Chapitre 2. Régression linéaire Simple

Cours de modèle linéaire gaussien par S. Donnet

Executive Master Statistique et Big-Data

Juin 2021

Jeu de données women

- Jeu de données women : poids et âge de 15 femmes
- Expliquer lien entre variables : `poids=f(taille)`
- Prédire le poids d'un nouvel individu à partir de sa taille

```
str(women)
```

```
## 'data.frame':    15 obs. of  2 variables:
##  $ height: num  58 59 60 61 62 63 64 65 66 67 ...
##  $ weight: num  115 117 120 123 126 129 132 135 139 142 ...
```

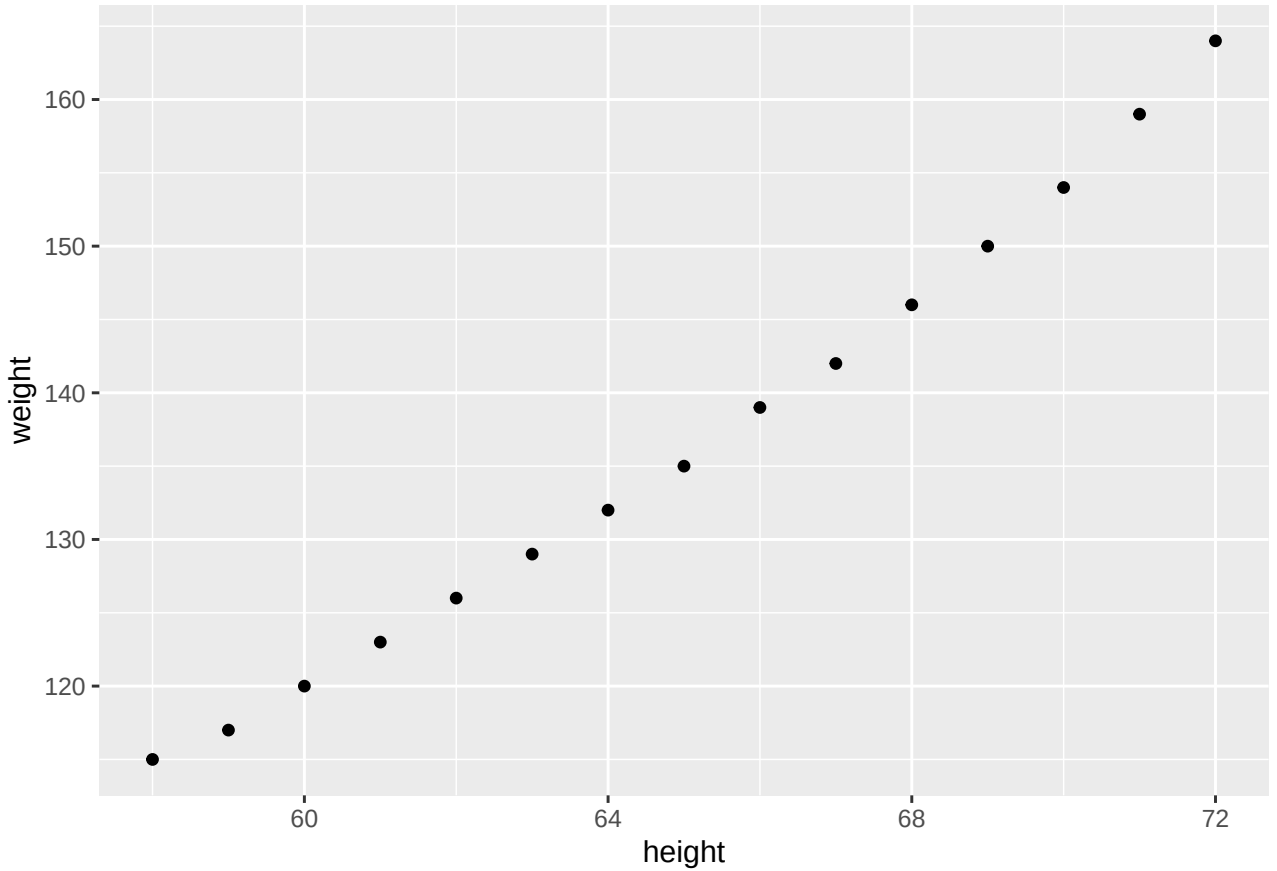
Jeu de données women

```
summary(women)
```

```
##      height          weight
##  Min.   :58.0   Min.   :115.0
##  1st Qu.:61.5   1st Qu.:124.5
##  Median :65.0   Median :135.0
##  Mean   :65.0   Mean   :136.7
##  3rd Qu.:68.5   3rd Qu.:148.0
##  Max.   :72.0   Max.   :164.0
```

Visualisation des données

```
library(ggplot2)
ggplot(women) + geom_point(aes(x = height, y = weight))
```



Relation linéaire

- D'après ce graphique, la relation est assez linéaire
- Si y_i=poids et x_i=taille de l'individu i, $1 \leq i \leq n = 15$, alors

$$\exists a, b | \forall i, \quad y_i \approx a x_i + b$$

- y=**variable à expliquer** et x=**variable explicative**

Démarche générale

- En statistique, on cherche à expliquer ou prédire, une variable d'intérêt y en fonction d'une autre variable x.
- On dipose de n valeurs $(x_1, \dots x_n)$ et $(y_1, \dots, y_n)$ pour *apprendre* la relation en x et y :

$$y \approx f(x)$$

- On ne sait pas *à l'avance*
  - si x explique correctement y (x est-elle en particulier suffisante pour prévoir y ?)
  - quelle est la forme de f : on essaie de la deviner en traçant le nuage (x_i, y_i) [1]

1. mais infaisable si plusieurs variables explicatives

Régression linéaire simple

Dans ce chapitre :

$$f(x) = a\,x + b$$

- Peu paraître naif
- Peut paraître simpliste (au regard des méthodes complexes vues plus tard)
- Parfois les modèles simples sont ceux qui fonctionnent le mieux
- Encore largement utilisée
- Méthodes plus sophistiquées = extensions du modèle linéaire

Fonction de perte (loss function)

Quantifier l'approximation $y \approx f(x)$
Fonction de perte ℓ (proximité de la droite au nuage de points)

$$\arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell\big(y_i - f(x_i)\big)$$

- Perte ou coût quadratique : $\ell(u) = u^2$
  Facilité de calcul, moindres carrés
- Perte ou coût absolu : $\ell(u) = |u|$
  Sensibilité moindre aux valeurs aberrantes, robustesse
- Dans les deux cas ℓ est positive, nulle en zéro, symétrique

Estimation par moindres carrés

Moindres carrés

On cherche a et b minimisant

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

Propriétés des estimations

Une fois a et b estimés sur NOS données dites d'apprentissage

- Que se serait-il passé si on avait pris un autre échantillon : grandes variations dans nos estimations ? Si oui, quelle validité donner à nos résultats ?
- *Pour répondre à ces questions*, démarche statistique :
  - y_i réalisation d'une variable aléatoire dont on spécifie (contrôle) la loi de probabilité,
  - Etudier les variations de notre estimation sous cette hypothèse.
  - Etudier les propriétés probabilistes de notre estimateur (biais, variance...)
  - Idée de l'incertitude que l'on a sur nos valeurs estimées de a et b.

Modélisation statistique

Pour tout i, on suppose que y_i est la réalisation de Y_i variable aléatoire telle que

$$Y_i = ax_i + b + \varepsilon_i, \quad i = 1, \ldots, n$$

où ε_i (bruit ou **erreur**) variables aléatoires.

A propos des ε_i

- D'espérance nulle [P1], de variance constante σ^2 [P2], indépendants [P3]

$$\mathbb{E}[\varepsilon_i] = 0 \quad \text{et} \quad \mathbb{V}[\varepsilon_i] = \sigma^2$$

- *Modèle linéaire gaussien*, ε_i gaussiens [P4]

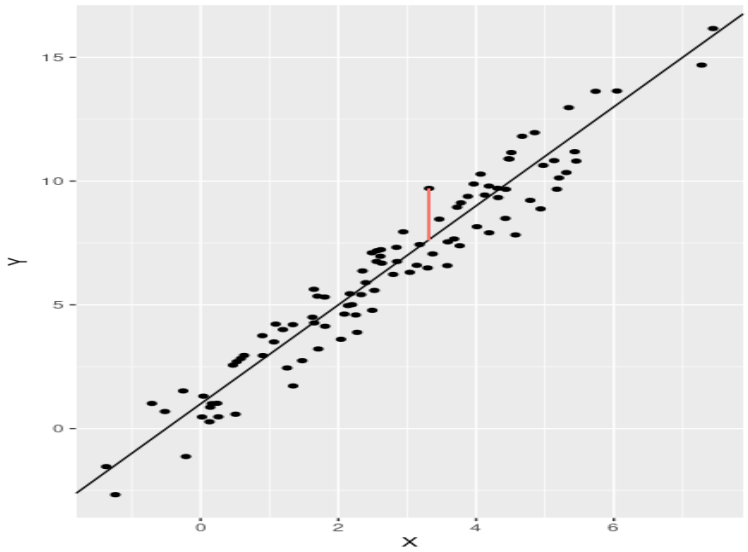
$$\varepsilon_i \sim_{i.i.d.} \mathcal{N}(0, \sigma^2)$$

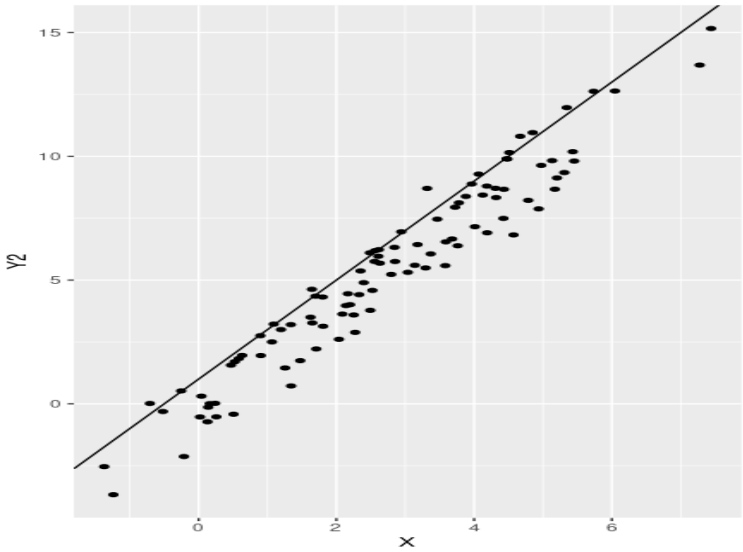
Modèle de régression linéaire simple, gaussien

A propos des postulats

$$\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$$

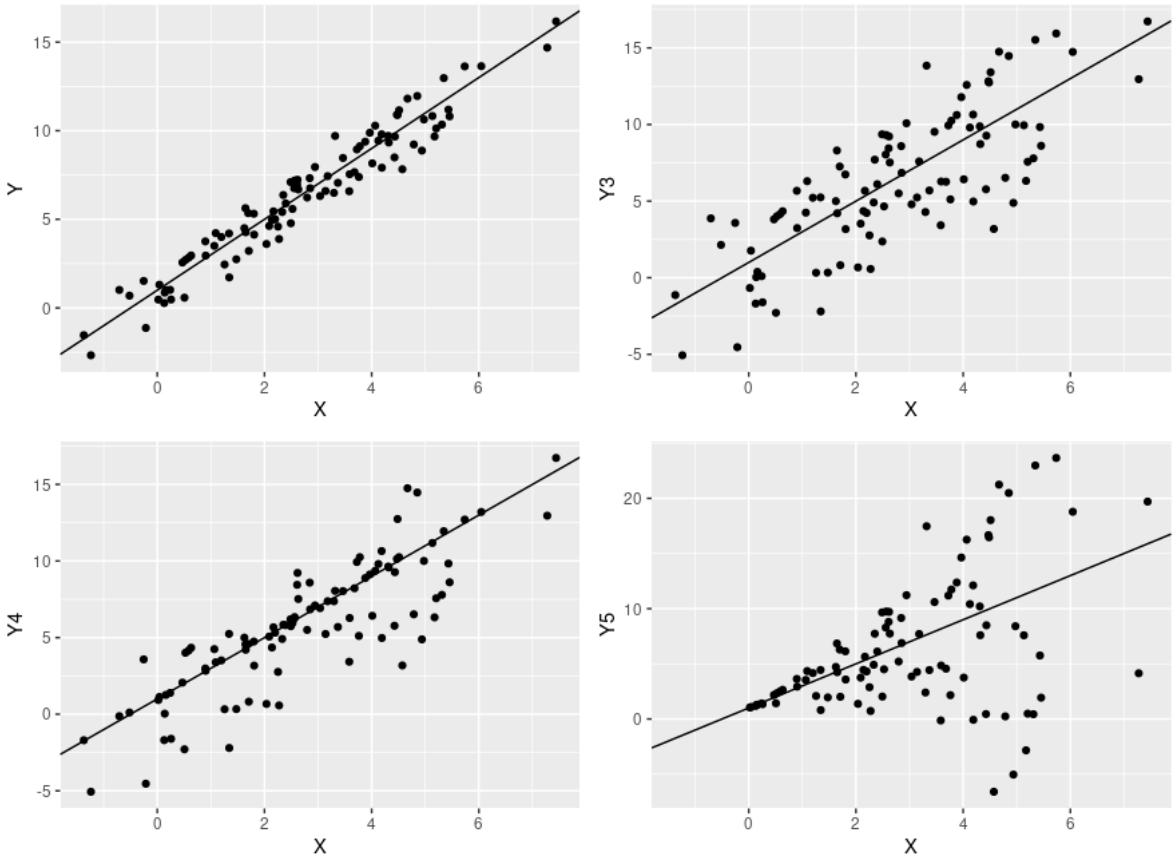
- [P1] $\mathbb{E}[\varepsilon] = 0_{\mathbb{R}}$: modèle est correct, on n'a pas oublié un terme pertinent





A propos des postulats

- [P2] $\mathbb{V}[\varepsilon_i] = \sigma^2, \forall i = 1 \dots n$ (*homoscédastique*, par opposition à *hétéroscédastique*)



ε de variance constante en haut, de variance variable en bas

A propos des postulats

- [P3] ε_i indépendants : les observations sont supposées indépendantes, i.e. correspondent à un échantillonnage indépendant ou aux résultats d'une expérience physique menée dans des conditions indépendantes.

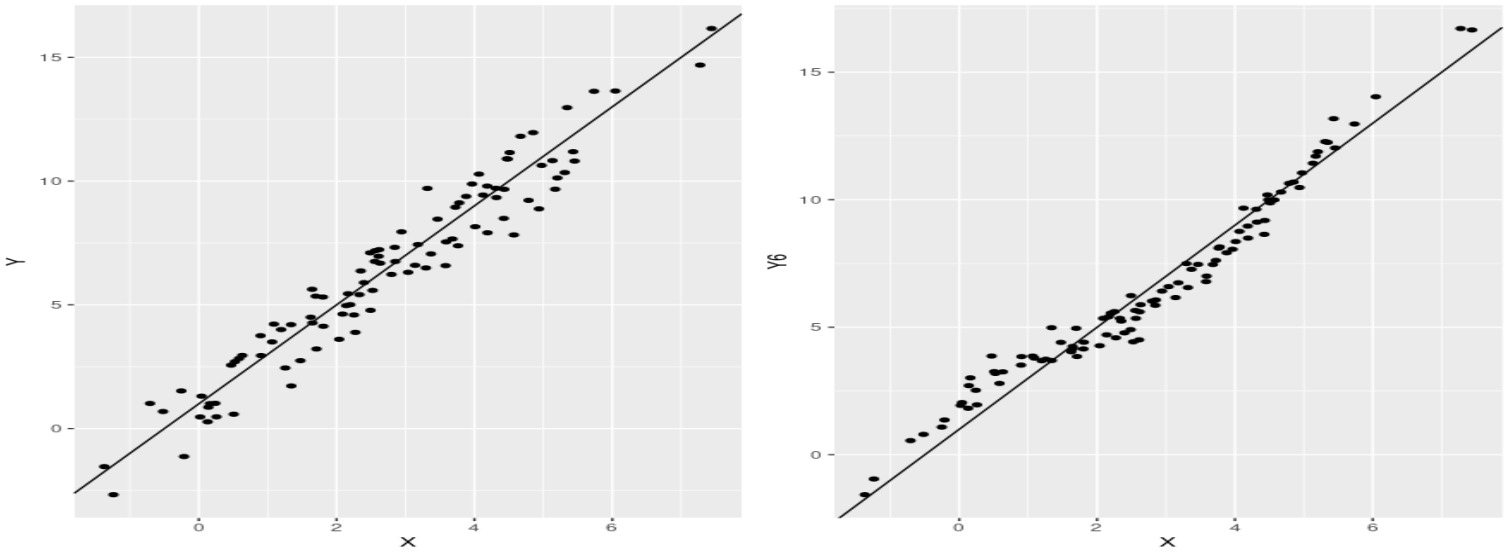


Figure – Erreurs indépendantes à gauche, dépendantes à droite

Propre aux données. Ne peut être remis en cause après l'inférence.

A propos des postulats

- [P4] ε_i gaussiennes.
  - Postulat est le moins important : on peut s'en passer si le nombre d'observations est grand (au delà de 20 ou 30 observations).
  - Difficile de détecter la non-gaussianité des erreurs.
  - **R** propose des outils graphiques pour tenter de valider ou non cette hypothèse.

Remarque

On parle de *postulats* en ce sens que nous ne pouvons pas formellement montrer qu'ils sont vérifiés par des tests statistiques : outils graphiques pour vérifier leur validité.

Définition de l'EMC

EMC = Estimateur des Moindres Carrés [2]

$$(\hat{a}, \hat{b}) = \underset{(a,b)\in\mathbb{R}\times\mathbb{R}}{\arg\min} \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

Ecriture matricielle

- $\boldsymbol{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} b \\ a \end{pmatrix}$
- $\begin{pmatrix} ax_1 + b \\ \vdots \\ ax_n + b \end{pmatrix} = X\boldsymbol{\beta}$
-

$$(\hat{a}, \hat{b}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^2}{\arg\min} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2$$

2. MCO = Moindres Carrés Ordinaires, (O)LS = (Ordinary) Least Squares

Calcul de l'estimation

Théorème (Calcul de l'estimation)

Si $\exists (i,j) : x_i \neq x_j$ *alors la solution du problème d'optimisation est*

$$\hat{a} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\sigma_{x,y}}{\sigma_x^2} \quad et \quad \hat{b} = \bar{y} - \hat{a}\bar{x}.$$

- Existence, unicité, calcul : démonstration au tableau ε_i
- Le vecteur $\hat{\beta} = \begin{pmatrix} \hat{b} \\ \hat{a} \end{pmatrix}$ est une combinaison linéaire des observations

$$\hat{a} = \sum_{i=1}^{n} \lambda_{i,1} y_i \quad \text{et} \quad \hat{b} = \sum_{i=1}^{n} \lambda_{i,2} y_i.$$

Exemple : hauteur des eucalyptus

```
euca<-read.table("eucalyptus.txt",header=T,sep=" ")
str(euca)
```

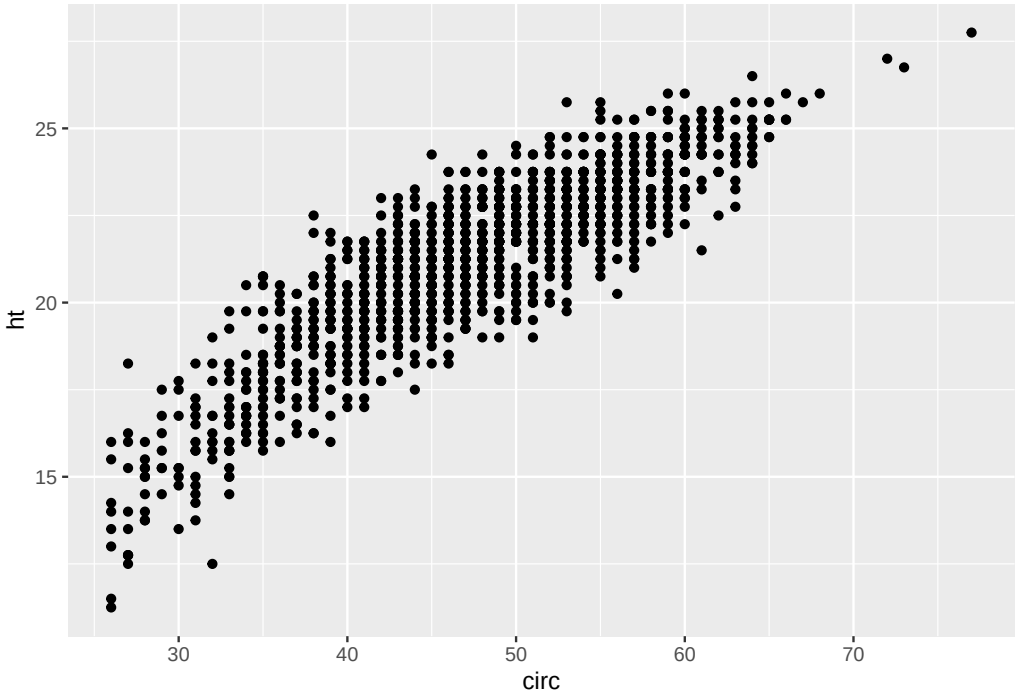
```
## 'data.frame':    1429 obs. of  3 variables:
##  $ ht   : num  18.2 19.8 16.5 18.2 19.5 ...
##  $ circ: int  36 42 33 39 43 34 37 41 27 30 ...
##  $ bloc: Factor w/ 3 levels "A1","A2","A3": 1 1 1 1 1 1 1 1 1
```

```
names(euca)
```

```
## [1] "ht"   "circ" "bloc"
```

Plot

- Une seule variable explicative
- Commencer par représenter graphiquement le nuage de points
- Les points sont disposés grossièrement le long d'une droite
- Une régression simple semble indiquée




```
gg_euca <- ggplot(euca, aes(x = circ, y = ht)) + geom_point()
gg_euca
```

Estimation des parametres avec R

```
reg <- lm(ht~circ,data=euca)
```

```
summary(reg)
```

Estimation des paramètres avec R

```
##
## Call:
## lm(formula = ht ~ circ, data = euca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.7659 -0.7802  0.0557  0.8271  3.6913
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.199 on 1427 degrees of freedom
## Multiple R-squared:  0.7683, Adjusted R-squared:  0.7682
## F-statistic:  4732 on 1 and 1427 DF,  p-value: < 2.2e-16
```

Exemple : hauteur des eucalyptus

Nous allons apprendre à interpréter toutes ces informations !

Exemple : hauteur des eucalyptus

- Liste des différents résultats de l'objet `reg` avec :

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

Exemple : hauteur des eucalyptus

- Liste des différents résultats de l'objet `reg` avec :

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

- $(\hat{b}, \hat{a})$: colonne `Estimate` du tableau `Coefficients` de la sortie de la fonction `summary`

Exemple : hauteur des eucalyptus

- Liste des différents résultats de l'objet `reg` avec :

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

- $(\hat{b}, \hat{a})$: colonne `Estimate` du tableau `Coefficients` de la sortie de la fonction `summary`
- La première ligne, `Intercept`, correspond au coefficient $\hat{b}$ et la seconde ligne au coefficient $\hat{a}$. C'est le vecteur $\hat{\beta}$!

Exemple : hauteur des eucalyptus

- Liste des différents résultats de l'objet `reg` avec :

```
names(reg)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

- $(\hat{b}, \hat{a})$: colonne `Estimate` du tableau `Coefficients` de la sortie de la fonction `summary`
- La première ligne, `Intercept`, correspond au coefficient $\hat{b}$ et la seconde ligne au coefficient $\hat{a}$. C'est le vecteur $\hat{\beta}$!
- On peut récupérer les coefficients avec la fonction `coef`

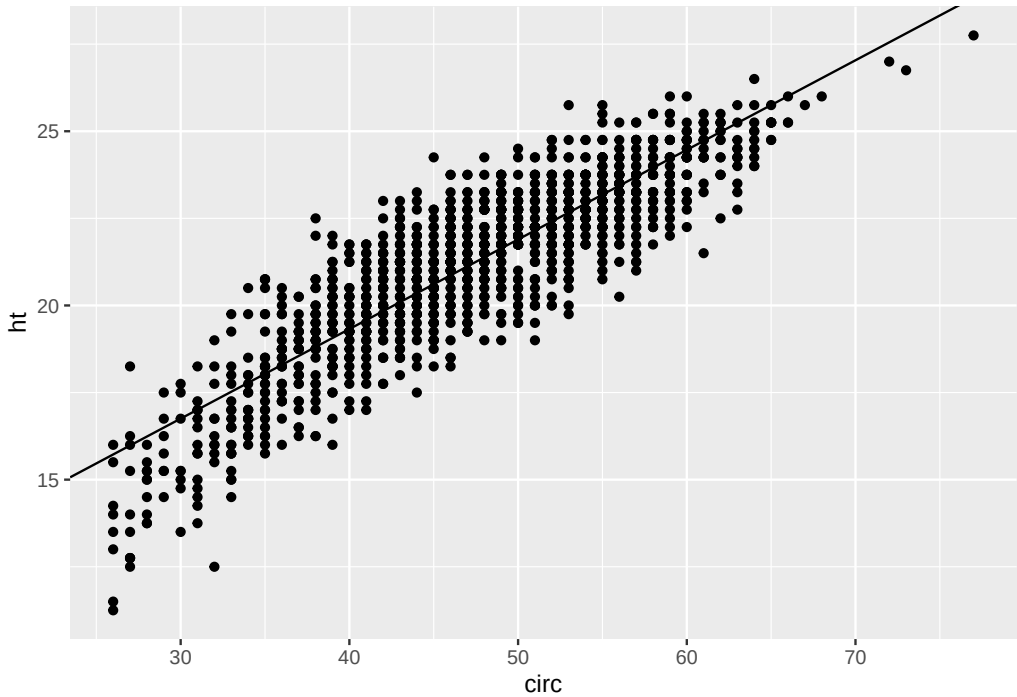
```
coef(reg)
```

```
## (Intercept)        circ
##   9.0374757   0.2571379
```

Exemple : hauteur des eucalyptus

Tracé de la droite de régression $y = \hat{b} + \hat{a}x$ sur le graphique du nuage

```
gg_euca + geom_abline(intercept=coef(reg)[1],
slope = coef(reg)[2])
```



De l'estimation à l'estimateur

- Doit-on avoir confiance en notre estimation ?
- Si on avait un autre échantillon, l'estimation aurait-elle beaucoup varié ?
- Etude de l'estimateur : on remplace y_i par Y_i et on étudie ses propriétés en tant que variable aléatoire.

$$\begin{aligned} \hat{A} &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{B} &= \bar{Y} - \hat{A}\bar{x} \\ \text{avec } \bar{Y} &= \frac{\sum_{i=1}^{n} Y_i}{n} \end{aligned}$$

Loi, espérance, variance ?

Loi de l'EMC

- Le paramètre et son EMC

$$\beta = \begin{pmatrix} b \\ a \end{pmatrix}, \quad \hat{\beta} = \begin{pmatrix} \hat{B} \\ \hat{A} \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{A}\bar{x} \\ \sigma_{x,Y}/\sigma_x^2 \end{pmatrix}$$

Théorème (Loi de l'EMC)

Si $Y_i = a + bx_i + \varepsilon_i$ *avec* $\varepsilon_i \sim_{i.i.d} \mathcal{N}(0, \sigma^2)$ *alors*

$$\hat{\beta} \sim \mathcal{N}_2(\beta, \sigma^2 V) \quad \text{où} \quad V = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{pmatrix}$$

Propriétés de l'EMC

- L'EMC $\hat{\beta}$ est **sans biais** : $\mathbb{E}(\hat{\beta}) = \beta$ i.e. $\mathbb{E}(\hat{A}) = a$ et $\mathbb{E}(\hat{B}) = b$.
- Ses composantes sont gaussiennes, de variance

$$\mathrm{Var}(\hat{A}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \mathrm{Var}(\hat{B}) = \sigma^2 \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

et de covariance

$$\mathrm{Cov}(\hat{A}, \hat{B}) = -\sigma^2 \frac{\bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Propriétés de l'EMC

$$\mathrm{Var}(\hat{A}) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{et} \quad \mathrm{Var}(\hat{B}) = \sigma^2 \frac{\frac{1}{n}\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

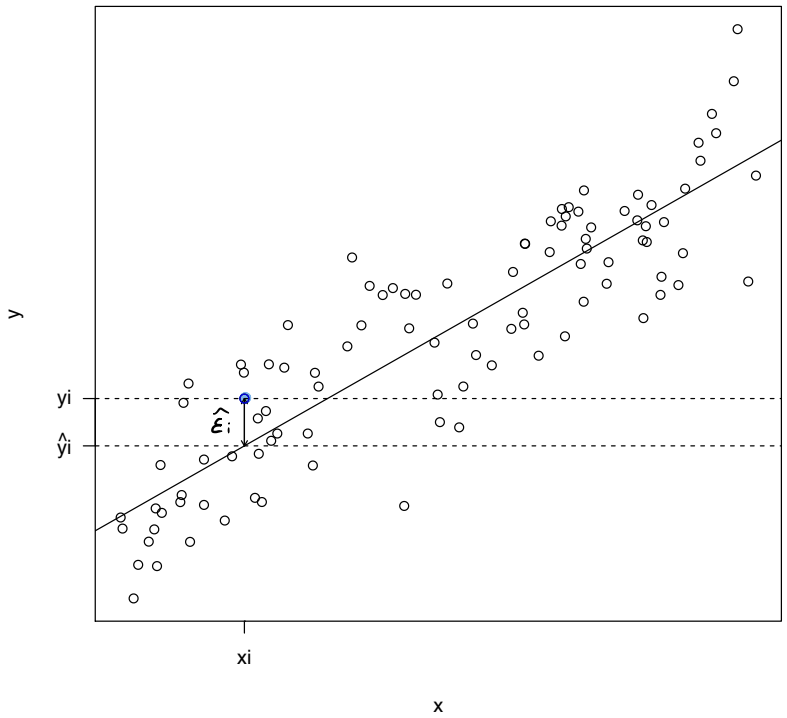
- Plus la variance est faible, plus l'estimateur est précis.
- Variances petites si numérateur petit et/ou dénominateur grand.
- Ces estimateurs $\hat{A}$ et $\hat{B}$ sont donc de faible variance lorsque
  - σ^2 est faible, signifie que le bruit est faible.
  - $\sum_{i=1}^n (x_i - \bar{x})^2$ est grande, signifie que x est dispersé
  - $\sum_{i=1}^n x_i^2$ n'est pas trop grande.

Valeurs ajustées

- Valeur ajustée de y_i par le modèle :

$$\hat{y}_i = \hat{a}x_i + \hat{b}$$

- Ordonnée du point de la droite des MC d'abscisse x_i.



Sous R

- Hauteur d'eucalyptus prédite par le modèle si circonférence=x_i.
- Vecteur des valeurs ajustées : `reg$fitted.values` ou `fitted(reg)`

Résidus : définitions

- Résidus estimés

$$\hat{e}_i = y_i - \hat{y}_i$$

- Résidus : estimateurs des erreurs inconnues ε_i

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i = Y_i - (\hat{B} + \hat{A}x_i)$$

où $\hat{Y}_i$ est la valeur ajustée de Y_i, c'est-à-dire $\hat{Y}_i = \hat{B} + \hat{A}x_i$.

Résidus : propriétés probabilistes

- $\hat{\varepsilon} = (\hat{\varepsilon}_1, \ldots, \hat{\varepsilon}_n)^T$ et $\hat{\beta}$ sont gaussiens et indépendants (admis)
- Donc les vecteurs $\hat{\varepsilon}$ et $(\hat{Y}_i)_{i=1\ldots n}$ sont gaussiens et indépendants

Estimateur de la variance du bruit

Théorème (Estimateur de σ^2)

L'estimateur de σ^2 suivant

$$S^2 = \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}\hat{\varepsilon}_i^2$$

est sans biais et vérifie

$$(n-2)\hat{\sigma}^2/\sigma^2 = \sum_{i=1}^{n}\hat{\varepsilon}_i^2/\sigma^2 \sim \chi^2(n-2)$$

Commentaires

- La loi des grands nombres donne $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i^2 = \sigma^2$ p.s.

Commentaires

- La loi des grands nombres donne $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \sigma^2$ p.s.
- On remplace les ε_i inconnus par les résidus $\hat{\varepsilon}_i$ et cela donne

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

Commentaires

- La loi des grands nombres donne $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \sigma^2$ p.s.
- On remplace les ε_i inconnus par les résidus $\hat{\varepsilon}_i$ et cela donne

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

- Cet estimateur est biaisé

Commentaires

- La loi des grands nombres donne $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \sigma^2$ p.s.
- On remplace les ε_i inconnus par les résidus $\hat{\varepsilon}_i$ et cela donne

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2.$$

- Cet estimateur est biaisé
- Pour débiaiser : remplacer $1/n$ par $1/(n-2)$

Sortie commande R

- Dans la sortie de `summary(reg)`, l'estimateur de σ^2 correspond au carré de la `Residual standard error` et vaut donc $(1.199)^2$, autrement dit "residual standard error" est l'estimateur de l'écart-type σ. On peut aussi y accéder par

```
> summary(reg)$sigma^2
[1] 1.438041
```

Remarques

- $\hat{\sigma}^2$ étant une fonction de $\hat{\varepsilon}$, on a aussi :

$$\hat{\sigma}^2 \text{ est indépendant de } \hat{Y}$$

Remarques

- $\hat{\sigma}^2$ étant une fonction de $\hat{\varepsilon}$, on a aussi :

$$\hat{\sigma}^2 \text{ est indépendant de } \hat{Y}$$

- $\tilde{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} \hat{\varepsilon}_i^2$ est l'EMV de σ^2 dans le modèle gaussien. Autrement dit l'EMV de σ^2 est $\frac{n-2}{n}\hat{\sigma}^2$.

Exercice

1. Charger le fichier "jouet1.txt" dans un dataframe nommé `jouet1`. Au vu du graphique de y contre x, une régression linéaire vous semble-t-elle indiquée ?
2. Faire la régression de y sur x et mettre le résultat dans un objet nommé `reg`. Afficher le résumé des résultats. Le résultat confirme-t-il la réponse à la question 1 ?
3. Afficher le graphique des résidus $\hat{\epsilon}_i$ contre les valeurs ajustées $\hat{y}_i$. Que penser de ce graphique au vu de ce que l'on sait sur ces deux quantités ?
4. Afficher le graphique des résidus contre x. Identifier le problème.

Principe

- Valeur ponctuelle d'un estimateur en général insuffisante
- Nécessaire de lui adjoindre un intervalle de confiance (IC) : donne une idée de la variabilité de l'estimateur.
- IC pour chaque paramètre a et b.

Definition

Soient $I_1(Y)$ et $I_2(Y)$ deux variables aléatoires construites à partir d'un vecteur aléatoire Y dont la loi dépend de θ.
$[I_1(Y), I_2(Y)]$ est un *intervalle de confiance* pour θ de niveau $1-\alpha$ si

$$\mathrm{P}_Y\left([I_1(Y), I_2(Y)] \ni \theta\right) = 1-\alpha$$

Pour une réalisation y de Y, $[I_1(y), I_2(y)]$ est la *fourchette de confiance* de θ. En général, on confond fourchette et intervalle.

Intervalles de confiance sur les paramètres

- Un IC de niveau $1-\alpha$ du paramètre A est donné par

$$[\hat{A} - \hat{\sigma}_{\hat{A}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}, \hat{A} + \hat{\sigma}_{\hat{A}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}]$$

- Un IC de niveau $1-\alpha$ du paramètre b est donné par

$$[\hat{B} - \hat{\sigma}_{\hat{B}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}, \hat{B} + \hat{\sigma}_{\hat{B}} q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}]$$

où $q^{1-\alpha/2}_{\mathcal{T}(n-2)}$ est le quantile d'ordre $1-\alpha/2$ d'une loi $\mathcal{T}(n-2)$, et $\mathcal{T}(n-2)$ est la loi de Student à $n-2$ degrés de liberté

Démonstration faîte au tableau

IC sous R

L'intervalle (bilatéral symétrique) de confiance à 95% est donné par défaut :

```
confint(reg)
```

```
##                 2.5 %    97.5 %
## (Intercept) 8.6847719 9.3901795
## circ        0.2498055 0.2644702
```

Donc, *sous l'hypothèse que nos données sont la réalisation d'un modèle linéaire gaussien*, on a 95% de chances pour que la fourchette 0.2498055 et 0.2644702 contienne le vrai paramètre a.

IC sous R

Changer le niveau de confiance avec l'argument `level`, par exemple pour un niveau de confiance de 97%

```
confint(reg,level=0.97)
```

```
##                   1.5 %    98.5 %
## (Intercept) 8.6468993 9.4280520
## circ        0.2490182 0.2652575
```

Remarque : Plus on demande une confiance élevée et plus l'intervalle de confiance est large. Si on avait demandé une confiance de 100%, on aurait eu pour intervalle de confiance $]-\infty, +\infty[$.

Test d'hypothèse

- Question importante : la variable x a une influence sur la variable d'intérêt y ?
- i.e. on cherche à tester l'hypothèse

$$\mathcal{H}_0 : a = 0 \quad \text{versus} \quad \mathcal{H}_1 : a \neq 0.$$

- Pour tester $a = 0$, on peut seulement comparer $\hat{a}$ à 0.
- On va rejeter $\mathcal{H}_0$ si $|\hat{a}| > s$ où s prend en compte la variabilité de notre estimation et l'erreur qu'on accepte de faire, (i.e. la probabilité (sous Y) de rejeter $\mathcal{H}_0$ alors que $\mathcal{H}_0$ est vraie)
- Besoin d'une loi pivotale (i.e. une loi ne dépendant pas des paramètres inconnus).

Construction d'une statistique de test d'hypothèse

- $\hat{a}$ est la réalisation de $\hat{A}$
- $\hat{A} \sim \mathcal{N}(a, \sigma^2 \rho_x)$.
- σ^2 est inconnu donc on a besoin de le remplacer par son estimateur.
- $$\frac{\hat{A} - a}{\hat{\sigma}_{\hat{A}}} \sim \mathcal{T}(n-2)$$
- En particulier, sous $\mathcal{H}_0$, $a = 0$ donc
$$T = \frac{\hat{A}}{\hat{\sigma}_{\hat{A}}} \sim \mathcal{T}(n-2).$$
- On va rejeter $\mathcal{H}_0$ si $|T| > q$

q tel que $\mathrm{P}_{\mathcal{H}_0}(\text{rejeter } \mathcal{H}_0) = \alpha$, i.e. $\mathrm{P}_{\mathcal{H}_0}(|T| > q) = \alpha$.
q est donc le quantile de niveau $1 - \frac{\alpha}{2}$ d'une $\mathcal{T}(n-2)$.

Test d'hypothèse

- **Test** : La région de rejet $|T| > q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}$ fournit un test de niveau $1-\alpha$ de l'hypothèse $\mathcal{H}_0 : a = 0$ versus $\mathcal{H}_1 : a \neq 0$.
- **p-valeur**
  - p-valeur, i.e. le niveau α le plus petit tel que pour tout niveau au dessus on rejette $\mathcal{H}_0$.
  - Si p-valeur est plus petite que 5% alors on rejette au niveau 5%.
  - Sinon, on n'a pas assez d'information suffisante pour rejetter $\mathcal{H}_0$, soit parce que $\mathcal{H}$ est fausse, soit parce que notre estimateur a une variance trop grande (trop d'incertitude).

Sous R

```
summary(reg)
```

```
##
...
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 9.037476   0.179802   50.26   <2e-16 ***
## circ        0.257138   0.003738   68.79   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
 ...
```

On constate que la p-valeur du test $a = 0$ est $< 2e - 16$. Donc même pour une erreur de $\alpha = 2e - 16$ on rejette $\mathcal{H}_0$. La circonférence du tronc a bien un effet sur la taille de l'arbre.

Définition

- Prédire / prévoir est l'un des but de la régression : prédire la valeur de y pour une nouvelle valeur de x
- **Exemple**
  - Soit x_{n+1} une nouvelle valeur de la variable explicative (circ)
  - Nous voulons prédire y_{n+1} (hauteur de l'arbre)
- Soit $\hat{y}_{n+1}^p$ la valeur prédite, y_{n+1} étant la vraie valeur, inconnue. On n'observe que x_{n+1}.
- Le modèle indique que

$$y_{n+1} = b + ax_{n+1} + \varepsilon_{n+1}$$

avec $\varepsilon_{n+1} \sim \mathcal{N}(0, \sigma^2)$ et ε_{n+1} indépendant des $(\varepsilon_i)_{1 \leq i \leq n}$.

- Nous prédisons la valeur correspondante grâce au modèle estimé

$$\hat{y}_{n+1}^p = \hat{a}x_{n+1} + \hat{b}$$

- La valeur pour laquelle nous effectuons la prévision, ici la $(n+1)$ème n'a pas servi pour le calcul de l'estimateur $\hat{\beta}$.

Prévision : code R

- Pour prédire de nouvelles valeurs $(y_{n+1}, \ldots, y_{n+p})$ à partir de $(x_{n+1}, \ldots, x_{n+p})$, on utilise la fonction `predict`

Prévision : code R

- Pour prédire de nouvelles valeurs $(y_{n+1}, \ldots, y_{n+p})$ à partir de $(x_{n+1}, \ldots, x_{n+p})$, on utilise la fonction `predict`
- Elle prend en entrée le résultat de la fonction `lm`, pour notre exemple `reg`, et les nouvelles valeurs des variables explicatives $(x_{n+1}, \ldots, x_{n+p})$ mises dans la colonne d'un data.frame. Cette colonne doit avoir pour nom le nom de la variable, ici `circ`.

Prévision : code R

- Pour prédire de nouvelles valeurs $(y_{n+1}, \ldots, y_{n+p})$ à partir de $(x_{n+1}, \ldots, x_{n+p})$, on utilise la fonction `predict`
- Elle prend en entrée le résultat de la fonction `lm`, pour notre exemple `reg`, et les nouvelles valeurs des variables explicatives $(x_{n+1}, \ldots, x_{n+p})$ mises dans la colonne d'un data.frame. Cette colonne doit avoir pour nom le nom de la variable, ici `circ`.
- Imaginons qu'on ait mesuré les circonférences de trois nouveaux arbres : 46, 35 et 67. Cela donne le code suivant :

Prévision : code R

- Pour prédire de nouvelles valeurs $(y_{n+1}, \ldots, y_{n+p})$ à partir de $(x_{n+1}, \ldots, x_{n+p})$, on utilise la fonction `predict`
- Elle prend en entrée le résultat de la fonction `lm`, pour notre exemple `reg`, et les nouvelles valeurs des variables explicatives $(x_{n+1}, \ldots, x_{n+p})$ mises dans la colonne d'un data.frame. Cette colonne doit avoir pour nom le nom de la variable, ici `circ`.
- Imaginons qu'on ait mesuré les circonférences de trois nouveaux arbres : 46, 35 et 67. Cela donne le code suivant :

```
xnew=c(46,35,67)
xnew=data.frame(circ=xnew)
predict(reg,new=xnew)
```

```
##        1        2        3
## 20.86582 18.03730 26.26571
```

Prévision : variance

Proposition (Variance de la prévision)

Posons : $\hat{Y}^p_{n+1} = \hat{A}x_{n+1} + \hat{B}$. $\hat{Y}^p_{n+1}$ *est une variable aléatoire gaussienne de variance :*

$$\text{Var}(\hat{Y}^p_{n+1}) = \sigma^2\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)$$

Prévision : variance avec R

- La variance de prévision fait intervenir σ^2 (inconnu)
- Remplacer σ^2 (inconnu) par un estimateur $\hat{\sigma}^2$
- Estimateur de la variance de prédiction :

$$\widehat{\mathrm{Var}}(\hat{y}^p_{n+1}) = \hat{\sigma}^2\left(\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right).$$

- La racine carrée de cette variance (l'écart-type="standard error") est calculée par R. Il suffit de rajouter l'argument `se.fit=T` dans l'appel de la fonction `predict`.

Prévision : variance avec R

```
predict(reg,new=xnew,se.fit=T)
```

```
## $fit
##        1        2        3
## 20.86582 18.03730 26.26571
##
## $se.fit
##          1          2          3
## 0.03212020 0.05600524 0.08001489
##
## $df
## [1] 1427
##
## $residual.scale
## [1] 1.199183
```

Erreur de Prévision

Proposition (Erreur de prévision)

L'erreur de prévision

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{Y}^p_{n+1}$$

vérifie $\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0$ *et* $\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$.

- Quantifie la capacité du modèle à prévoir.

$$\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$$

Erreur de Prévision

Proposition (Erreur de prévision)

L'erreur de prévision

$$\hat{\varepsilon}^p_{n+1} = y_{n+1} - \hat{Y}^p_{n+1}$$

vérifie $\mathbb{E}(\hat{\varepsilon}^p_{n+1}) = 0$ *et* $\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$.

- Quantifie la capacité du modèle à prévoir.

$$\mathrm{Var}(\hat{\varepsilon}^p_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}\right)$$

- Mauvais quand x_{n+1} est "loin" de $\bar{x}$.

IC pour y_{n+1}

Proposition (IC pour y_{n+1})

Un IC de y_{n+1} est donné par :

$$\left[\hat{Y}^p_{n+1} \pm q^{1-\frac{\alpha}{2}}_{\mathcal{T}(n-2)}\hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x_{n+1}-\bar{x})^2}{\sum_{i=1}^n (x_i-\bar{x})^2}}\right]$$

Autrement dit, avec une probabilité $1-\alpha$, cet intervalle contient la vraie valeur y_{n+1}.

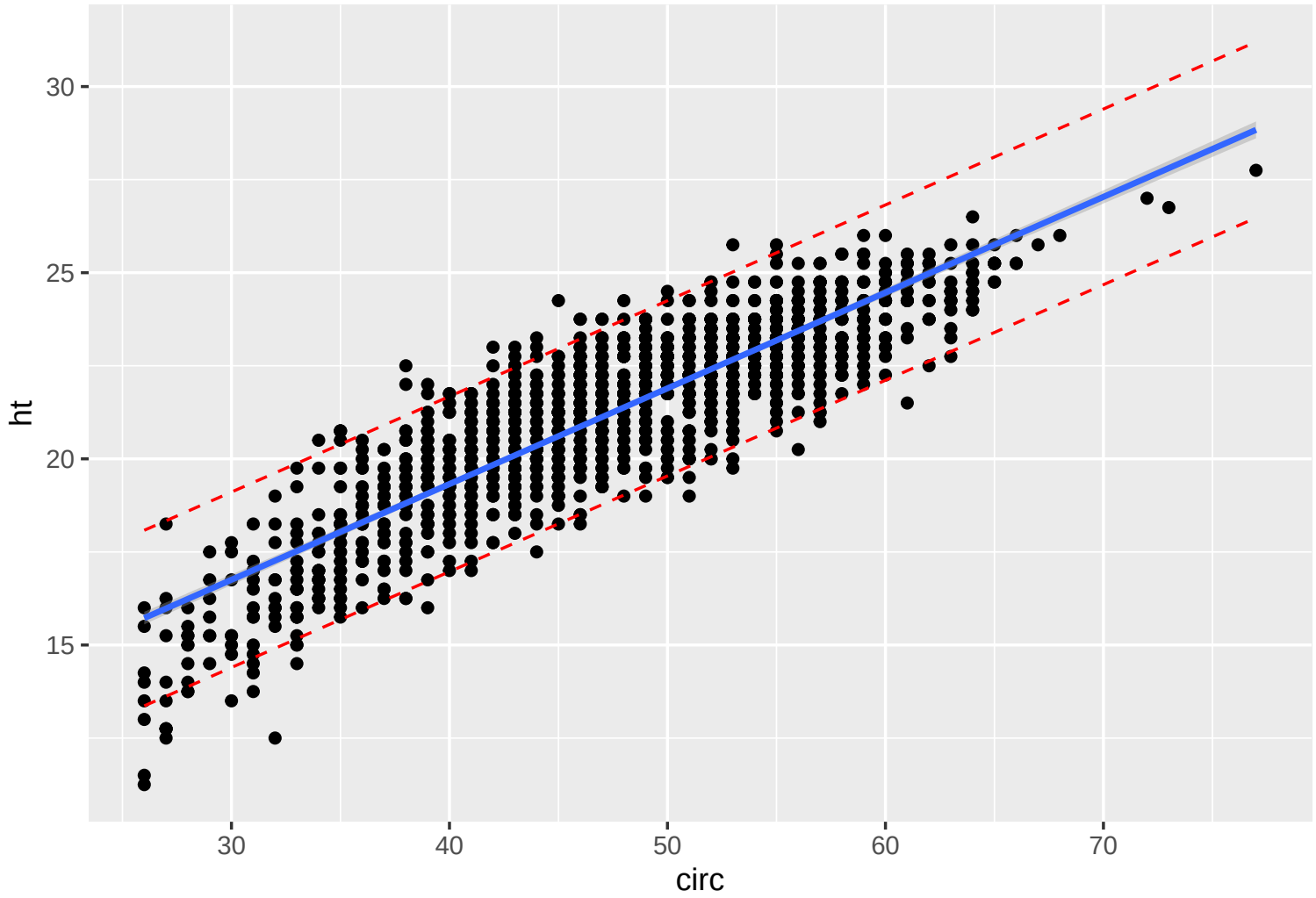
Sous **R**,

```
> predict(reg,new=xnew,interval="pred",level=0.95)
       fit      lwr      upr
1 20.86582 18.51262 23.21901
2 18.03730 15.68239 20.39222
3 26.26571 23.90813 28.62329
```

Tracé de la droite de prédiction avec intervalle de confiance

```
xnew <- seq(min(euca$circ), max(euca$circ),len=1000)
xnew <- data.frame(xnew);
names(xnew) = 'circ'
ICpred <- as.data.frame(predict(reg,xnew,interval="pred",
level=0.95))
res_pred <- cbind(ICpred,xnew)
ggplot(euca, aes(x = circ, y = ht))+ geom_point() +
 geom_smooth(method = lm)
+  geom_line(data=res_pred,aes(x=circ,y=lwr), color = "red",
linetype = "dashed")
+ geom_line(data=res_pred, aes(x = circ,y=upr), color = "red",
linetype = "dashed")
```

Tracé de la droite de prédiction avec intervalle de confiance



Validité des résultats

- Tests et intervalles de confiance valables *sous l'hypothèse que nos observations sont la réalisation d'un modèle linéaire gaussien*
- Première chose à faire : vérifier **sur nos données** que les postulats que l'on a défini sur les erreurs ε_i sont vérifiés :
  - [P1] Les erreurs sont centrées : $\mathbb{E}[\boldsymbol{\varepsilon}] = 0_{\mathbb{R}}$.
  - [P2] Les erreurs $\boldsymbol{\varepsilon}$ sont de variance constante : $\mathbb{V}[\varepsilon_i] = \sigma^2, \forall i = 1 \dots n$.
  - [P3] Les termes d'erreur sont supposés indépendants.
  - [P4] Les erreurs sont supposées gaussiennes.

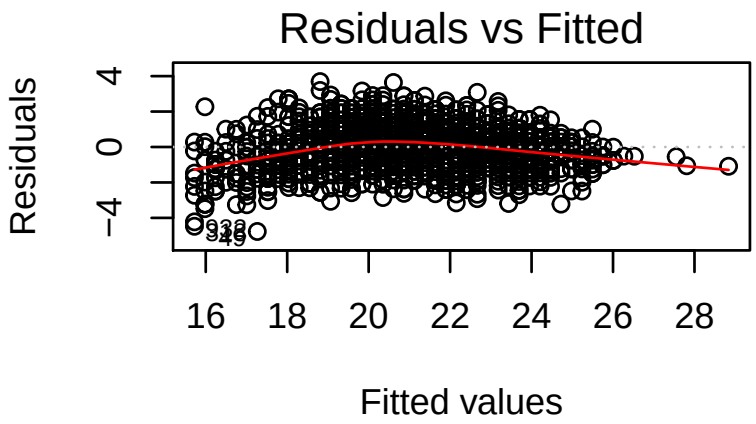
Vérification des postulats

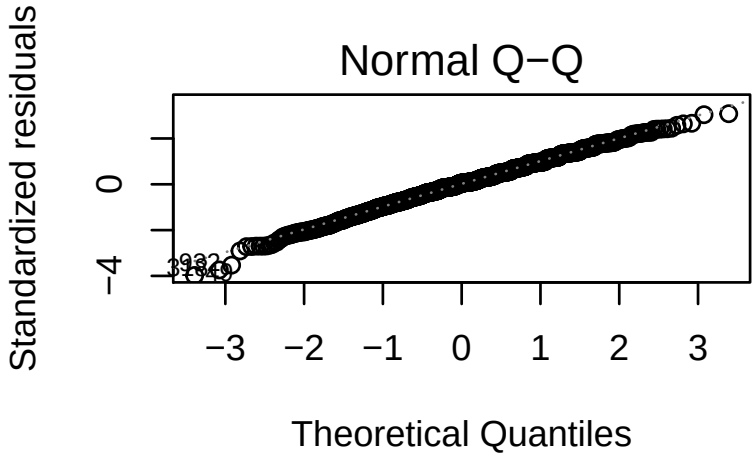
- Important de s'assurer qu'ils sont vérifiés *sur nos données*.
- ε_i non observés $\Rightarrow$ pas de tests statistiques
- Validation grâce à des outils graphiques utilisant les résidus estimés $\hat{e}_i = y_i - \hat{y}_i$.
- Les 4 graphiques utiles sont donnés par la commande **R** suivante :

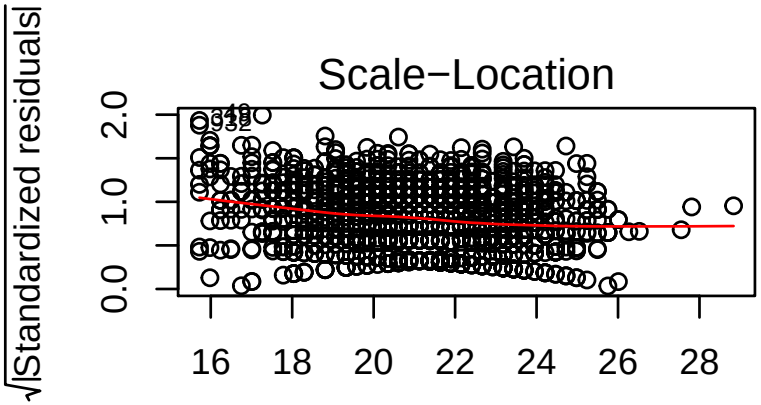
```
par(mfrow=c(2,2))
plot(reg)
```

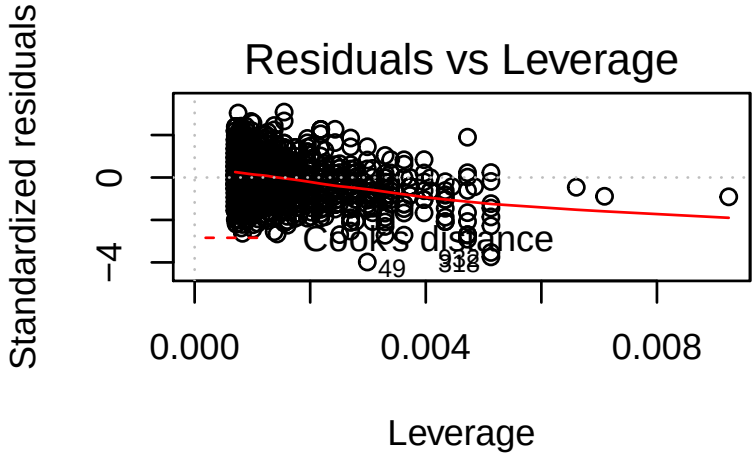
Graphes des résidus

Sur le jeu de données Eucalyptus.









Validation du postulat [P3]

L'indépendance des données ne peut être assurée que par le protocole expérimental.

Validation du postulat [P1]

- Revient à valider que la relation en y et x est bien affine.
- Par construction $\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i = 0 \Rightarrow$ ne peut servir pour remettre en cause [P1]
- Graphe $(\hat{y}_i, \hat{\varepsilon}_i)$ (en haut à gauche).
  - Si l'on observe un nuage de points centré et aligné sans structure particulière, alors on se satisfait.
  - Par contre, si on observe une structure particulière (croissance des résidus en fonction des données prédites par exemple ou autre) alors on pourra penser que le modèle n'est pas adapté aux données, qu'il manque une tendance dans le modèle.
  - Plusieurs solutions sont possibles : travailler avec le log des observations, travailler avec le log des variables explicatives...

Dans le cas de nos données Eucalyptus, le premier graphe en haut à gauche semble montrer une tendance. Nous pourrons essayer travailler sur le log de la hauteur.

Validation du postulat [P2]

Les résidus sont-ils homoscédastiques (de même variance) ?

- $\hat{e}_i$ se sont pas de variance constante $\Rightarrow$ version standardisée
- Graphe de $(\hat{y}_i, \hat{r}_i)$ avec $\hat{r}_i = \frac{\hat{e}_i}{\hat{s}e_i}$ où se_i est l'écart-type estimé des $\hat{e}_i$) : Graphique en bas à gauche).
  - Si l'on observe un nuage de points centré et aligné sans structure particulière, alors on se satisfait.
  - Par contre, si on observe une structure particulière (croissance des résidus standardisés en fonction des données prédites par exemple ou autre) alors on pourra penser que le modèle n'est pas homoscédastique.

Validation du postulat [P4]

Les résidus sont-ils gaussiens ?

- Le QQ-plot des résidus estimés (graphique en haut à droite) : tester le caractère gaussien des résidus.
- Cependant, ce postulat [P4] n'est pas obligatoire car si le nombre d'observations n est grand on peut obtenir des propriétés asymptotiques des estimateurs et tests.

Détection des points abhérrants

- Quatrième graphe de diagnostique (en bas à droite) : détection des observations y_i ayant eu une grande influence sur l'estimation.
- Si un y_i très influente sur l'inférence : problématique car conclusions non stables si on considère un autre échantillon. *Graphe "residuals versus leverage"*.
- Rappel : prédictions combinaisons linéaires des observations

$$\hat{y}_i = \sum_{j=1}^{n} h_{ij} y_j$$

- Si $|h_{ij}|$ est grand alors l'observation j pèse beaucoup dans la prédiction $\hat{y}_i$.
- Le graphe "residuals versus leverage" trace $\hat{r}_i$ en fonction de h_{ii}.

Residuals versus leverage

Si $\hat{r}_i$ grand, le point est mal ajusté, il est atypique par rapport aux autres données.

- *Si $\hat{r}_i$ grand et h_{ii} petit*, il est atypique mais peu influent sur l'estimation des paramètres. Donc ce n'est pas très grave
- *Si $\hat{r}_i$ grand et h_{ii} grand*, alors il est atypique ET influe beaucoup sur l'estimation. Cela devient problématique.

Distance de Cook

$$d_i = \frac{\sum_{j=1}^{n} (\hat{y}_j^{(-i)} - \hat{y}_j)^2}{(p+1)\hat{\sigma}^2} = \frac{\hat{r}_i^2 h_{ii}}{(p+1)(1-h_{ii})^2}$$

où $\hat{y}_i^{(-j)}$ est la prédiction de la i-ème observation lorsque l'analyse a été faite en retirant l'observation j.

- Plus cette distance est grande et plus le point i est influent.
- En général on regarde de près les points dont la distance de Cook dépasse 1.
- On pourra décider d'enlever les points aberrants.

Version plus complète de la validation de modèle dans les annexes du poly.

En pratique

Les grandes étapes

1. Exploration des données et calcul de statistiques descriptives
2. Tracé du nuage de points (x_i, y_i)
3. Appliquer `lm`
4. Validation des postulats par l'étude des graphes de résidus.
   Possiblement modifier le modèle pour obtenir des résidus satisfisants
5. Tests et intervalles de confiance, interprétation des résultats...

Exercice sur les eucalyptus

Au vu des graphiques de validation de modèle sur les Eucalyptus, on propose de changer de modèle et d'expliquer la hauteur en fonction du logarithme de la circonférence du tronc.

1. Ecrire les instructions **R**.
2. Afficher les 4 graphes de validation. A-t-on amélioré les choses ?

Exercice sur les données ozone

On va utiliser les données contenues dans le fichier ozone.txt. On va étudier l'influence de la température sur la concentration en ozone : `max03` est la concentration en ozone et `T12` la température à 12H. On choisit donc `max03` comme variable à expliquer et `T12` comme variable explicative.

1. Importer le fichier texte ozone.txt dans un data.frame qu'on nommera `ozone`. Afficher ses variables. Que vaut n ici ?
2. Vous semble-t-il opportun d'utiliser le modèle linéaire ?
3. Afficher le résumé des informations de la régression. Donner l'EMC $\hat{\beta}$. Donner les intervalles de confiance à 95%.
4. Est-ce que la variable `T12` vous semble bien expliquer linéairement la variable `max03` ? Donner le résultat du test.
5. Donner la prédiction d'ozone pour une température égale à 27 degrés ainsi que l'intervalle de confiance à 95% correspondant.