehat{b} = \overline{y} - \widehat{a}\overline{x}.}$$

- Droite $y = \widehat{a}x + \widehat{b}$: droite de régression D de Y en X (ou droite des moindres carrés ou droite d'ajustement linéaire de Y en X).
- La droite passe par le point moyen $(\overline{x}, \overline{y})$: $\bar{y} = \widehat{a}\overline{x} + \widehat{b}$.
- $\widehat{a}$ représente la pente de la droite et $\widehat{b}$ l'ordonnée à l'origine (intercept en anglais).

intercept

- $\widehat{a} = -0.5019$, $\widehat{b} = 211.354$.
- Pour une augmentation de 1 an de l'âge, la FCM a tendance à diminuer de -0.5019.

Une deuxième droite de régression

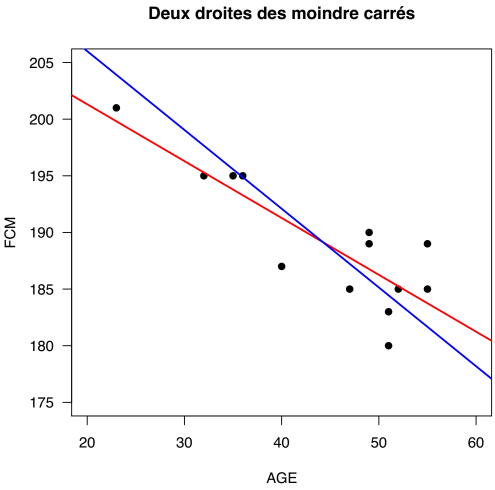
- Expliquer X en fonction de Y.
- Critère des moindres carrés :
  minimiser $\varphi(a_1, b_1) = \sum_{i=1}^{n}(x_i - a_1 y_i - b_1)^2$. Solution

$$\boxed{\begin{aligned} \widehat{a_1} &= \frac{\mathrm{Cov}(X,Y)}{Var(Y)} \\ \widehat{b_1} &= \overline{x} - \widehat{a_1}\,\overline{y} \end{aligned}}$$

- $(\widehat{a_1}, \widehat{b_1})$ sont les coefficient de la droite de régression D_1 de X en fonction de Y.

Retour à l'exemple

$\widehat{a_1} = -1.438$, $\widehat{b_1} = 316.265$. Ci-dessous, nuage de point et D, D_1.



Droites des moindres carrés et coefficient de corrélation de Pearson

- D et D_1 sont en général distinctes.
- Le point d'intersection se situe au centre de gravité du nuage $(\overline{x}, \overline{y})$.
- L'angle entre D et D_1 mesure de la dépendance entre X et Y : plus cet angle est ouvert, moins la liaison est forte.
- $(r(X, Y))^2 = \widehat{a} \times \widehat{a_1}$.
- $D = D_1$ ssi liaison linéaire exacte entre X et Y.
- D et D_1 perpendiculaires ssi X et Y sont non corrélées (dans ce cas, la droite D est la droite d'équation $y = \overline{y}$ parallèle à l'axe des x).

Prévision avec la droite des moindres carrés

- Soit x^* une nouvelle valeur de x.
- La prévision de la valeur de y vaut : $\widehat{y}^* = \widehat{a}x^* + \widehat{b}$.
- La prévision n'a de sens que pour des x^* proches des $(x_i)_{i=1,\dots,n}$.
- Exemples :
  - Prévision de la FCM d'un homme âgé de 27 ans.
  - Prévision de votre FCM.

Validation du modèle et qualité de l'ajustement

- Questions :
  - Valider le modèle.
  - Juger de la qualité de l'ajustement réalisé pour faire de la prévision.
- Analyse des résidus.
- Coefficient de détermination.

Définition et étude des résidus

- On appelle valeur ajustée de la i-ième observation de la variable Y l'approximation $\widehat{y}_i = \widehat{a}x_i + \widehat{b}$. C'est la valeur prévue par la droite de régression.
- On appelle résidu $\widehat{e}_i$, l'erreur que l'on commet en approchant y_i par $\widehat{y}_i$: $\widehat{e}_i = y_i - \widehat{y}_i$.

modèle

ce que j'observe

Propriétés des résidus

- Les $(y_i)_{i=1,\dots,n}$ et les valeurs prévues $(\widehat{y_i})_{i=1,\dots,n}$ ont même moyenne $\bar{y}$. En effet $\overline{\widehat{y}} = \widehat{a}\overline{x} + \widehat{b} = \overline{y}$.
- Les résidus sont centrés : $\bar{\hat{e}} = 0$.
- $\mathrm{Var}(\widehat{e}) = \frac{1}{n}\sum_{i=1}^{n}(\widehat{e}_i - \bar{\hat{e}})^2 = \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2$.

MMC

- Les résidus ne sont pas corrélés avec X: $\mathrm{Cov}(X, \widehat{e}) = 0$.
- Les résidus ne sont pas corrélés avec les valeurs prédites: $\mathrm{Cov}(\widehat{Y}, \widehat{e}) = 0$ (cela découle facilement de l'item précédent, car $\widehat{Y} = \widehat{a}X + \widehat{b}$).

Preuve de $\mathrm{Cov}(X, \widehat{e}) = 0$

On écrit

$$\begin{aligned}
\mathrm{Cov}(X, \widehat{e}) &= \mathrm{Cov}(X, Y - \widehat{Y}) \\
&= \mathrm{Cov}(X, Y) - \mathrm{Cov}(X, \widehat{Y}) \\
&= \mathrm{Cov}(X, Y) - \mathrm{Cov}(X, \widehat{a}X + \widehat{b}) \\
&= \mathrm{Cov}(X, Y) - \widehat{a}\,\mathrm{Var}(X) = 0\,.
\end{aligned}$$

Modèle linéaire avec R : lecture des données

```
fish <- read.table(file = "fish_linsimple.csv",
                   sep =";", header = TRUE)
kable(fish[1:10,])
```

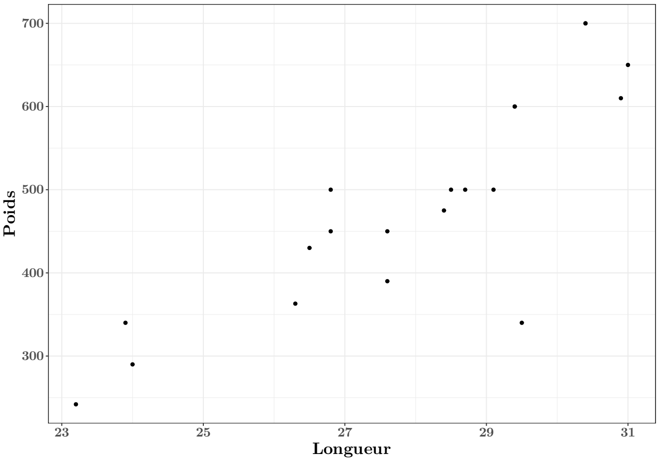
Poids	Longueur
242	23.2
290	24.0
340	23.9
363	26.3
430	26.5
450	26.8
500	26.8
390	27.6
450	27.6
500	28.5

Modèle linéaire avec R : représentation des données

```
summary(fish)
```

```
     Poids          Longueur    
 Min.   :242.0   Min.   :23.20  
 1st Qu.:383.2   1st Qu.:26.73  
 Median :487.5   Median :28.45  
 Mean   :481.5   Mean   :27.92  
 3rd Qu.:600.0   3rd Qu.:29.43  
 Max.   :700.0   Max.   :31.00  
```

Modèle linéaire avec R : représentation des données



Modèle linéaire avec R

```r
modele_reg_simple <- lm(Poids~ Longueur, data = fish)
coef(modele_reg_simple)
```

```
(Intercept)    Longueur
 -876.48191    48.63832
```

valeur de a ? b ?

Modélisation des données avec R

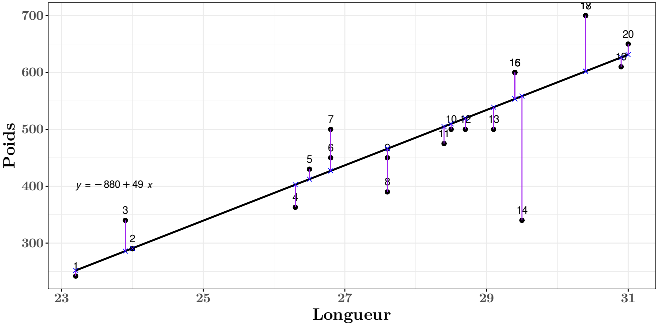
```
summary(modele_reg_simple)
```

```
Call:
lm(formula = Poids ~ Longueur, data = fish)

Residuals:
     Min       1Q   Median       3Q      Max
-218.349  -22.040   -5.274   46.515   97.877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -876.482    198.380  -4.418 0.000332 ***
Longueur      48.638      7.082   6.868    2e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 71.27 on 18 degrees of freedom
Multiple R-squared:  0.7238,	Adjusted R-squared:  0.7084
F-statistic: 47.16 on 1 and 18 DF,  p-value: 2.004e-06
```

Résidus observés

```
summary(modele_reg_simple)$residuals %>% head()
```

```
         1          2          3          4          5          6
-9.9271269 -0.8377834 54.0260487 -39.7059207 17.5664152 22.9749190
```



Fonction summary

On peut récupérer les infos:

```
Coef <- summary(modele_reg_simple)$coefficients
class(Coef)
```

```
[1] "matrix" "array"
```

```
Coef
```

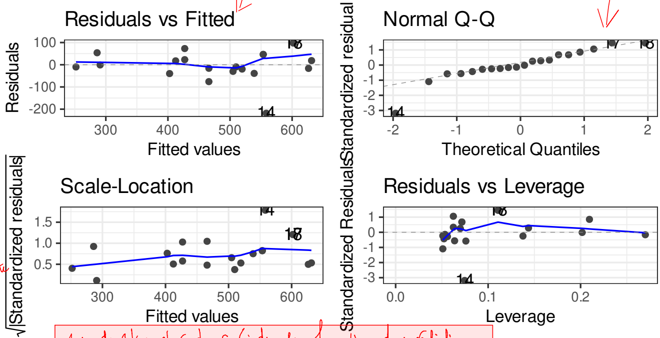
```
             Estimate Std. Error   t value     Pr(>|t|)
(Intercept) -876.48191 198.379689 -4.418204 3.318341e-04
Longueur      48.63832   7.082325  6.867565 2.004153e-06
```

4 graphes

structure aléatoire des Résidus

E(ε) = 0 validé

qq-plot validation de la normalité des Résidus points alignés selon droite Henry



valeur var constante

pas de points aberrants ou influents

pas de structure des résidus en fonction des prédictions

Donc on valide les hypothèses du modèle pour notre exemple.

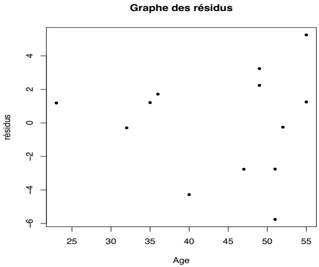
On peut maintenant tester la pertinence du modèle.

Exemple

Individu i	Age x_i	FCM y_i	$\widehat{y_i} = \widehat{a}x_i + \widehat{b}$	$\widehat{e_i} = y_i - \widehat{y_i}$
1	40	187	191.28	-4.28
2	36	195	193.28	1.72
3	51	180	185.76	-5.76
4	49	190	186.76	3.24
5	47	185	187.76	-2.76
6	51	183	185.76	-2.76
7	32	195	195.29	-0.29
8	55	185	183.75	1.25
9	55	189	183.75	5.25
10	23	201	199.81	1.19
11	49	189	186.76	2.24
12	52	185	185.25	-0.25
13	35	195	193.79	1.21

Graphes des résidus : Nuage de points $(x_i, \widehat{e}_i)_{i=1,\ldots,n}$.

- Modèle non adapté s'il y a une structure particulière dans le nuage des résidus.
- Il reste alors une information dans les résidus que la relation proposée entre y et x ne prend pas en compte.



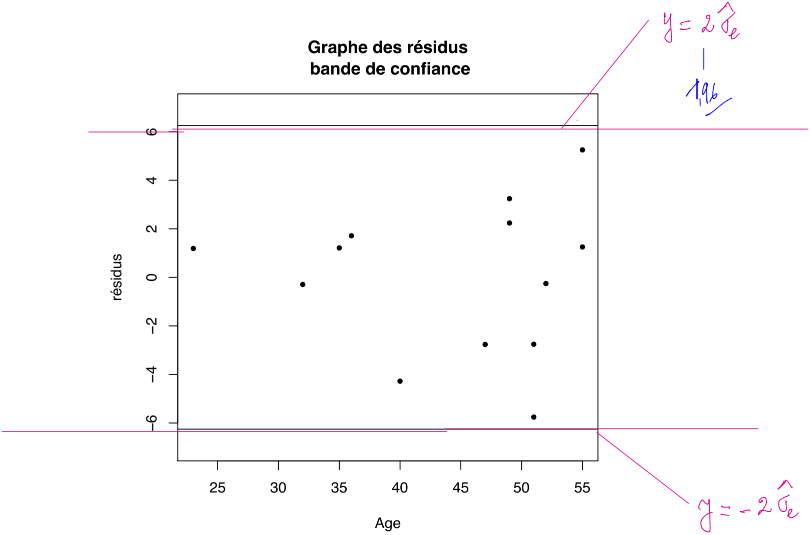
Lorsque le nombre d'observation est assez grand, ont peut aussi tracer la boîte à moustache ou l'histogramme des résidus.

Points du nuage mal expliqués par la relation linéaire

Écart-type des Résidus

- Localisation des points du nuage mal expliqués par la relation linéaire avec graphe des résidus.
- Tracé des deux droites parallèles à l'axe des abscisses d'équations $y = 2\sigma_{\widehat{e}}$ et $y = -2\sigma_{\widehat{e}}$.
- Points mal expliqués par le modèle : ceux situés hors de la bande déterminée par ces deux droites.
- Modèle mal choisi ou mal spécifié si beaucoup de points mal expliqués.
- Qualité de l'ajustement: on mesure l'erreur moyenne qu'on commet par l'écart-type des résidus:
  Erreur « moyenne » : $\sigma_{\widehat{e}} = \sigma_{Y-\widehat{Y}}$

Exemple


Graphe des résidus
bande de confiance
résidus
Age
6
4
2
0
−2
−4
−6
25
30
35
40
45
50
55
y = 2 σ̂e
1,96
y = -2 σ̂e

Décomposition de la variance :

$$
\begin{aligned}
\mathrm{Var}(Y) &= \mathrm{Var}(\widehat{Y}) + \mathrm{Var}(\widehat{e}) \\
\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 &= \frac{1}{n}\sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \frac{1}{n}\sum_{i=1}^{n}\widehat{e}_i^2
\end{aligned}
$$

Décomposition de la variance de la série observée en :

- Variance de la série ajustée $\widehat{Y}$: variance *expliquée* par le modèle.
- Variance de la série des résidus : variance *résiduelle* ou *non expliquée* par le modèle.

Preuve de la décomposition de la variance

On part de l'égalité élémentaire $Y = \widehat{Y} + \widehat{e}$. On en déduit que

$$\mathrm{Var}(Y) = \mathrm{Var}(\widehat{Y}) + 2\mathrm{Cov}(\widehat{Y}, \widehat{e}) + \mathrm{Var}(\widehat{e})\,.$$

Or, on a vu que $\mathrm{Cov}(\widehat{Y}, \widehat{e}) = 0$, et par conséquent,

$$\mathrm{Var}(Y) = \mathrm{Var}(\widehat{Y}) + \mathrm{Var}(\widehat{e})\,.$$

Coefficient de détermination

- Part de la variance de Y expliquée par la régression linéaire :

$$\frac{\mathrm{Var}(\widehat{Y})}{\mathrm{Var}(Y)}.$$

- Coefficient de détermination :

$$R^2 = \frac{\mathrm{Var}(\widehat{Y})}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(\widehat{e})}{\mathrm{Var}(Y)}.$$

Coefficient de détermination

- $0 \leq R^2 \leq 1$.
- Si $R^2 = 0$, la variable X n'a aucune influence linéaire sur Y et $\widehat{y}_i = \overline{y}$ pour tout i.
- Si $R^2 = 1$: $y_i = \widehat{y}_i = \widehat{a}x_i + \widehat{b}$ pour tout i. Relation linéaire exacte entre Y et X.

Coefficient de détermination et coefficient de corrélation de Pearson

Dans le cas de la régression linéaire : $R^2 = (r(X,Y))^2$.
En effet, on a que

$$R^2 = \frac{\mathrm{Var}(\widehat{Y})}{\mathrm{Var}(Y)} = \frac{\widehat{a}^2\mathrm{Var}(X)}{\mathrm{Var}(Y)} = \frac{(\mathrm{Cov}(X,Y))^2}{\mathrm{Var}(X)\mathrm{Var}(Y)} = (r(X,Y))^2\,.$$

Exemple

- $\mathrm{Var}(Y) = 35,14$ et $\mathrm{Var}(\widehat{e}) = 9,77$
- $\mathrm{Var}(\widehat{Y}) = 35,14 - 9,77 = 25,37$
- $R^2 = 25,37/35,14 = 0,7218$
- La part de variance expliquée par la relation linéaire $y = -0,5019\ x + 211,354$ est de 72,18%.

Généralisation: ajustement polynomial

- Lorsque le nuage de point ne présente pas une forme allongée le long d'une droite, mais qu'une relation semble exister entre X et Y, il peut être judicieux d'expliquer Y par un polynôme de degré p en X (p étant un entier strictement positif).
- Idée: si la relation entre Y et X est complexe, on peut s'approcher de la « vraie » relation en augmentant le degré du polynôme (en accord avec la théorie de l'approximation).
- Toutefois, il ne faut pas non plus que p soit trop grand pour éviter « l'overfitting » (et obtenir une courbe qui passe par beaucoup de points du nuage sans refléter la tendance).

Ajustement polynômial par moindres carrés

- Pour ajuster un polynôme de degré p par la méthodes des moindres carrés, on cherche à minimiser

$$\boxed{\varphi(a_0, a_1, \ldots, a_p) = \sum_{i=1}^{n} (y_i - a_0 - a_1 x_i - \cdots - a_p x_i^p)^2 \, .}$$

- On sait montrer que cette fonction φ admet un unique minimum (grâce au théorème de Pythagore dans $\mathbb{R}^n$).
- De plus, ce minimum est atteint pour un seul jeu de paramètres $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$. pourvu que les $p+1$ vecteurs de $\mathbb{R}^n$: $(1, \ldots, 1)$, $(x_1, \ldots, x_n)$, ..., $(x_1^p, \ldots, x_n^p)$ forment une famille de vecteurs linéairement indépendants.
- Si c'est le cas, on dispose d'une formule explicite pour $(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_p)$, faisant intervenir $(y_1, \ldots, y_n)$ ainsi que la matrice composée des p vecteurs décrits ci-dessus.

On peut montrer en introduisant les notations

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \text{et} \quad M = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ 1 & x_2 & x_2^2 & \dots & x_2^p \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^p \end{pmatrix}$$

que la solution (bien connue) de ce problème de minimisation est :

$$\boxed{(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_p)' \;=\; (M'M)^{-1}M'Y}$$

où M' désigne la matrice transposée de M et :

$$M'M = \begin{pmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^p \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{p+1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum x_i^p & \sum x_i^{p+1} & \sum x_i^{p+2} & \dots & \sum x_i^{2p} \end{pmatrix}, M'Y = \begin{pmatrix} \sum y_i \\ \sum x_i y_i \\ \vdots \\ \sum x_i^p y_i \end{pmatrix}.$$

En particulier, on peut vérifier que :

- Lorsque $p = 0$, alors $M'M = n$ et $M'Y = \sum_{i=1}^{n} y_i$, et donc $\widehat{a}_0 = \overline{y}$.
- Lorsque $p = 1$, alors :

$$M'M = \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix} \quad \text{et} \quad M'Y = \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

D'où :

$$\begin{pmatrix} \widehat{a}_0 \\ \widehat{a}_1 \end{pmatrix} = \begin{pmatrix} \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2} \\ \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2} \end{pmatrix} = \begin{pmatrix} \frac{\overline{y} \sum x_i^2 - \overline{x} \sum x_i y_i}{\sum x_i^2 - n\overline{x}^2} \\ \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2} \end{pmatrix} = \begin{pmatrix} \overline{y} - \widehat{a}_1 \overline{x} \\ \frac{\sum (x_i - \overline{x})(y_i - \overline{y})}{\sum (x_i - \overline{x})^2} \end{pmatrix}$$

Ajustement polynômial par moindres carrés

- On peut ensuite définir les valeurs ajustées $\hat{y}_i = \hat{a}_0 + \hat{a}_1 x_i + \cdots + \hat{a}_p x_i^p$ et les résidus $\hat{e}_i = y_i - \hat{y}_i$.
- Les résidus sont centrés: $\overline{\hat{e}} = 0$.
- Les résidus ne sont pas corrélés avec les valeurs prédites: $\mathrm{Cov}(\widehat{Y}, \widehat{e}) = 0$.
- Par conséquent, la formule de décomposition de la variance est encore valable: $\mathrm{Var}(Y) = \mathrm{Var}(\hat{Y}) + \mathrm{Var}(\hat{e})$
- Le coefficient de détermination s'écrit donc toujours

$$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} .$$

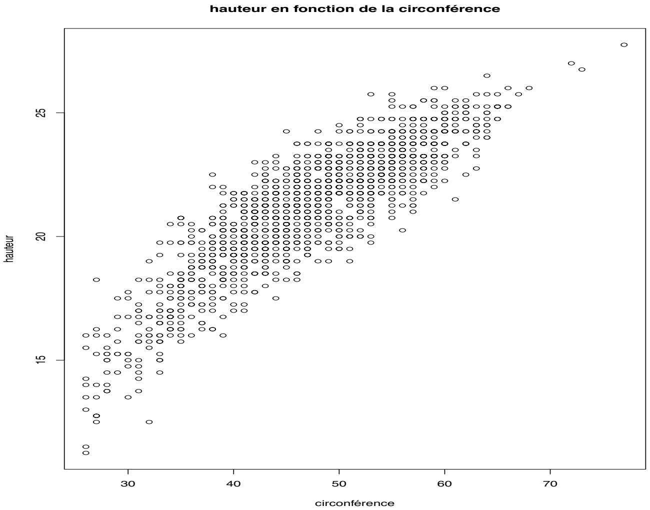
C'est la part de la variance de Y expliquée par la relation polynomiale en X.

- Attention : si $p > 1$ l'égalité $R^2 = (r(X,Y))^2$ n'est plus vraie. On peut montrer que le R^2 augmente toujours avec le degré du polynôme.

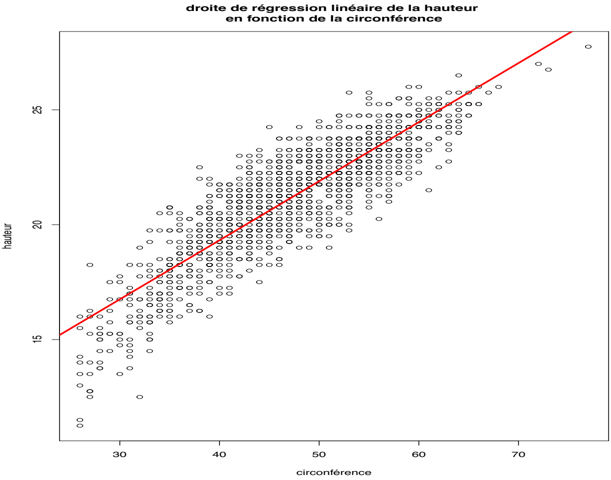
Necéssité de l'ajustement d'un polynôme: un exemple

- On souhaite pouvoir prédire la hauteur des eucalyptus (quantité difficile à mesurer, variable à expliquer) à partir de la mesure de la circonférence du tronc à 1m30 du sol (variable explicative).
- Pour cela, on a mesuré la hauteur (en m) et la circonférence (en cm) de 1429 eucalyptus. On note Y la variable hauteur, et X la variable circonférence.
- Données fournies par le CIRAD (centre international de recherche agronomique pour le développement).

Necéssité de l'ajustement d'un polynôme: un exemple

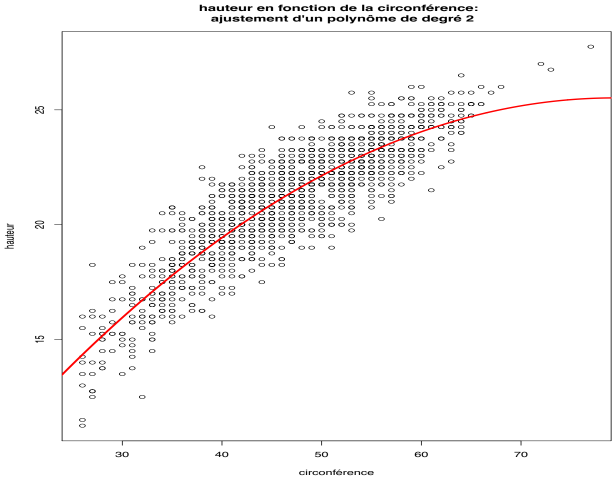


Ajustement d'un polynôme de degré 1: droite de régression

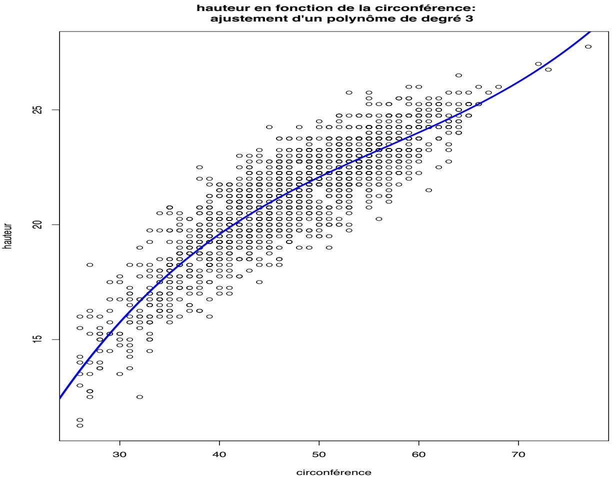


$\hat{a} = 0.257, \hat{b} = 9.037, R^2 = 0.768.$

Ajustement d'un polynôme de degré 2

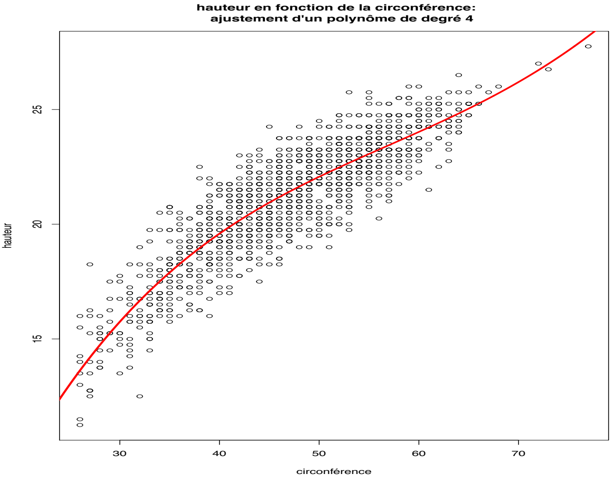


$$\hat{a}_0 = 0.803, \hat{a}_1 = 0.623, \hat{a}_2 = -0.004, R^2 = 0.789.$$



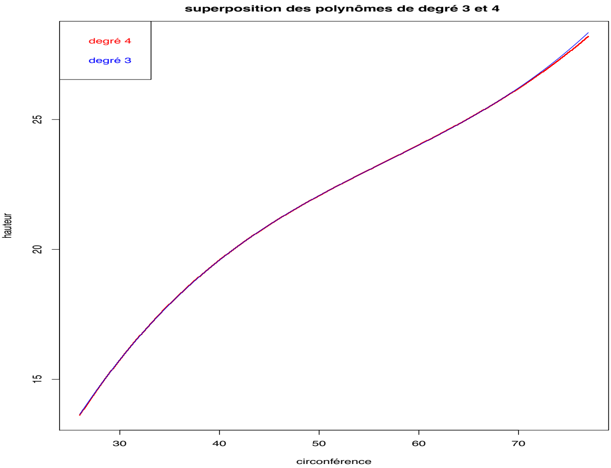
$$\hat{a}_0 = -11.92,\ \hat{a}_1 = 1.49,\ \hat{a}_2 = -0.023,\ \hat{a}_3 = 0.00013,\ R^2 = 0.7943.$$

Ajustement d'un polynôme de degré 4



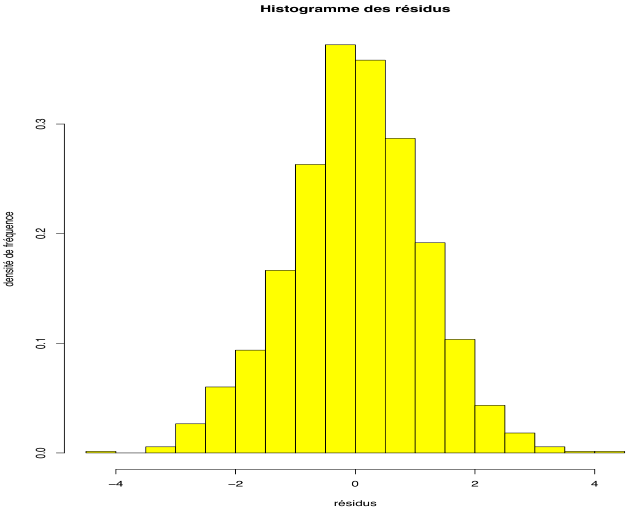
$R^2 = 0.7943.$

Comparaison des ajustements (degré 3 et 4)

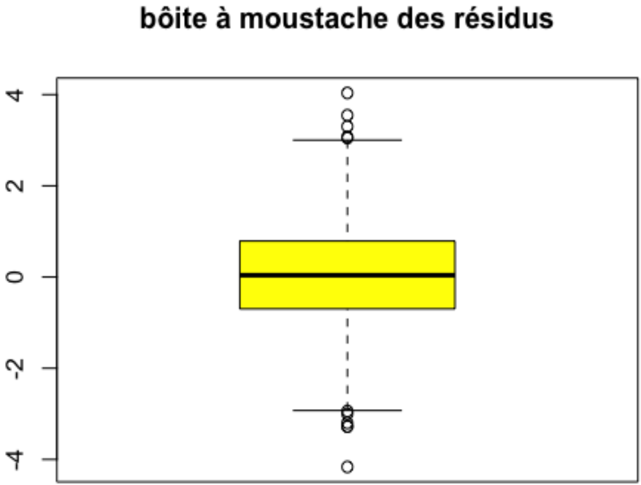


$R^2 = R^2 = 0.7943$

Histogramme des résidus (polynôme de degré 3)



Boîte à moustache des résidus (polynôme de degré 3)



Prévision de la hauteur (polynôme de degré 3)

Rappel

$\hat{a}_0 = -11.92, \hat{a}_1 = 1.49, \hat{a}_2 = -0.023, \hat{a}_3 = 0.00013, R^2$

- La part de variance de Y expliquée par la relation polynomiale de degré 3 est de 79.4%.
- Pour une circonférence de 30cm, on prévoit une hauteur de

$$\underbrace{-11.92}_{\hat{a}_0} + \underbrace{1.49}_{\hat{a}_1}x30 - \underbrace{0.023}_{\hat{a}_2}x30^2 + \underbrace{0.00013}_{\hat{a}_3}x30^3 = 15.59m$$

- Pour une circonférence de 60cm, on prévoit une hauteur de

$$-11.92 + 1.49x60 - 0.023x60^2 + 0.00013x60^3 = 22.76m$$

En général, le choix du degré du polynôme est basé sur l'examen :

- Du graphe de la série $(y_i)_{i=1,\dots,n}$, qui permet de se faire une première idée du degré.
- Des résidus obtenus après l'ajustement.

  On voudra :
  - Obtenir des résidus qui fluctuent autour de 0 et de faible amplitude.
  - Utiliser un polynôme de degré le plus faible possible.
- Du critère AIC (dans le but de faire de la prévision).

Le critère AIC

- Le critère AIC (Akaike Information Criterion) fait partie des méthode dites « de pénalisation ».
- Le critère se calcule de la façon suivante (à des constantes près, qui ne dépendent pas de p). Notons SCR_p la somme des carrés des résidus de la régression polynomiale de degré p. Alors
$$AIC(p) = n \log(SCR_p) + 2p$$
- Sous certaines conditions, et si n est assez grand, on peut montrer que le degré pour lequel l'AIC est le plus faible donnera le « meilleur » ajustement pour la prévision.

Le critère AIC, suite

- Lorsque p augmente, R^2 augmente et donc SCR_p baisse. On rajoute le terme $2p$ qui augmente avec p (on « pénalise » les modèles avec beaucoup de paramètres).

- L'AIC(p) minimal choisit donc le degré pour lequel à la fois SCR_p et p ne sont pas trop grands: il s'agit d'un compromis pour trouver « le meilleur modèle pas trop grand ».

Exemple: choix du degré à partir de l'AIC

On reprend l'exemple des Eucalyptus, et on calcule l'AIC pour les degrés 1,2,3 et 4. On obtient le tableau suivant:

degré	1	2	3	4
R^2	76,8%	79%	79,4%	79,4%
AIC	4578.454	4440.56	4412.794	4414.737

Ici, le degré pour lequel l'AIC est minimal est le degré 3.

La part de variance de Y expliquée par l'ajustement polynomial de degré 3 est de $79,4\%$.