Modèles de prévision
Partie 1 - régression - # 2

Arthur Charpentier

charpentier.arthur@uqam.ca

http ://freakonometrics.hypotheses.org/

UQÀM
Université du Québec à Montréal

Été 2014

Plan du cours - données individuelles

- ◦ Rappels de statistique
- • **Motivation et introduction aux modèles de régression**
- • **Le modèle linéaire simple**
- ◦ Résultats généraux
- ◦ Approche matricielle
- • **Le modèle linéaire multiple**
- ◦ Résultats généraux
- ◦ Tests, choix de modèle, diagnostique
- • **Aller plus loin**
- ◦ Les modèles non linéaires paramétriques
- ◦ Les modèles non linéaires nonparamétriques

Petit rappel sur la significativité, test de $H_0 : \beta_j = 0$

Les résultats précédants permettent de proposer un test simple de

$$H_0 : \beta_j = 0 \text{ contre l'hypothèse } H_1 : \beta_j \neq 0.$$

La statistique de test

$$T_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{Var}(\widehat{\beta}_j)}} \sim \mathcal{S}t(n-k) \text{ sous } H_0.$$

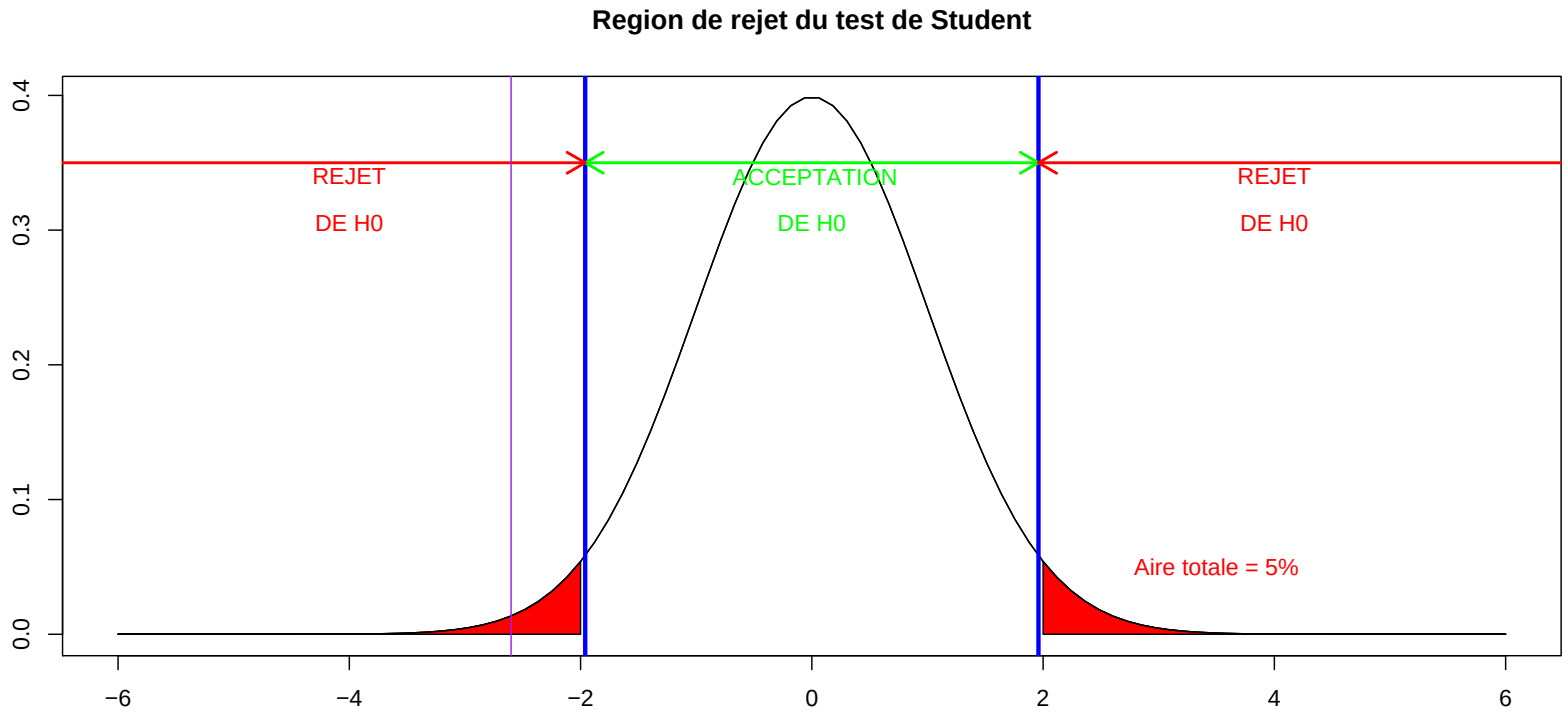
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
```

Les deux lectures possibles d'un test

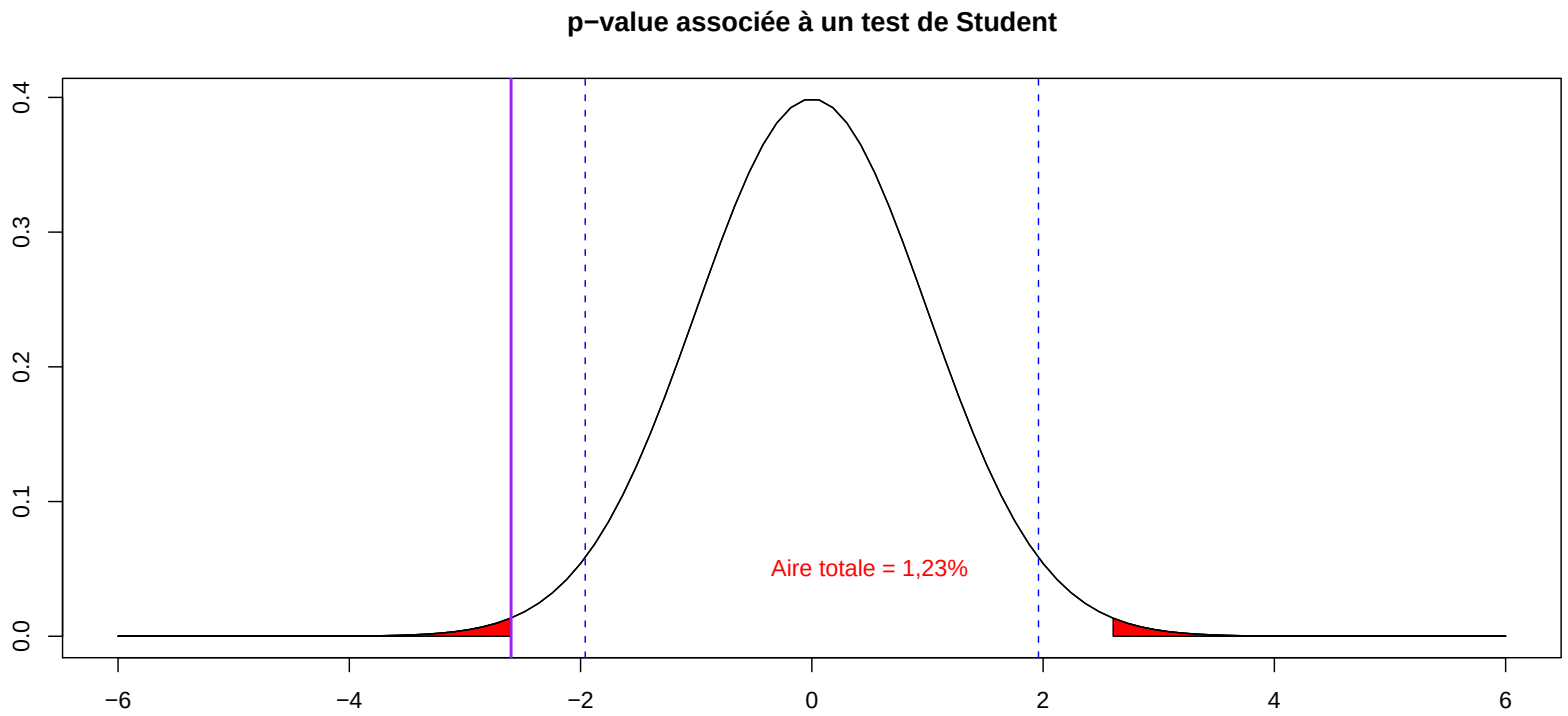
- donner la region de rejet, de la forme $[\pm T_{1-\alpha/2}]$, avec un seuil α fixé arbitrairement (par dfaut 95%)
- donner le seuil α tel que la région de rejet soit $[\pm\widehat{t}]$ (la plus petite region de rejet à laquelle appartienne la statistique observée), i.e. la probabilité que de rejeter H_0 si H_0 était vraie.

Dans ce dernier cas, on parle de p-value, $p = \mathbb{P}(\text{rejeter } H_0 | H_0 \text{ vraie})$: si p est faible, on rejette H_0, car il y a peu de chances qu'H_0 soit vraie.

Lecture du test de Student



Lecture du test de Student



Analyse d'une sortie de régression

Des tests de student de $H_0 : \beta_i = 0$, contre $H_1 : \beta_i \neq 0$ sont proposés, avec

$$t_0 = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\widehat{Var}(\widehat{\beta}_0)}} = \frac{-17.5791 - 0}{6.7584} = -2.601 \text{ sous} H_0$$

$$t_1 = \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var}(\widehat{\beta}_0)}} = \frac{3.9324 - 0}{0.4155} = 9.464 \text{ sous} H_0$$

Ces valeurs sont à comparer avec le quantile de Student à 95% (à 49 degrés de liberté).

Une alternative est d'utiliser la p-value, i.e. si $Z \sim \mathcal{S}t(49)$,

$$p_0 = \mathbb{P}(|Z| > t_0) = 0.0123 \text{ et } p_1 = \mathbb{P}(|Z| > t_1) = 1.49 \times 10^{-12}.$$

La p value est alors donnée par

```
> 2*(1-pt(abs(REG$coefficients[1]/summary(REG)$coefficients[1,2]), df=n-2))
(Intercept)
 0.01231882
```

$\widehat{\sigma} = 15.38$, i.e. `summary(reg)$sigma`

```
> confint(reg)
                 2.5 %    97.5 %
(Intercept) -31.167850 -3.990340
speed         3.096964  4.767853
```

Pour la constante, par exemple, l'intervalle de confiance est donné par

```
> REG$coefficients[1]+qt(c(.025,.975),n-2)* summary(REG)$coefficients[1,2]
[1] -31.16785  -3.99034
```

La matrice de variance-covariance des coefficients, $\text{Var}(\widehat{\boldsymbol{\beta}})$ est ici

```
> vcov(reg)
             (Intercept)      speed
(Intercept)   45.676514 -2.6588234
speed         -2.658823  0.1726509
```

Introduction aux tests multiples, e.g. $H_0 : \beta_1 = \cdots = \beta_j = 0$

On a vu comment tester $H_0 : \beta_2 = 0$ et $H_0 : \beta_3 = 0$, mais ces deux tests peuvent être validé, sans pour autant avoir $H_0 : \beta_2 = \beta_3 = 0$.

```
> US=read.table("http://freakonometrics.free.fr/US.txt",
+ header=TRUE,sep=";")
> US$Density=US$Population/US$Area
> model1 = lm(Murder ~ Income + HS.Grad + Frost +
+ Population + Illiteracy + Life.Exp +
+ Area + Density, data=US)

> summary(model1)

Call:
lm(formula = Murder ~ Income + HS.Grad + Frost + Population +
Illiteracy + Life.Exp + Area + Density, data = US)

Residuals:
Min       1Q   Median       3Q      Max
-3.10973 -0.92363 -0.07636  0.74884  2.92362
```

```
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.121e+02  1.684e+01   6.657 5.04e-08 ***
Income       1.018e-03  6.642e-04   1.532 0.133084
HS.Grad      1.318e-02  5.315e-02   0.248 0.805412
Frost       -7.301e-03  7.074e-03  -1.032 0.308040
Population   2.180e-04  6.051e-05   3.602 0.000845 ***
Illiteracy   2.208e+00  8.184e-01   2.699 0.010068 *
Life.Exp    -1.579e+00  2.374e-01  -6.652 5.12e-08 ***
Area        -9.413e-07  4.228e-06  -0.223 0.824911
Density     -4.369e+00  1.499e+00  -2.915 0.005740 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 1.608 on 41 degrees of freedom
Multiple R-squared: 0.8412,        Adjusted R-squared: 0.8102
F-statistic: 27.14 on 8 and 41 DF,  p-value: 4.813e-14
```

Sur cette exemple, on valide les tests $H_0 : \beta_1 = 0$, $H_0 : \beta_2 = 0$ et $H_0 : \beta_3 = 0$.
Mais peut-on valider $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$?

Ce test peut s'écrire de manière très générale $H_0 : \boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{q}$ (contre $H_1 : \boldsymbol{R}\boldsymbol{\beta} \neq \boldsymbol{q}$) avec ici

$$\underbrace{\begin{pmatrix} 0 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 1 & 0 & \cdots & 0 \end{pmatrix}}_{\boldsymbol{R}} \underbrace{\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \vdots \\ \beta_k \end{pmatrix}}_{\boldsymbol{\beta}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}}_{\mathbf{0}}$$

La stratégie est de comparer deux modèles : le modèle non-contraint (sous H_1),

$$\widehat{\boldsymbol{\beta}} = \text{argmin}\{(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{R}^{k+1}\}$$

et le modèle non-contraint (sous H_1),

$$\widehat{\boldsymbol{\beta}}_\star = \text{argmin}\{(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}), \boldsymbol{\beta}\in\mathbb{R}^{k+1}, \boldsymbol{R}\boldsymbol{\beta}=\boldsymbol{q}\}$$

Pour le premier modèle, on cherche à minimiser

$$h(\boldsymbol{\beta}) = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$$

et dans le second modèle, c'est de la minimisation sous-contrainte. On optimise le Lagrangien,

$$\ell(\boldsymbol{\beta},\lambda) = (\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \lambda(\boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{q})$$

Dans ce cas, les conditions du premier ordre sont

$$\frac{\partial \ell(\boldsymbol{\beta},\lambda)}{\partial \boldsymbol{\beta}} = 2\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{R}'\boldsymbol{\lambda} = \boldsymbol{0}$$

et

$$\frac{\partial \ell(\boldsymbol{\beta},\lambda)}{\partial \lambda} = \boldsymbol{R}\boldsymbol{\beta}-\boldsymbol{q} = \boldsymbol{0},$$

pour $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_\star$. On a finallement un système de deux (systèmes d') équations

$$\begin{pmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{R}' \\ \boldsymbol{R} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_\star \\ \boldsymbol{\lambda} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}'\boldsymbol{Y} \\ \boldsymbol{q} \end{pmatrix}$$

Comme $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, on peut écrire

$$\widehat{\boldsymbol{\beta}}_\star = \widehat{\boldsymbol{\beta}} - \boldsymbol{C}[\boldsymbol{R}\widehat{\boldsymbol{\beta}}_\star - \boldsymbol{q}]$$

où

$$\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}']^{-1}.$$

Si on pose $\widehat{\boldsymbol{\varepsilon}}_\star = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}_\star$ et $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$, alors

$$\widehat{\boldsymbol{\varepsilon}}_\star'\widehat{\boldsymbol{\varepsilon}}_\star - \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = [\boldsymbol{R}\widehat{\boldsymbol{\beta}}_\star - \boldsymbol{q}]'(\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}')[\boldsymbol{R}\widehat{\boldsymbol{\beta}}_\star - \boldsymbol{q}]$$

Or d'après la seconde condition du premier ordre, $\boldsymbol{R}\widehat{\boldsymbol{\beta}}_\star = \boldsymbol{q}$. Donc sous H_0, la statistique de test est

$$F = \frac{\widehat{\boldsymbol{\varepsilon}}_\star'\widehat{\boldsymbol{\varepsilon}}_\star - \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{\dim(\boldsymbol{q})} \cdot \frac{n-k}{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}$$

qui doit suivre une loi de Fisher, $\mathcal{F}(\dim(\boldsymbol{q}), n-k)$.

```
> (EE=sum(residuals(model1)^2))
[1] 106.0532
>  model2 = lm(Murder ~
+  Population + Illiteracy + Life.Exp +
+  Area + Density, data=US)
>  (EEc=sum(residuals(model2)^2))
[1] 119.6924
>  (F=(EEc-EE)/3*(nrow(US)-9)/(EE))
[1] 1.757643
>  1-pf(F,3,nrow(US)-9)
[1] 0.170363
```

Pour savoir si on rejette, ou si on accepte H_0, on calcule la p-value,

```
>  1-pf(F,3,nrow(US)-9)
[1] 0.170363
```

i.e. on peut accepter ici H_0 (les trois coefficients sont nuls simultanément).

Cette analyse de variance peut se faire via

```
> library(car)
> linearHypothesis(model1,
+ c("Income","HS.Grad","Frost"),c(0,0,0))
Linear hypothesis test

Hypothesis:
Income = 0
HS.Grad = 0
Frost = 0

Model 1: restricted model
Model 2: Murder ~ Income + HS.Grad + Frost +
Population + Illiteracy + Life.Exp + Area + Density

Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     44 119.69
2     41 106.05  3    13.639 1.7576 0.1704
```

Diagnostique et régression, le R^2

Le coefficient de détermination R^2 défini à partir le rapport entre la variance des résidus et la variance de Y,

$$R^2 = 1 - \frac{\text{Variance non expliquée}}{\text{Variance totale}} = \frac{\text{Variance expliquée}}{\text{Variance totale}}.$$

ou pour la version empirique

$$R^2 = 1 - \frac{\text{somme des carrés des résidus}}{\text{somme des carrés de la régression}} = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \widehat{Y}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

Diagnostique et régression, le R^2

On utilise pour cela la formule de décomposition de la variance

$$\underbrace{Var(Y)}_{\text{variance totale}} = \underbrace{Var[\mathbb{E}(Y|X)]}_{\text{variance expliquée par } X} + \underbrace{\mathbb{E}[Var(Y|X)]}_{\text{variance réisudelle}} .$$

On notera que cette grandeur est un estimateur baisé du *vrai* R^2,

$$\mathbb{E}(R^2) = R^2 + \frac{k-1}{n-1}[1-R^2] + O\left(\frac{1}{n^2}\right)$$

Le coefficient d'ajustement est $R^2 = 0.6511$ et $\overline{R}^2 = 0.6438$.

```
> summary(reg)$r.squared
[1] 0.6510794
```

Le calcul se fait de la manière suivante

```
> 1-deviance(REG)/sum((Y-mean(Y))^2)
[1] 0.6510794
```

Afin de prendre en compte le nombre de paramètre, et de corriger du biais, on peut utiliser le R^2 ajusté,

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-(k-1)-1} = \frac{(n-1)R^2 - (k-1)}{n-k-2}$$

où $(k-1)$ est le nombre de variables explicatives (sans la constante). Notons que ce $\overline{R}^2$ peut être négatif.

Remarque En rajoutant des variables explicatives, on ne peut *que* augmenter le R^2, mais si ces dernières sont peu corrélées avec Y.

Remarque Dans un modèle sans constante, le R^2 n'a plus aucun sens. En fait, sans constante, rien ne garantit que le plan de régression passe par le centre de gravité du nuage, $(\overline{\boldsymbol{x}}, \overline{y})$. Et donc la somme des résidus n'est alors pas forcément nulle. La formule de décomposition de la variance n'est alors plus valide.

De l'utilisation du R^2

Considérons une régression linéaire

$$TIN_t = \beta_0 + \beta_1 TIF_t + \varepsilon_t,$$

où TIN désigne le taux d'intérêt nomial, TIF le taux d'inflation et TIR le taux d'intérêt réel, i.e. $TIN = TIR + TIF$. Au lieu de modéliser le taux d'intérêt *nominal* en fonction de l'inflation, supposons que l'on cherche à modéliser le taux d'intérêt *réel*,

$$TIR_t = \alpha_0 + \alpha_1 TIF_t + \eta_t.$$

Notons que de la première équation $TIN_t = \beta_0 + \beta_1 TIF_t + \varepsilon_t = TIR_t + TIF_t$, on en déduit

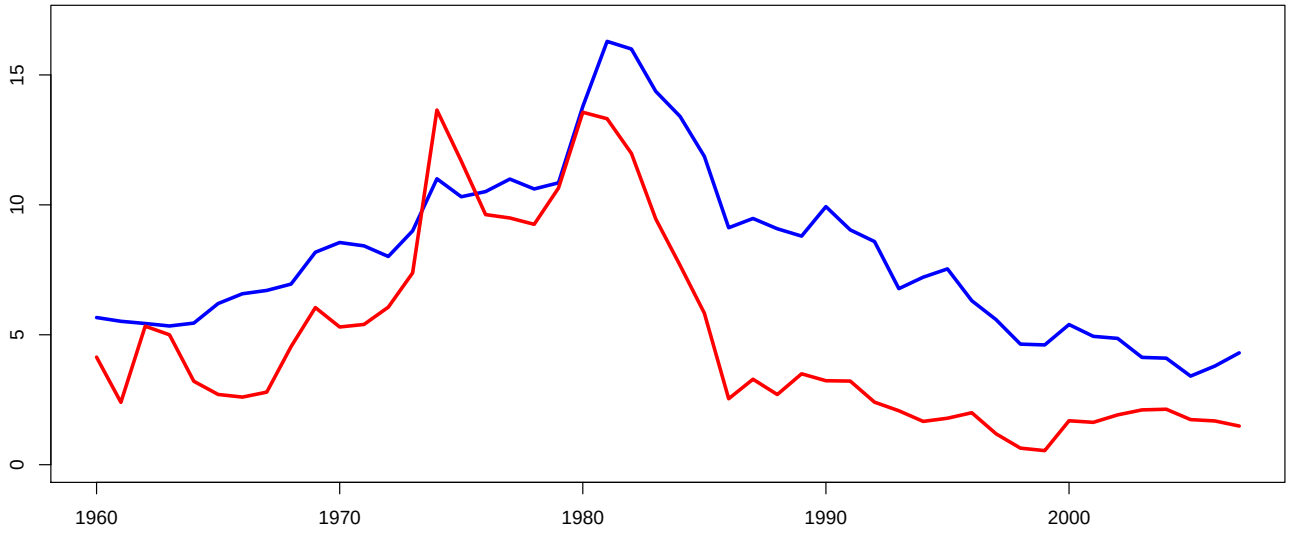
$$TIR_t = \underbrace{\beta_0}_{=\alpha_0} + \underbrace{[\beta_1 - 1]}_{=\beta_0} TIF_t + \underbrace{\varepsilon_t}_{=\eta_t},$$

autrement dit les deux équations sont équivalentes.

Pourtant

$$R^2_{\text{nominal}} = 1 - \frac{Var(\eta)}{Var(TIN)} = 1 - \frac{Var(\eta)}{Var(TIR + TIF)} \geq 1 - \frac{Var(\eta)}{Var(TIR)} = R^2_{\text{réel}}$$

aussi, on peut artificiellement augmenter un R^2, tout en étudiant un modèle rigoureusement équivalent.



De l'utilisation du R^2

Les sorties montrent que les deux sorties sont effectivement équivalence entre les deux modèles

```
> summary(lm(TIR~TIF,data=D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.65040    0.44301  10.497  8.5e-14 ***
TIF         -0.29817    0.07211  -4.135 0.000149 ***

> summary(lm(TIN~TIF,data=D))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.65040    0.44301  10.497 8.50e-14 ***
TIF          0.70183    0.07211   9.733 9.55e-13 ***
```

De l'utilisation du R^2

Mais surtout, on note que le R^2 du premier modèle est beaucoup plus faible que le second

```
> summary(lm(TIR~TIF,data=D))

Multiple R-Squared: 0.271,      Adjusted R-squared: 0.2551

> summary(lm(TIN~TIF,data=D))

Multiple R-Squared: 0.6731,     Adjusted R-squared: 0.666
```

Diagnostique dans le modèle linéaire

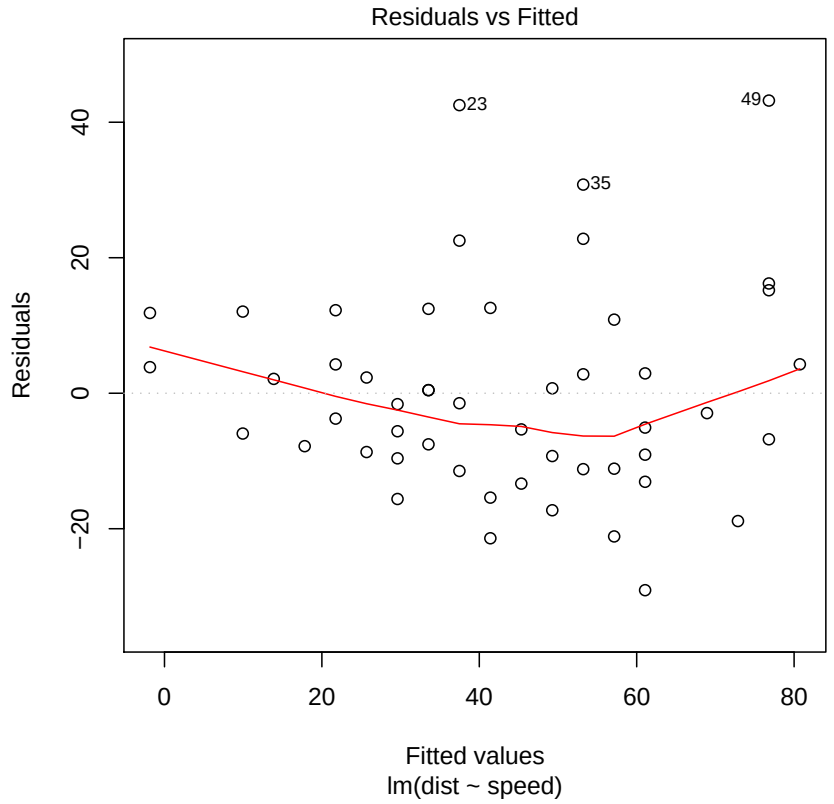
La fonction `plot(REG)` produit 6 graphiques de diagnostique

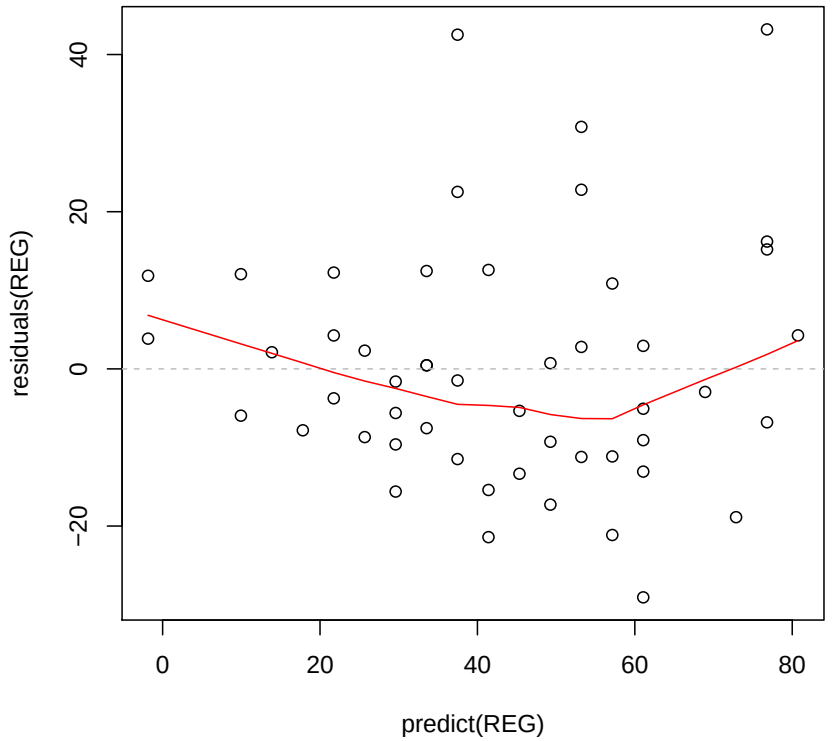
1. résidus contre valeurs estimées, $(\widehat{Y}_i, \widehat{\varepsilon}_i)$ (*plot of residuals against fitted values*)
2. $(\widehat{Y}_i, \sqrt{|\widetilde{\varepsilon}_i|})$ (*Scale-Location plot*),
3. un graphique quantile-quantile des résidus (*Normal Q-Q plot*),
4. un graphique de distances de Cook (*plot of Cook's distances versus row labels*),
5. un graphique de *leverage* (*plot of residuals against leverages*)
6. (*plot of Cook's distances against leverage/(1-leverage)*)

Remarque dans la plupart des graphiques, on utilise les résidus standardisés, i.e. ε/σ, centrés et de variance unitaire.

Diagnostique dans le modèle linéaire

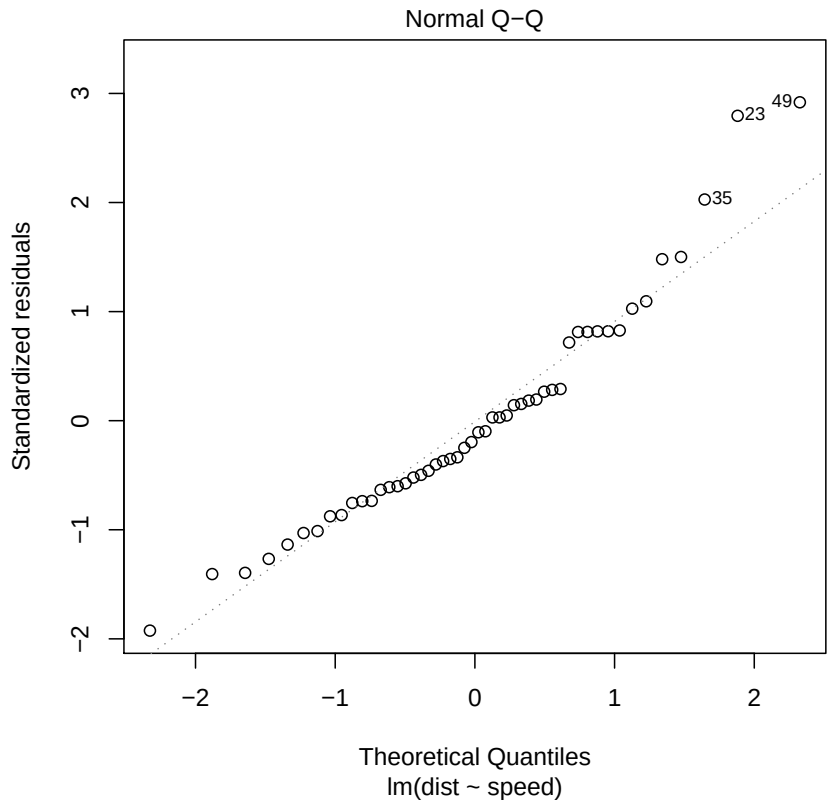
```
> plot(predict(REG),residuals(REG))
> abline(h=0,lty=2,col="grey")
> lines(lowess(predict(REG),residuals(REG)),col="red")
>
```

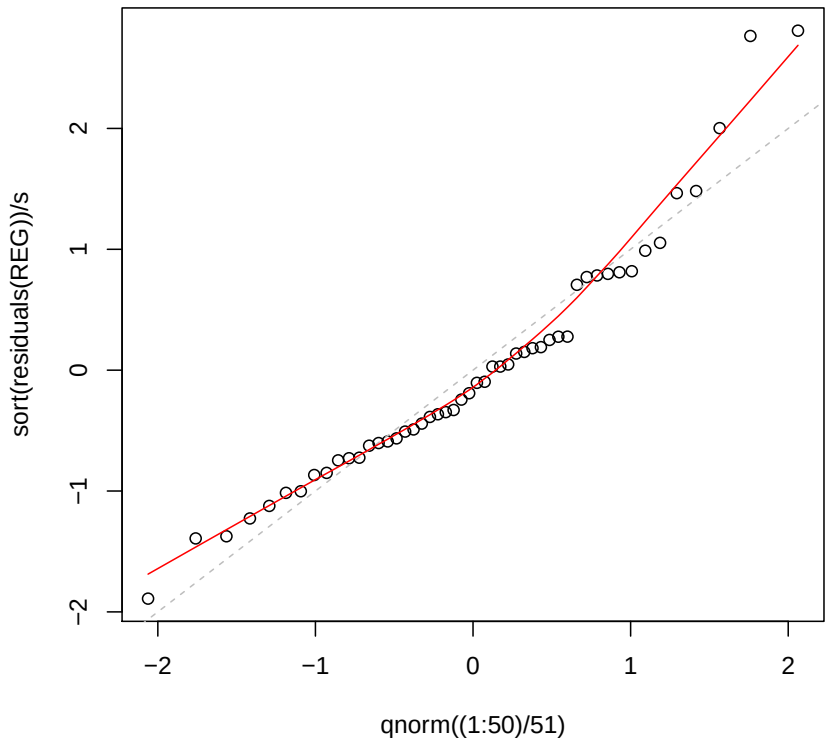




Diagnostique dans le modèle linéaire

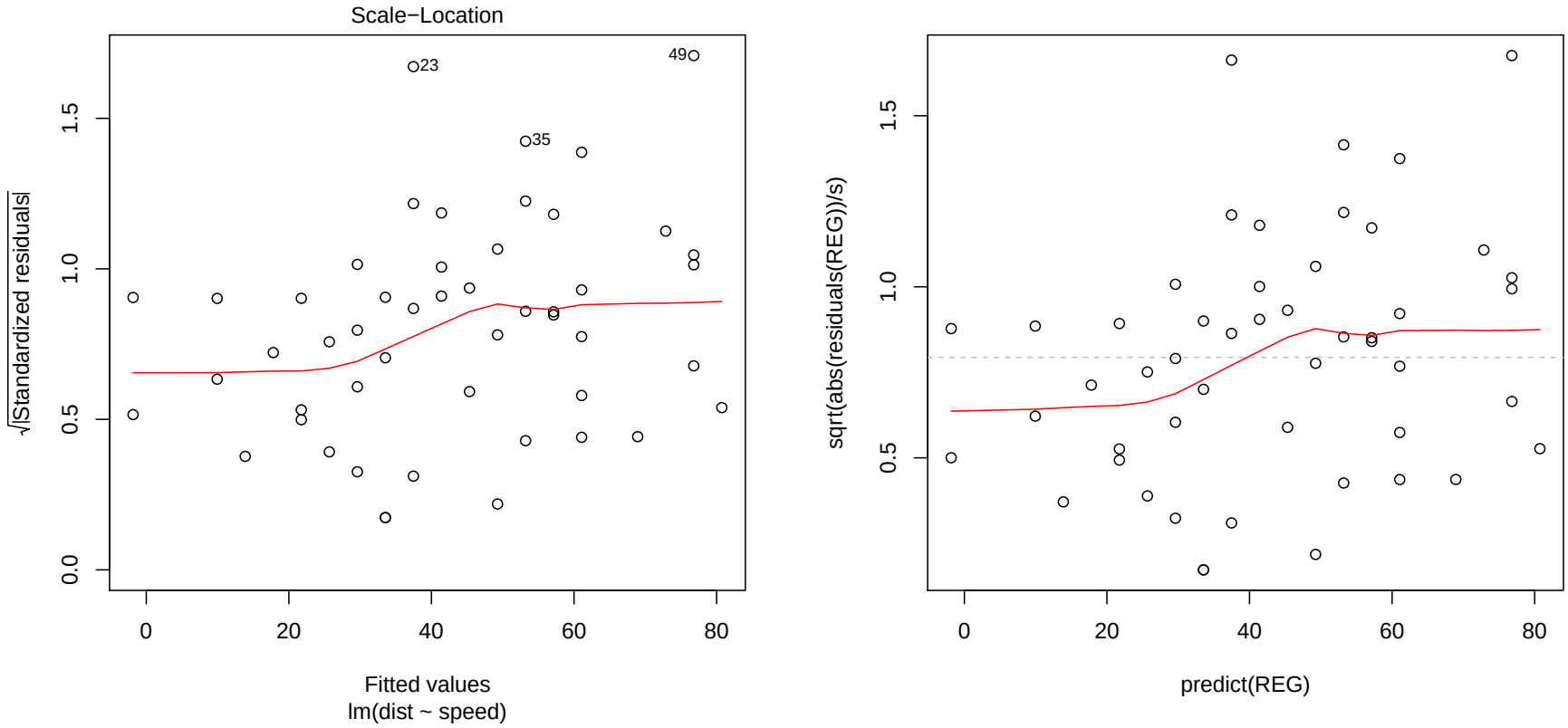
```
> s=summary(REG)$sigma
> plot(qnorm((1:50)/51),sort(residuals(REG))/s)
> abline(a=0,b=1,lty=2,col="grey")
> lines(lowess(qnorm((1:50)/51),sort(residuals(REG))/s),col="red")
```





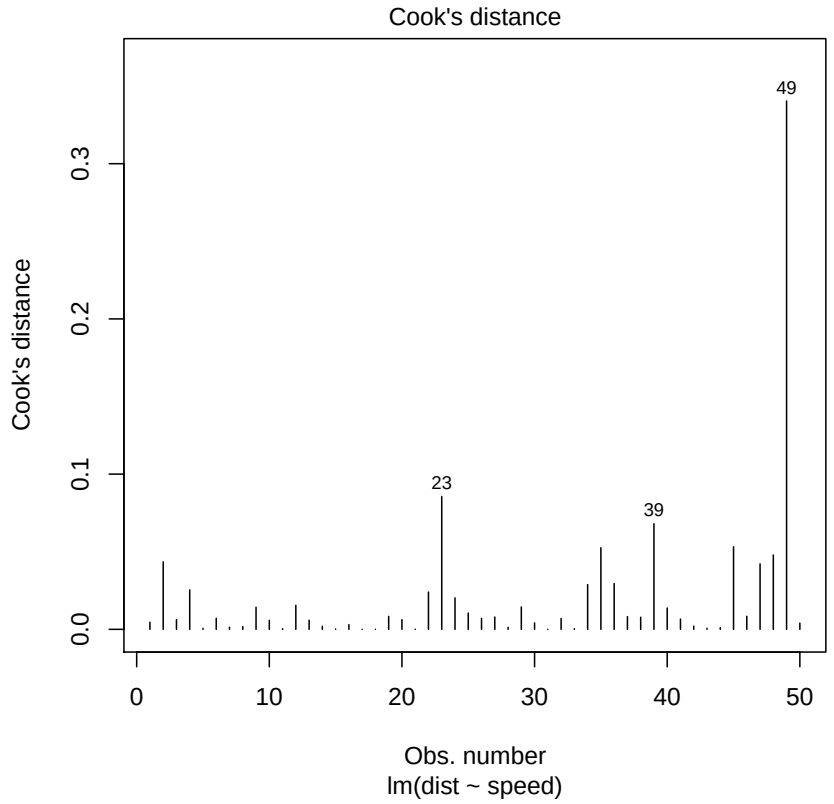
Diagnostique dans le modèle linéaire

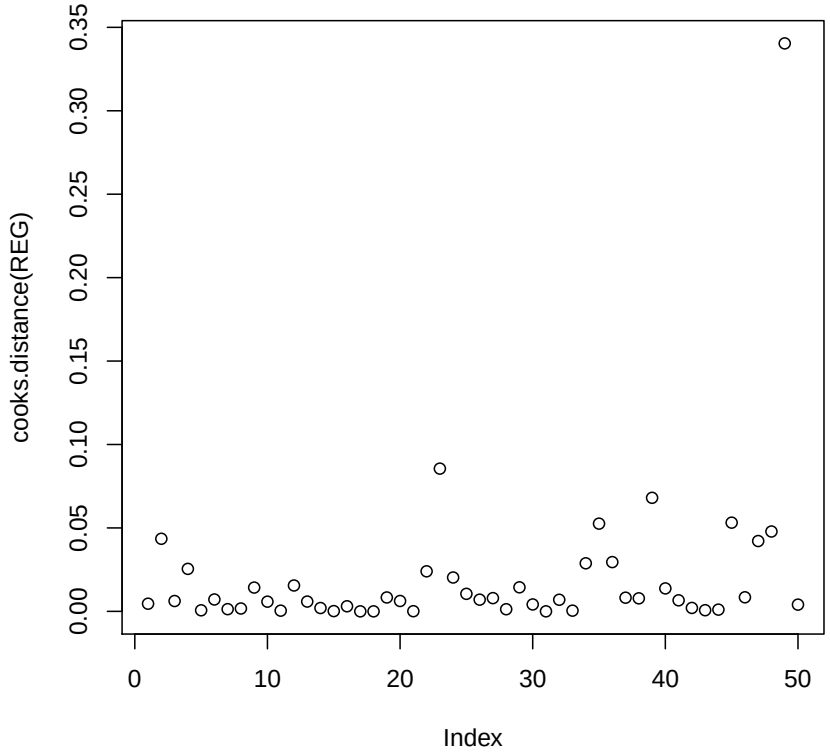
```
> plot(predict(REG),sqrt(abs(residuals(REG))/s))
> abline(h=mean(sqrt(abs(residuals(REG))/s)),lty=2,col="grey")
> lines(lowess(predict(REG),sqrt(abs(residuals(REG))/s)),col="red")
>
```



Diagnostique dans le modèle linéaire

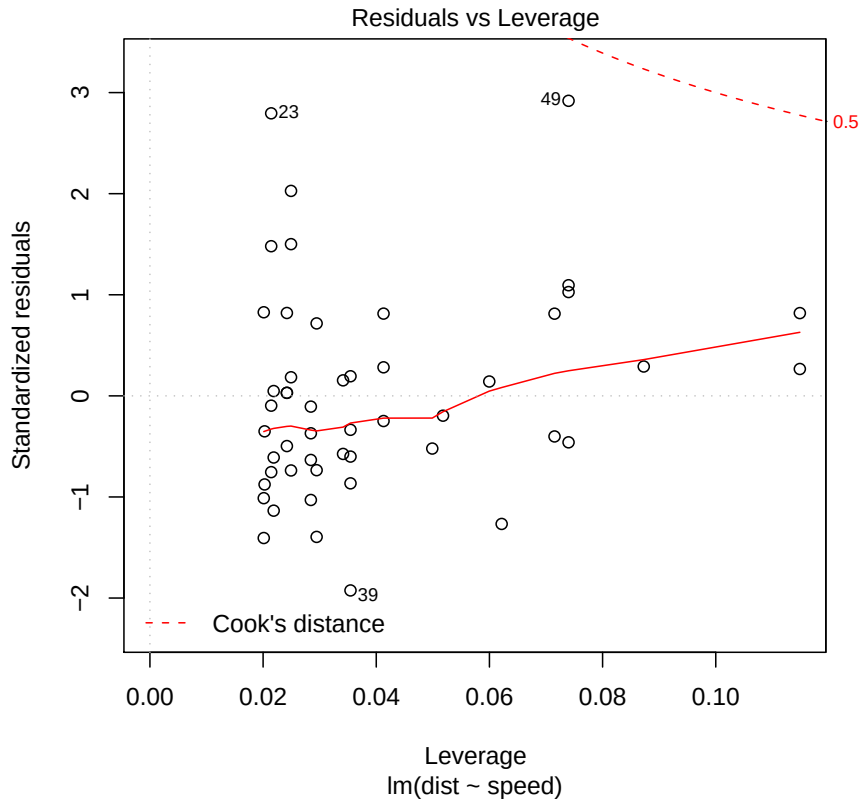
```
> library(car)
> plot(cooks.distance(REG))
>
>
```

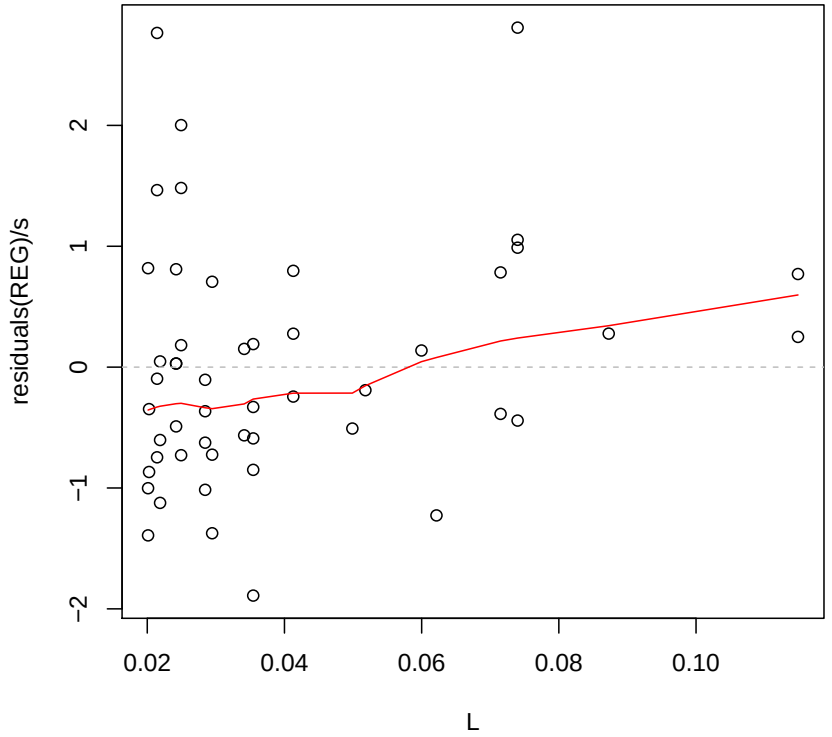




Diagnostique dans le modèle linéaire

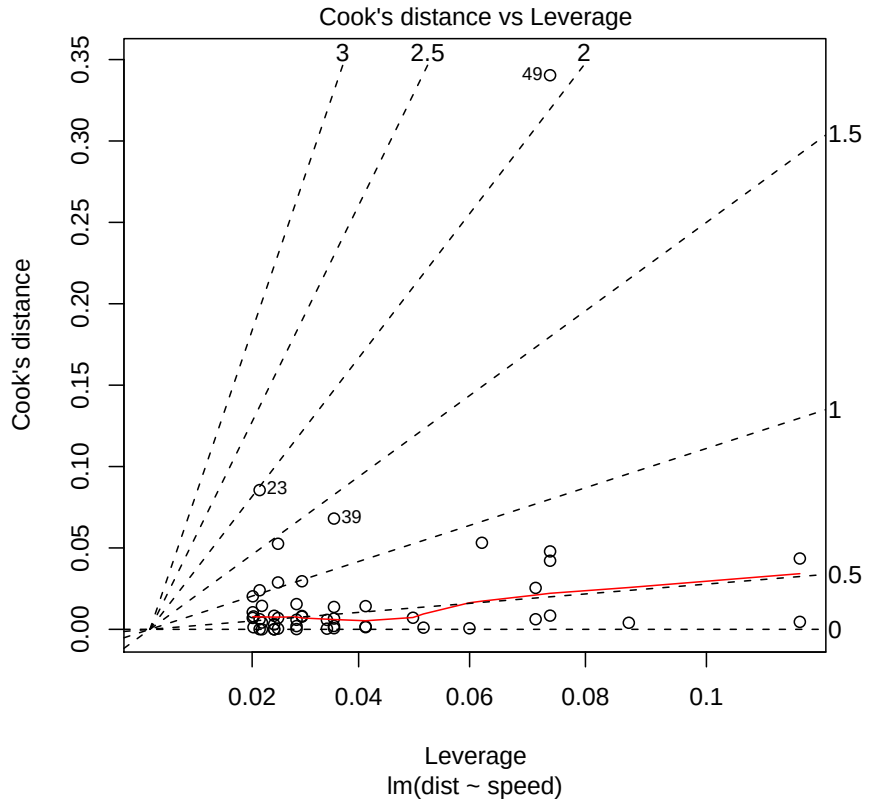
```
> X=cbind(1,cars$speed); L=diag(X%*%solve(t(X)%*%X)%*%t(X))
> plot(L,residuals(REG)/s)
> abline(h=0,lty=2,col="grey")
> lines(lowess(L,residuals(REG)/s),col="red")
```

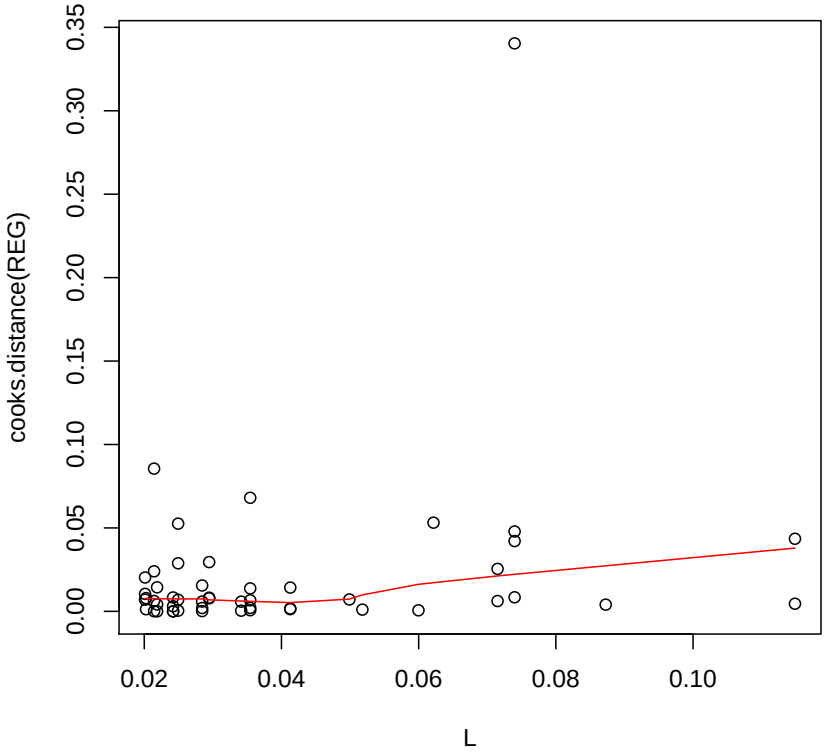




Diagnostique dans le modèle linéaire

```
> plot(L,cooks.distance(REG))
> lines(lowess(L,cooks.distance(REG)),col="red")
>
>
```





Les points atypiques et influents

La notion d'outliers ou de points abérants.

La distance de Cook mesure l'impact sur la régression de l'absence d'une observation. Aussi

$$C_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \cdot MSE}$$

ou encore,

$$C_i = \frac{\widehat{\varepsilon}_i^2}{p \cdot MSE}\left[\frac{h_{ii}}{(1-h_{ii})^2}\right]$$

où $h_{i,i}$ est lélément diagonale de la matrice $H = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ (que l'on notera parfois h_i). Les $h_i = [\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}]_{i,i} = H_{i,i}$ sont appelés (leverage).

Le vecteur des leverages $\boldsymbol{h} = (h_1, \cdots, h_n)$ est obtenu aisément sous R,

```
> diag(X%*%solve(t(X)%*%X)%*%t(X))[1:6]
 [1] 0.11486131 0.11486131 0.07150365 0.07150365 0.05997080 0.04989781
 > influence(REG)[1:6]
$hat
        1         2         3         4         5         6
0.11486131 0.11486131 0.07150365 0.07150365 0.05997080 0.04989781
```

Les hypothèses sont que $\mathbb{E}(\varepsilon_i) = 0$ et $Var(\varepsilon_i) = \sigma^2$. En réalité, $\mathbb{E}(\widehat{\varepsilon}_i) = 0$ mais $Var(\widehat{\varepsilon}_i) = [\mathbb{I} - H]_{i,i}\sigma^2 \neq \sigma^2$

Notons que puisque $Y = HY + \varepsilon$,

$$\widehat{\varepsilon} = Y - \widehat{Y} = [\mathbb{I} - H]Y = [\mathbb{I} - H](\boldsymbol{X}\boldsymbol{\beta} + \varepsilon) = [\mathbb{I} - H]\varepsilon,$$

et donc $Var(\widehat{\varepsilon}) = Var([\mathbb{I} - H]\varepsilon) = [\mathbb{I} - H]\sigma^2$. Aussi, $Var(\widehat{\varepsilon}_i) = [1 - h_i]\sigma^2$.

Les résidus Studentisés sont les

$$\widetilde{\varepsilon}_i = \frac{\widehat{\varepsilon}}{\widehat{\sigma}\sqrt{1 - h_i}}.$$

Notons que $\text{Var}(\widetilde{\varepsilon}_i) = 1$.

```
> diag(X%*%solve(t(X)%*%X)%*%t(X))
> rstudent(REG)[1:6]
          1           2           3           4           5           6
 0.26345000  0.81607841 -0.39781154  0.81035256  0.14070334 -0.51716052
> mean(rstudent(REG))
[1] 0.01347908
> sd(rstudent(REG))
[1] 1.045681
```

Sur la matrice de leverage (matrice de projection orthogonale), notons que

$$\widehat{Y}_i = HY = h_{i,i}Y_i + \sum_{j \neq i} h_{i,j}Y_j.$$

Aussi, $h_{i,i}$ est le poids accordé à Y_i pour sa propre prédiction.

- si $h_{i,i} = 1$, $\widehat{Y}_i$ est uniquement déterminé par Y_i ($h_{i,j} = 0$ pour $j \neq i$),
- si $h_{i,i} = 0$, $\widehat{Y}_i$ est n'est nullement influencé par Y_i.

On parelera de point levier i si $h_{i,i}$ est *trop* grand, i.e.

- si $h_{i,i} > 2k/n$, d'après Hoaglin & Welsch (1978),
- si $h_{i,i} > 3k/n$ pour $k > 6$ et $n - k > 12$, d'après Welleman & Welsch (1981),
- si $h_{i,i} > 1/2$, d'après Huber (1981).

Cette méthode permet de détecter des points atypiques, ou plutôts des points influents.

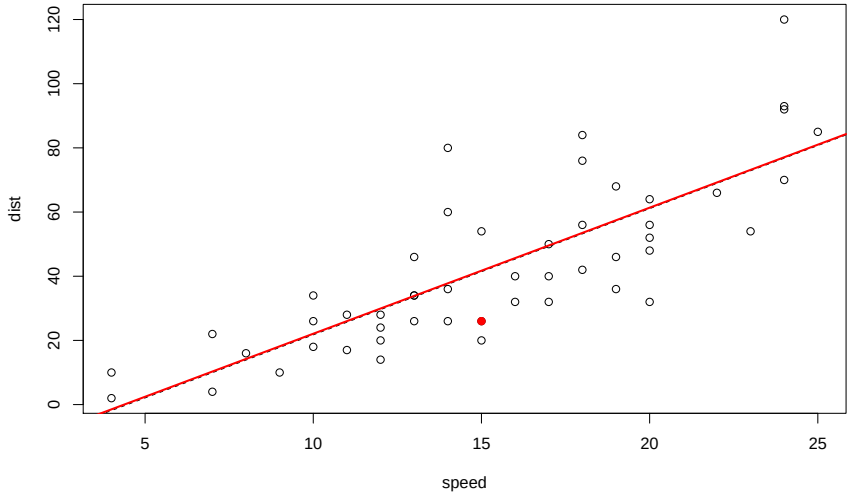
Afin de mesurer l'impact d'une observation sur la régression, il peut aussi être utile de regarder les résultats de la régression si l'on supprime une des observations.

Après suppression de la $i$ème observation, les estimateurs des moindres carrés s'écrivent

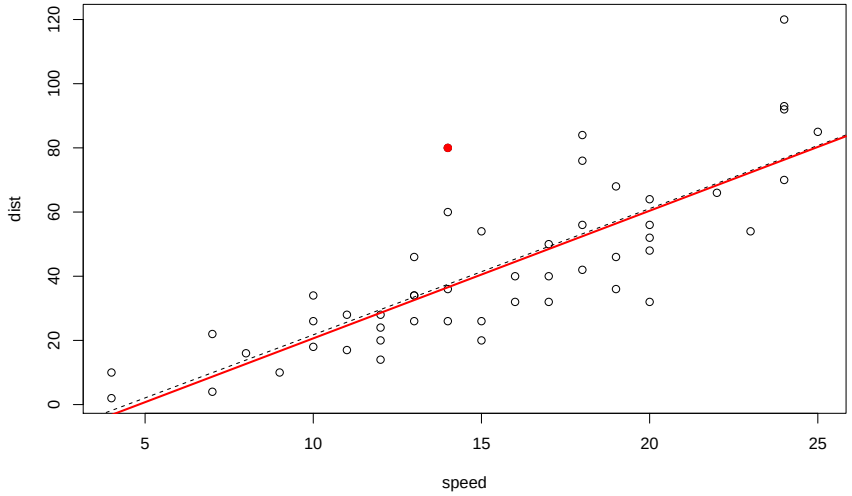
$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_i \cdot \frac{\widehat{\varepsilon}_i}{1-h_{i,i}}$$

$$\widehat{\sigma}^2_{(i)} = \frac{1}{n-k-1}\left((n-k)\widehat{\sigma}^2 \frac{\widehat{\varepsilon}_i^2}{1-h_{i,i}}\right)$$

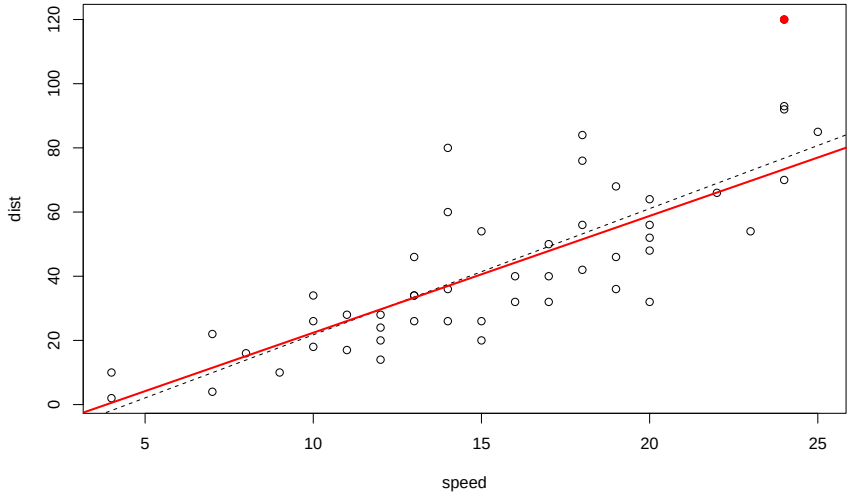
```
> i=25
> REGi=lm(dist~speed,data=cars[-i,])
> plot(cars); points(cars[i,],col="red",pch=19)
> abline(REG,lty=2); abline(REGi,lwd=2,col="red")
```



```
> i=23
> REGi=lm(dist~speed,data=cars[-i,])
> plot(cars); points(cars[i,],col="red",pch=19)
> abline(REG,lty=2); abline(REGi,lwd=2,col="red")
```



```
> i=49
> REGi=lm(dist~speed,data=cars[-i,])
> plot(cars); points(cars[i,],col="red",pch=19)
> abline(REG,lty=2); abline(REGi,lwd=2,col="red")
```



Remarque Beaucoup d'autres distances, basées sur la fonction d'influence, ont été proposées

$$\text{Cook} : C_i = \frac{(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})' \boldsymbol{X}' \boldsymbol{X} (\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})}{k\widehat{\sigma}^2}$$

$$\text{Welsh-Kuh} : WK_i = \frac{|\boldsymbol{X}_i'(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)})|}{\widehat{\sigma}^2_{(i)}\sqrt{h_{i,i}}}$$

$$\text{Welsh} : W_i = WK_i \sqrt{\frac{n-1}{1-h_{i,i}}}$$

$$\text{vraisemblance} : LD_i = 2\left(\mathcal{L}(\widehat{\boldsymbol{\beta}}, \widehat{\sigma}^2) - \mathcal{L}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\sigma}^2_{(i)}\right)\right)$$

Remarque : les points aberrants ont des valeurs de Y aberrantes, mais on pourrait aussi vouloir tester une abération en X.

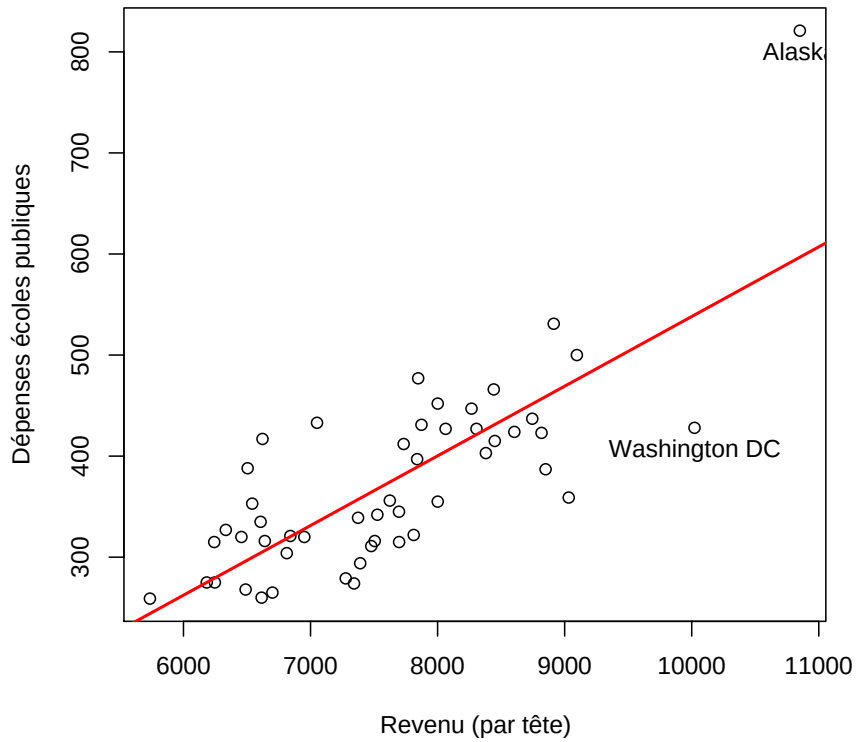
Analyse graphique des résidus, example

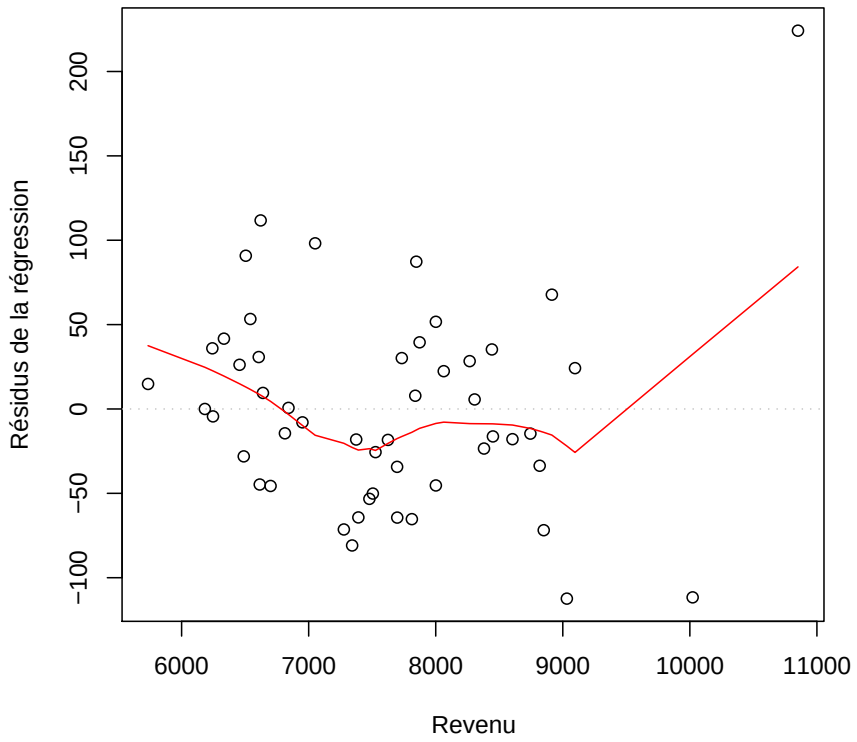
Pour illustrer, considérons les dépenses dans les écoles publiques, par état (aux U.S.A.)

```
> library(sandwich)
>  data(PublicSchools)
>
> tail(PublicSchools)
              Expenditure Income
Virginia              356   7624
Washington            415   8450
Washington DC         428  10022
West Virginia         320   6456
Wisconsin              NA   7597
Wyoming               500   9096
```

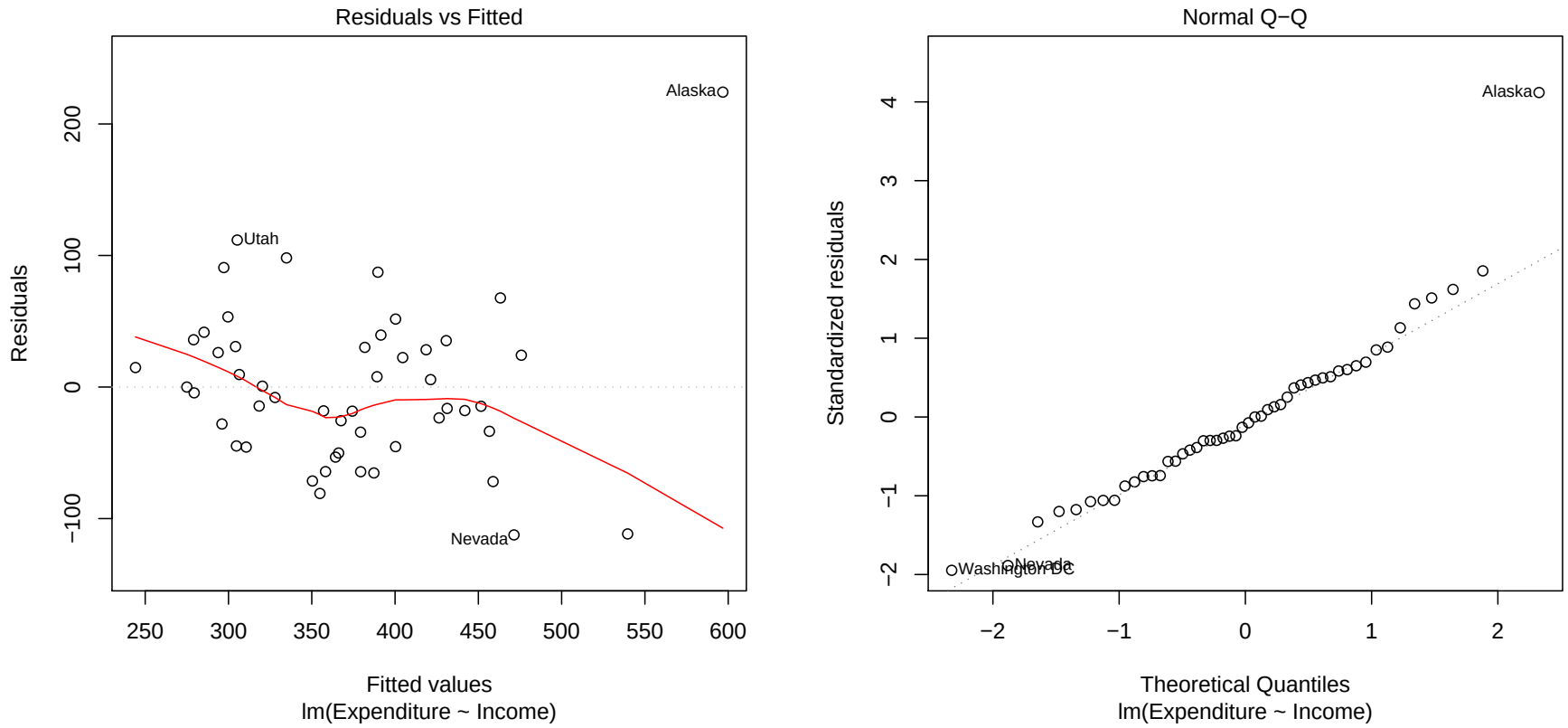
Analyse graphique des résidus, example

```
> plot(PublicSchools$Income,PublicSchools$Expenditure)
> id=c(2,49)
> text(PublicSchools$Income[id],PublicSchools$Expenditure[id],
+ rownames(PublicSchools)[id],pos=1)
```

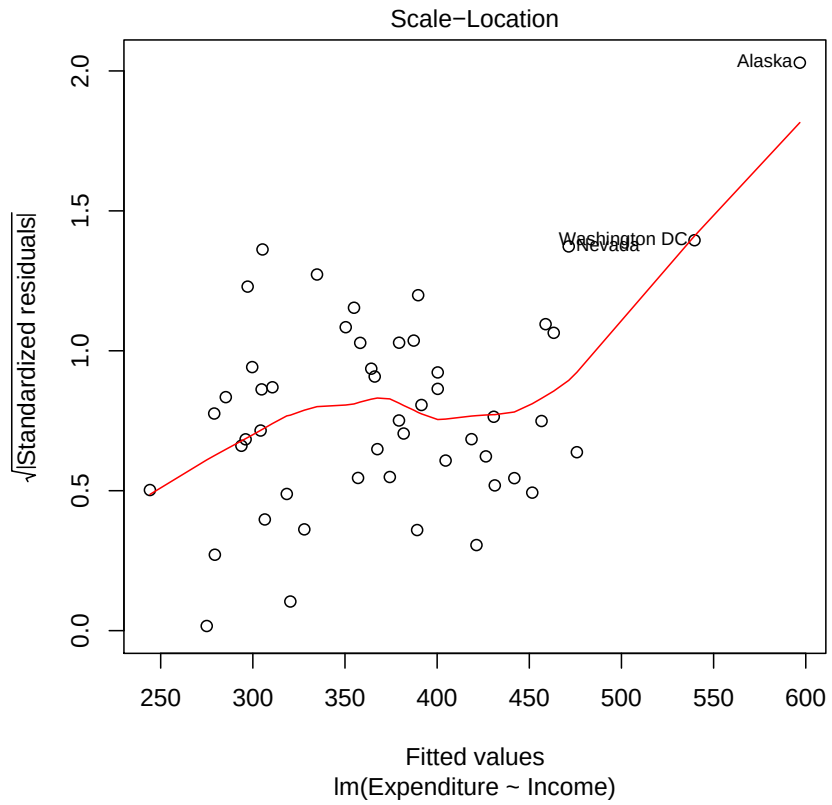


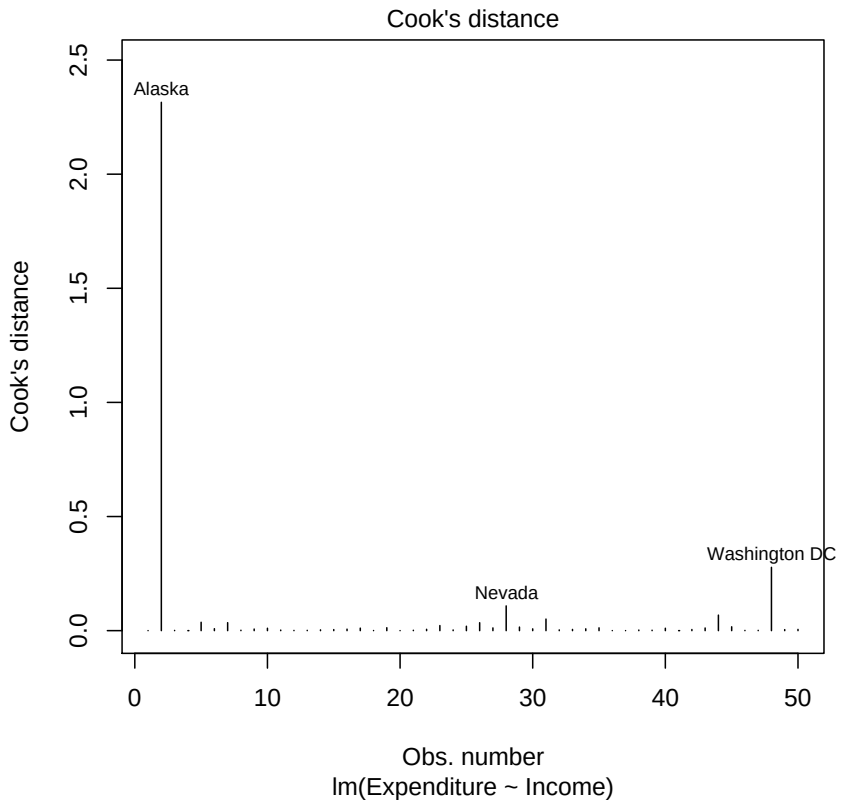


Analyse graphique des résidus, example

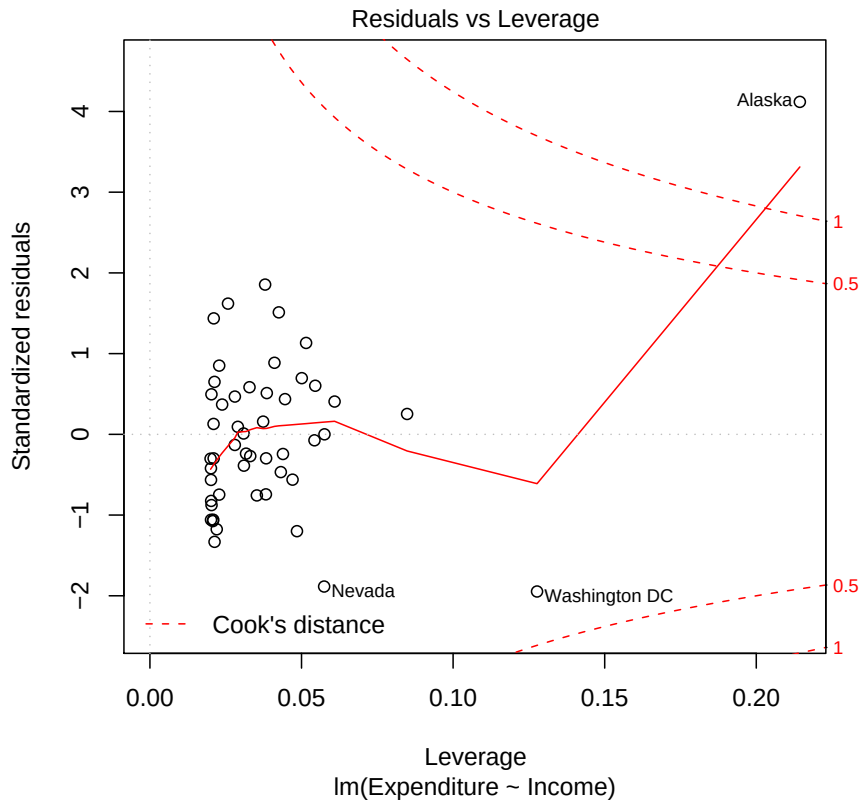


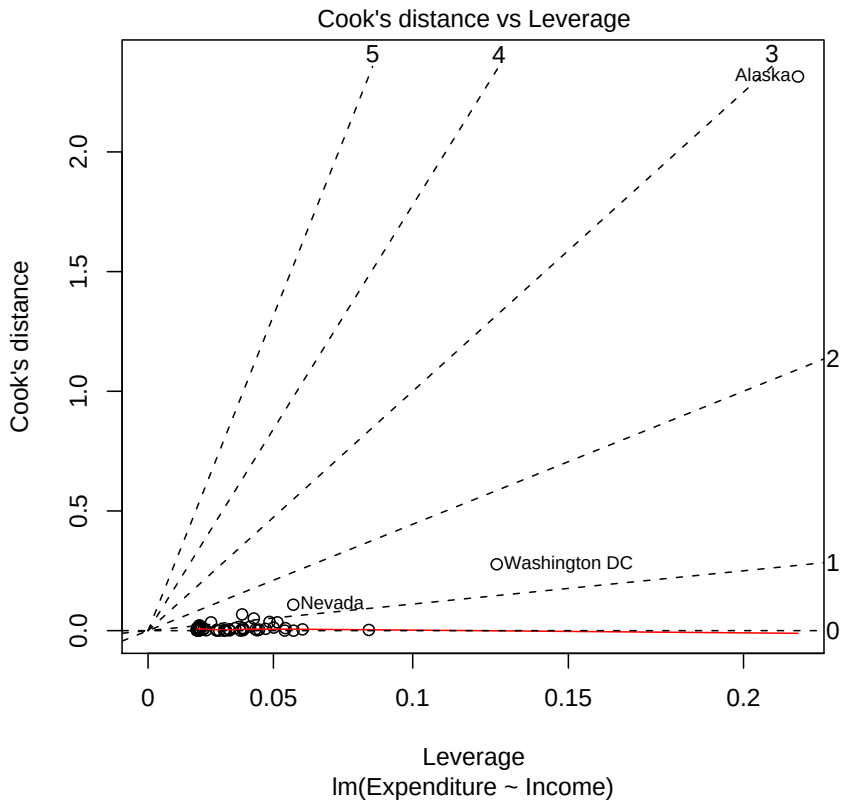
Analyse graphique des résidus, example





Analyse graphique des résidus, example





Analyse des résidus : quelles alternatives

Les hypothèses fondamentales sur les résidus sont

- $\mathbb{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \mathbf{0}$
- $\text{Var}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = \sigma^2\mathbb{I}$

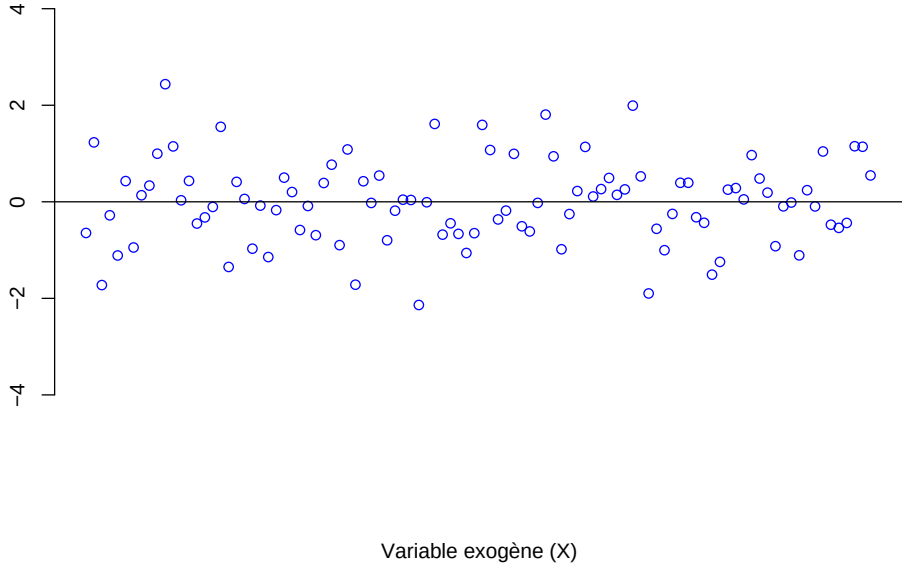
et éventuellement

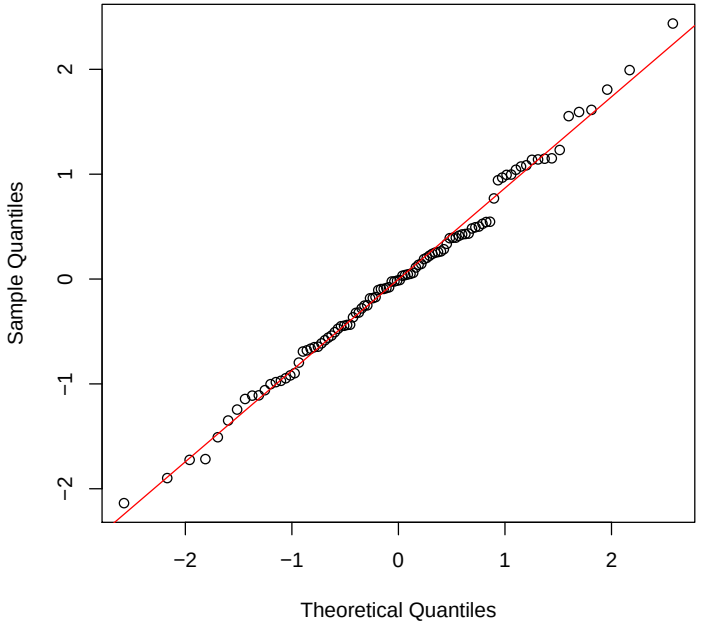
- $\boldsymbol{\varepsilon}|\boldsymbol{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbb{I})$

Un "*mauvais modèle* est un modèle pour lequel une de ces hypothèses n'est pas valide

Analyse graphique des résidus, suite

Graphique *normal* des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

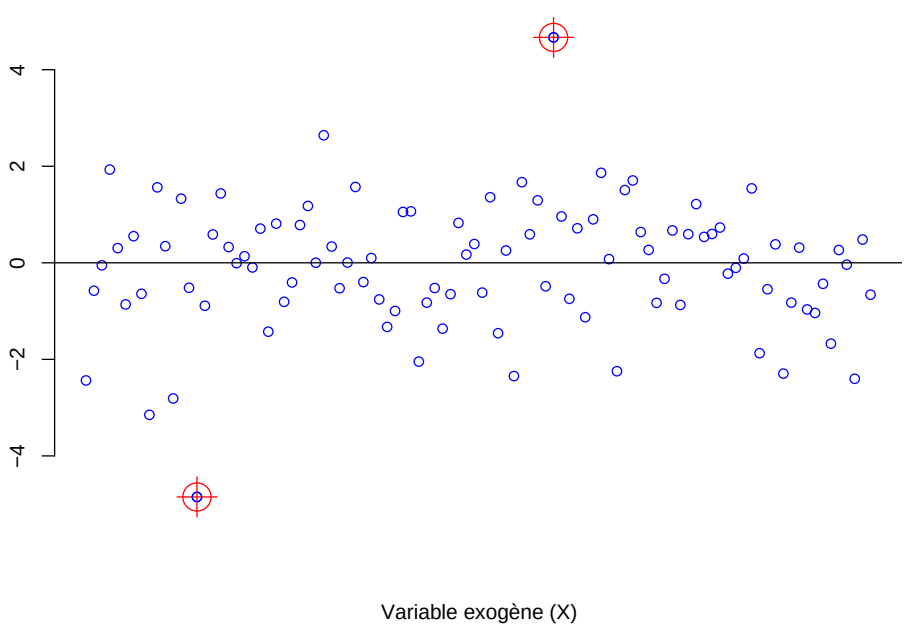


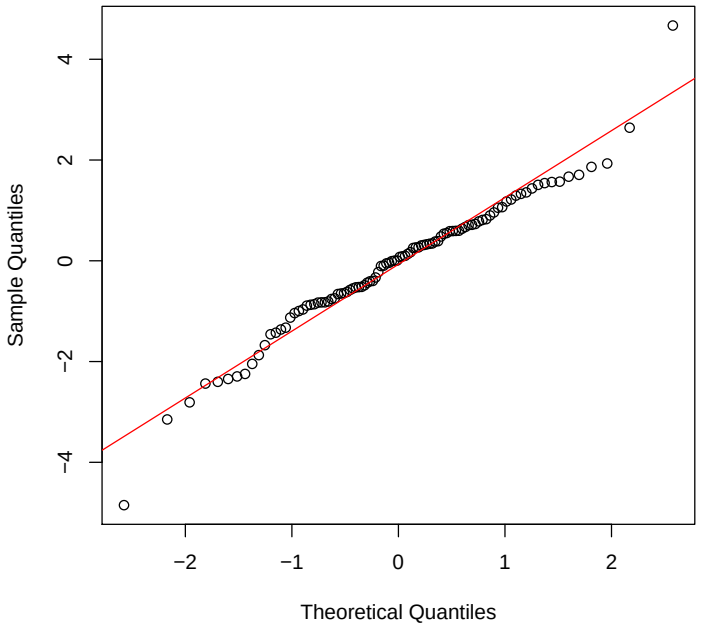


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

Présence de deux points atypiques,

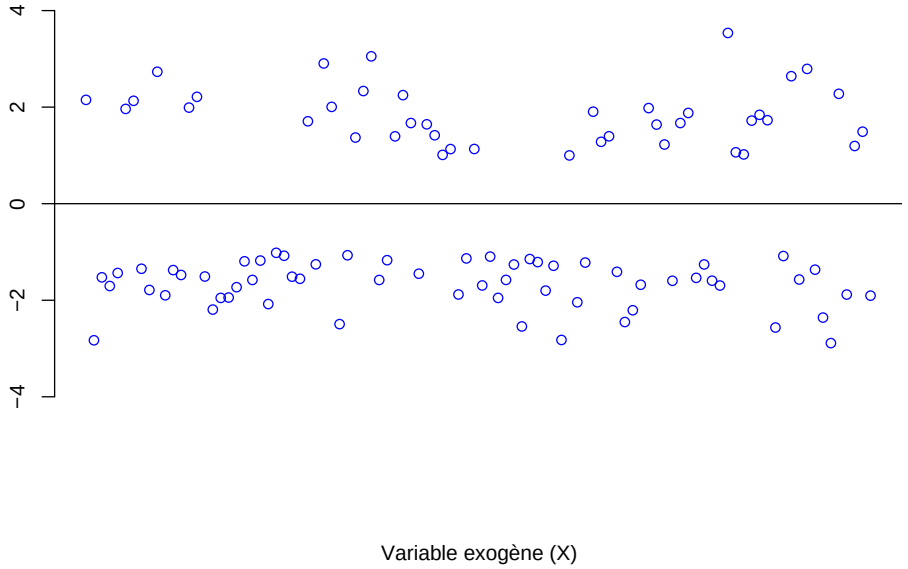


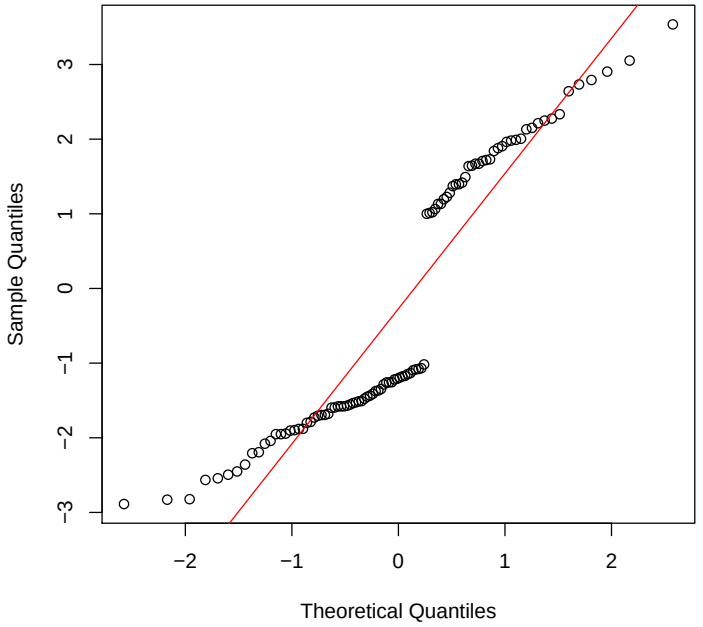


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

Distribution asymétrique des résidus, présence d'hétérogénéité (e.g. hommes/femmes)

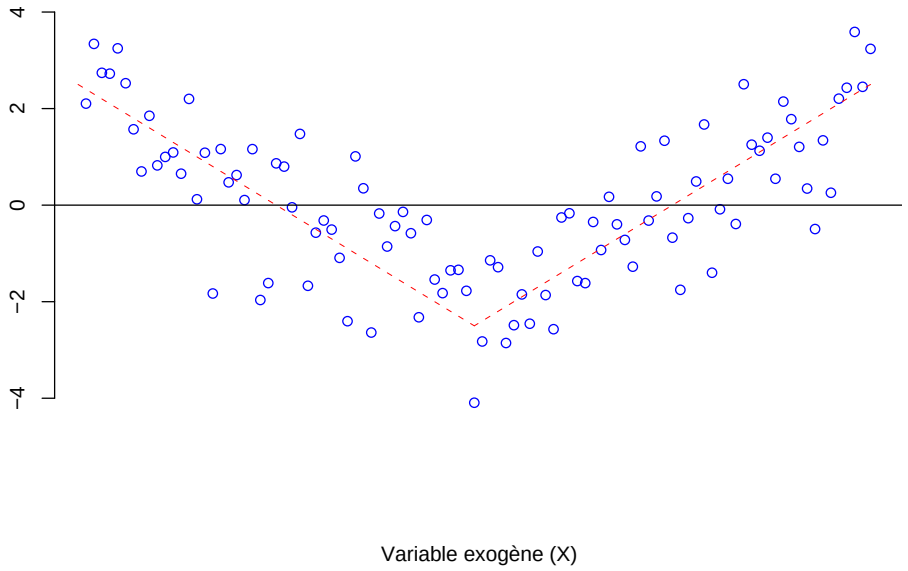


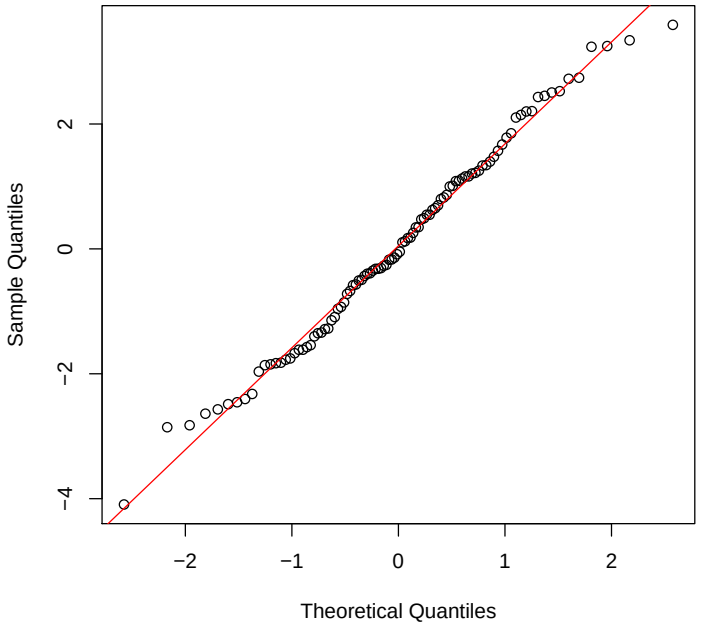


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

Présence d'une tendance, i.e. relation nonlinéaire pour X

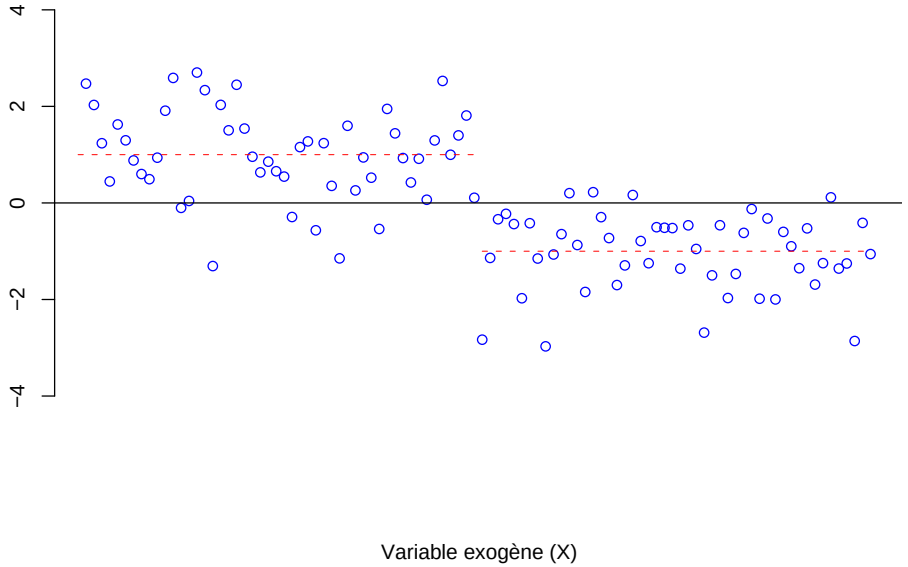


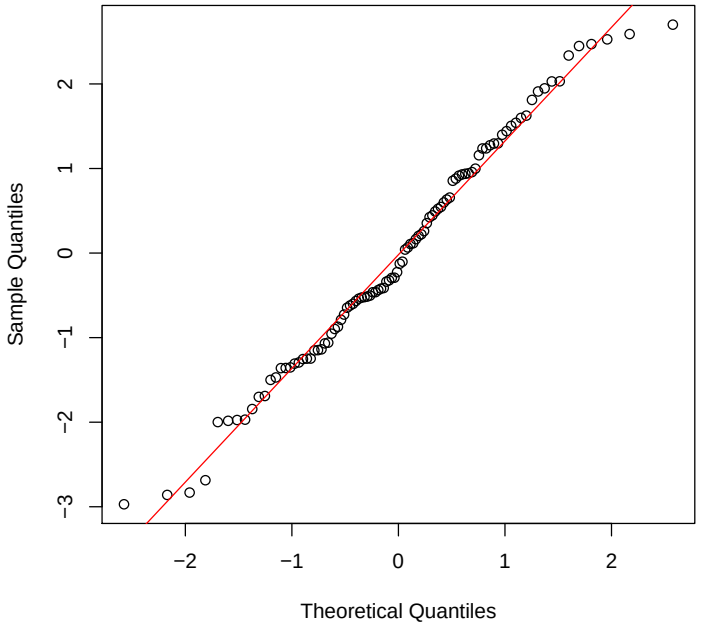


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon_i} = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

Rupture de la série, i.e. présence d'un seuil (relation non-linéaire en X, distinguer $X_i < u$ et $X_i > u$

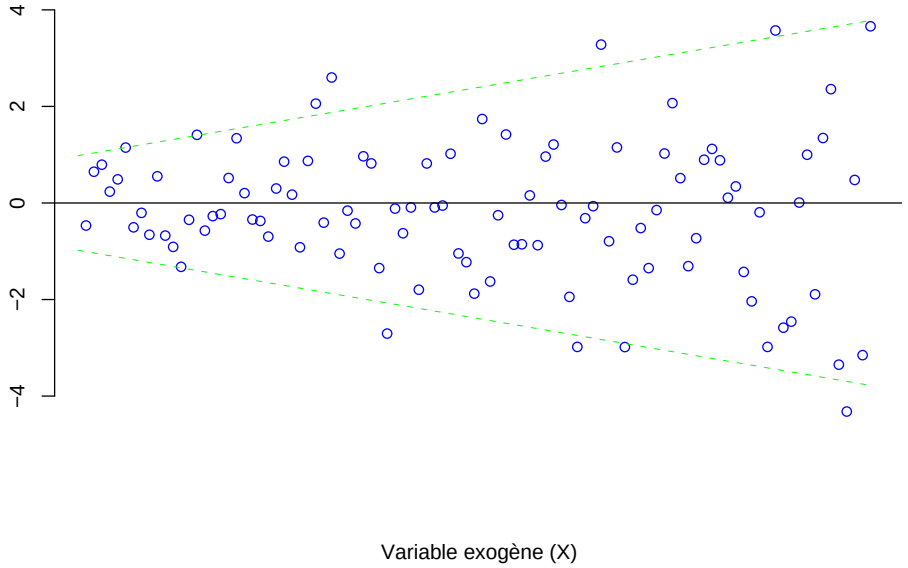


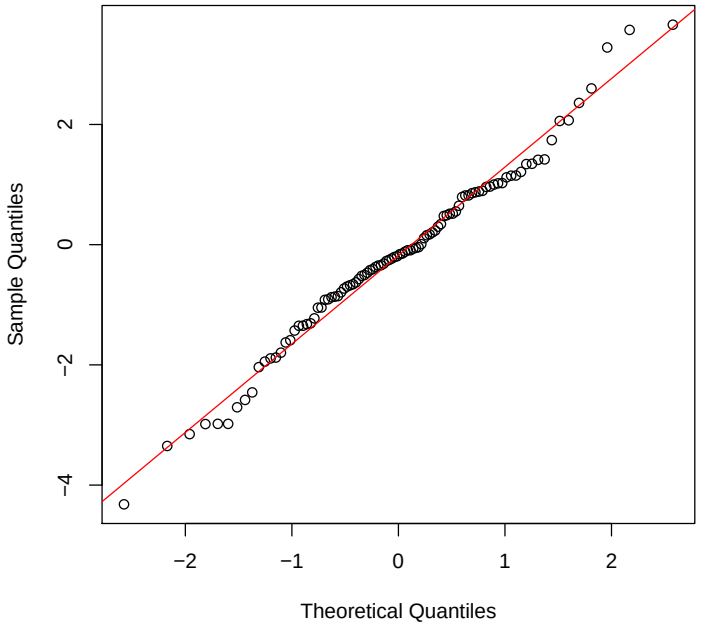


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

résidus autoscédastiques, la variance croît avec X, $Var(\varepsilon_i) = \sigma^2 X_i$ ou $Var(\varepsilon_i) = \sigma^2 X_i^2$, ...etc.

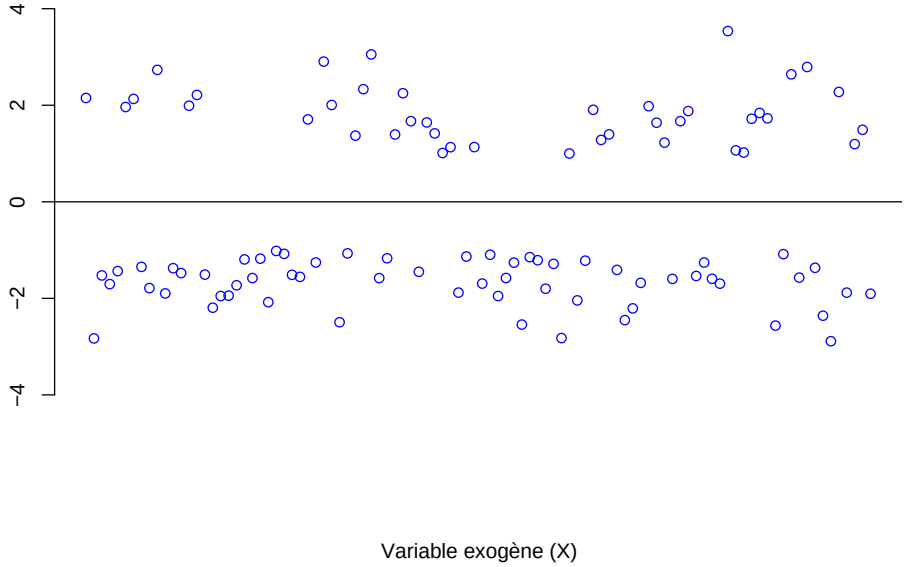


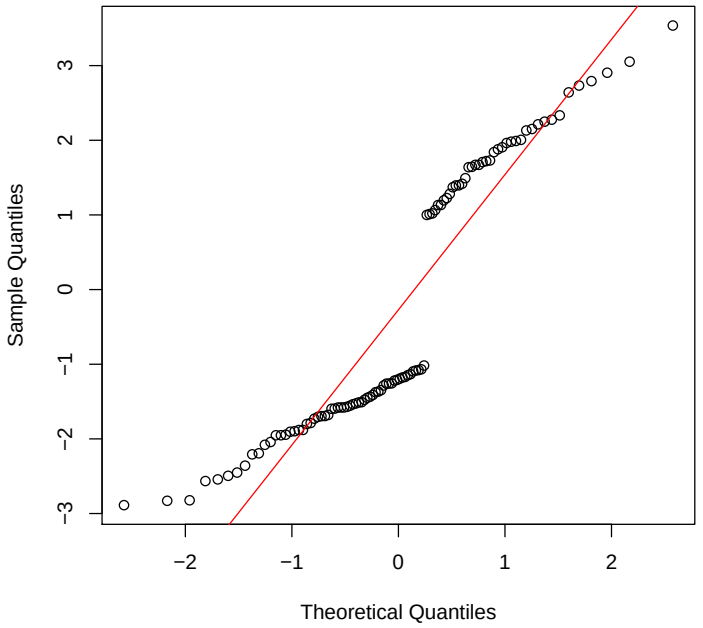


Analyse graphique des résidus, suite

Graphique des résidus $\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}$ versus X_i

résidus autocorrélés, ce qui arrive souvent avec des séries temporelles





Exemple : tester l'hypothèse d'homoscédasticité

Tester l'hypothèse d'homoscédasticité est délicat : il faut spécifier une alternative. Par exemple, au lieu d'avoir $\text{Var}(\varepsilon|X) = \sigma^2$ on pourrait avoir $\text{Var}(\varepsilon|X) = \sigma^2 h(\alpha_0 + \alpha_1 X^2)$ Le test de Breusch-Pagan revient à tester $H_0 : \alpha_1 = 0$ (homoscédasticité) contre $H_0 : \alpha_1 \neq 0$ (hétéroscédasticité).

L'idée est d'utiliser un test du score (appelé aussi test du multiplicateur de Lagrange), dans un modèle

$$Y = \beta_0 + \beta_1 X + \varepsilon \text{ où } \varepsilon^2 = \gamma_0 + \gamma_1 X + \eta.$$

La procédure générale est simple,

- faire la régression $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$
- sur $\widehat{\varepsilon}$, faire la régression $\widehat{\varepsilon}^2 = \gamma_0 + \gamma_1 X_1 + \cdots + \gamma_k X_k + \eta$
- tester si $\gamma_1 = \cdots = \gamma_k = 0$, i.e. test de significativité globale de Fisher, i.e. $T = nR^2$ qui suit (sous H_0) une loi $\chi^2(k)$

```
> library(lmtest)
> bptest(REG)
```

```
studentized Breusch-Pagan test

data:  model1
BP = 10.2903, df = 8, p-value = 0.2452
```

On peut le faire à la main pour vérifier

```
> E=residuals(REG)
> REG2=lm(E^2~cars$speed)
> summary(REG2)$r.squared*50
[1] 3.21488
> 1-pchisq(summary(REG2)$r.squared*50,1)
[1] 0.07297155
> summary(REG2)

Call:
lm(formula = E^2 ~ cars$speed)

Residuals:
    Min      1Q  Median      3Q     Max
-388.45 -175.07  -96.03   22.56 1607.53
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -61.05     167.56  -0.364   0.7172
cars$speed     18.71      10.30   1.816   0.0756 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 381.3 on 48 degrees of freedom
Multiple R-squared: 0.0643,        Adjusted R-squared: 0.0448
F-statistic: 3.298 on 1 and 48 DF,  p-value: 0.0756
```

Remarque On reviendra par la suite sur les modèle permettant de prendre en compte de l'hétéroscédasticité.

Tester l'indépendance des résidus

Tester l'indépendance des résidus nécessite d'expliciter la forme de l'alternative. Par exemple, au lieu d'avoir $\text{cor}(\varepsilon_i, \varepsilon_{i+1}) = 0$ on pourrait avoir $\text{cor}(\varepsilon_i, \varepsilon_{i+1}) = r$ $(\neq 0)$.

La statistique de Durbin-Watson est basée sur

$$DW = \frac{\sum(\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i-1})^2}{\sum(\widehat{\varepsilon}_i)^2}$$

qui est dans le package `lmtest`, via la fonction `dwtest`.

```
> library(lmtest)
> REG=lm(dist~speed,data=cars)
> dwtest(REG)

        Durbin-Watson test

data:  reg
DW = 1.6762, p-value = 0.09522
alternative hypothesis: true autocorrelation is greater than 0
```

Les résidus sont ici ordonnés comme dans la base. On pourrait suspecter un lien avec les X. Dans ce cas, on peut faire une régression sur la base réordonnée.

```
>  indice=rank(cars$speed,ties.method="random")
>  reg=lm(dist~speed,data=cars[indice,])
>  dwtest(reg)

        Durbin-Watson test

data:  reg
DW = 1.7993, p-value = 0.1935
alternative hypothesis: true autocorrelation is greater than 0
```

On va accepter ici l'hypothèse d'absence d'autocorrélation des résidus.

Tester l'indépendance spatiale des résidus

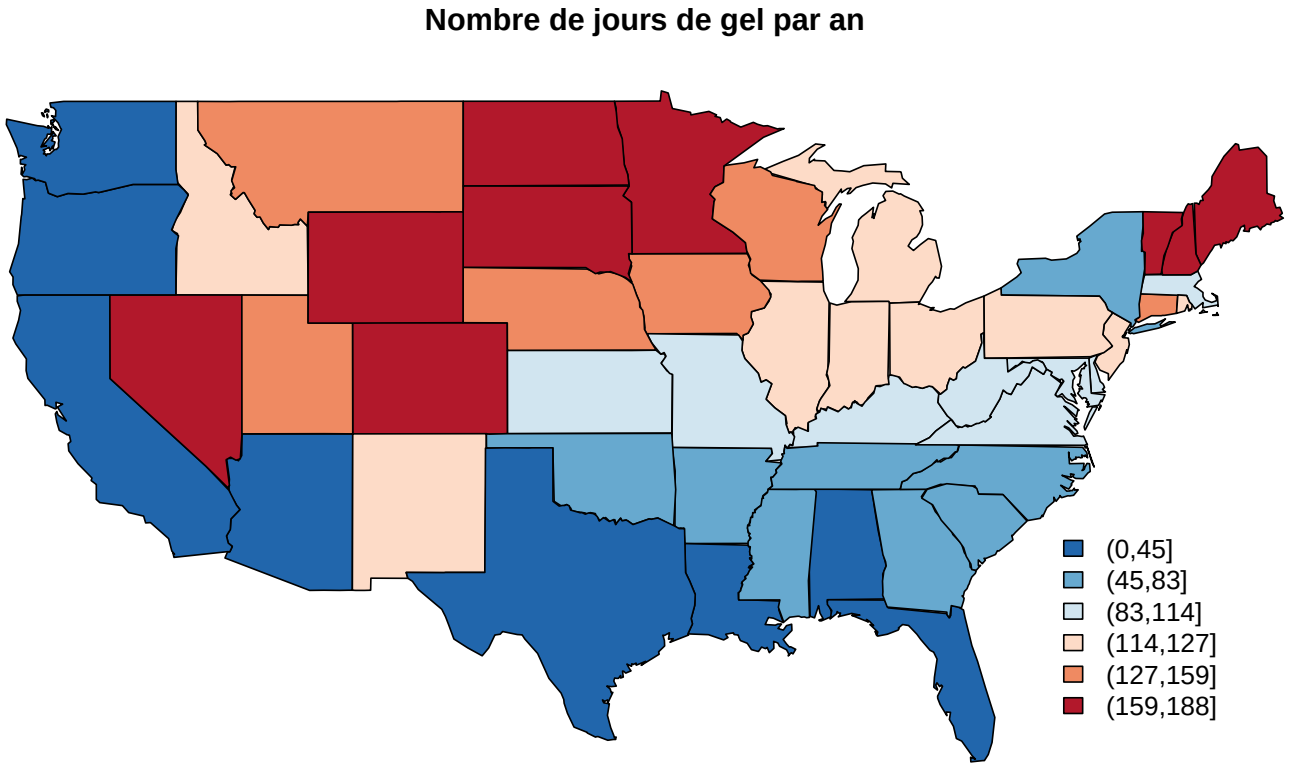
Pour des données spatiales, on devra se contenter de regarder si, visuellement, l'indépendance spatiale est une hypothèse qui semble valide,

```
> US=read.table("http://freakonometrics.free.fr/US.txt",
+ header=TRUE,sep=";")
> abreviation=read.table(
+ "http://freakonometrics.blog.free.fr/public/data/etatus.csv",
> header=TRUE,sep=",")
> US$USPS=rownames(US)
> US=merge(US,abreviation)
> US$nom=tolower(US$NOM)
> library(maps)
> VL0=strsplit(map("state")$names,":")
> VL=VL0[[1]]
> for(i in 2:length(VL0)){VL=c(VL,VL0[[i]][1])}
> ETAT=match(VL,US$nom)
```

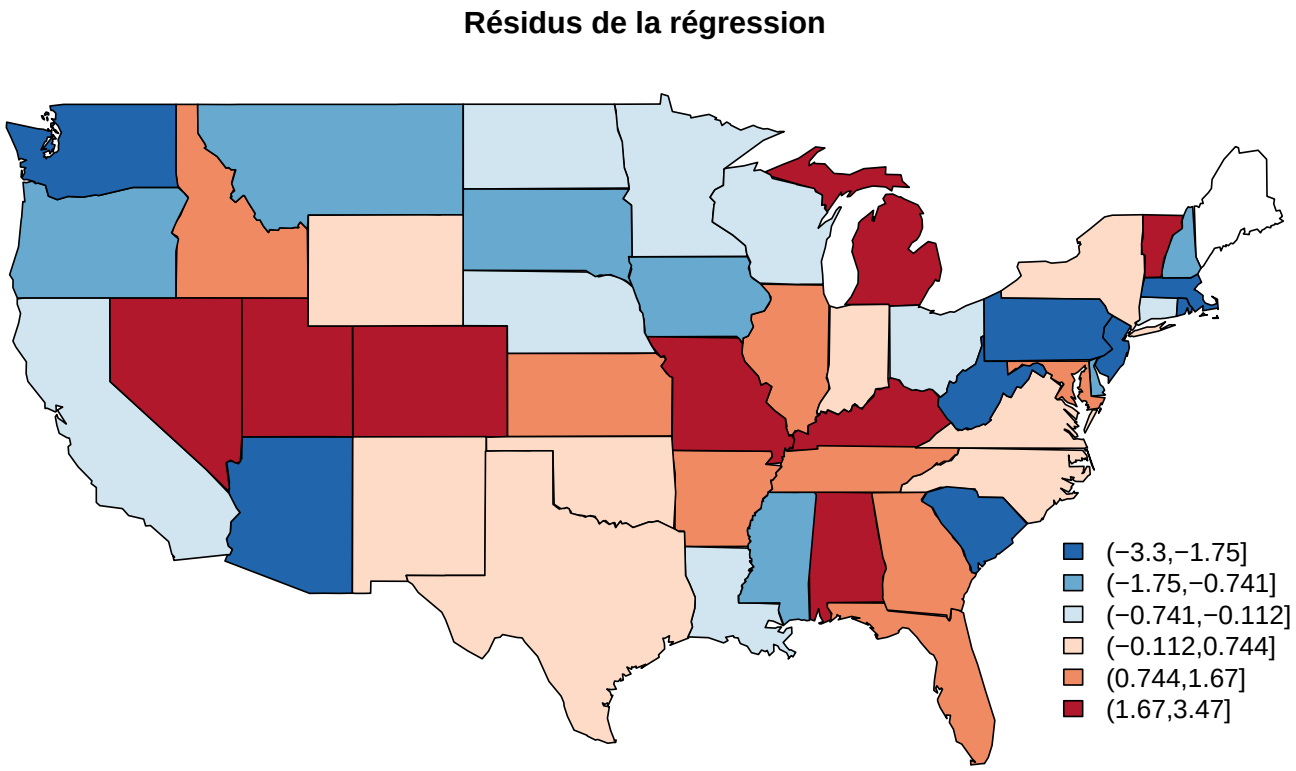
On peut faire une fonction mettant de la couleur appropriée dans chacune des états,

```
> library(RColorBrewer)
> carte=function(V=US$Murder,titre=
+ "Taux d'homicides aux Etats-Unis"){
+ variable=as.numeric(as.character(cut(V,
+ quantile(V,seq(0,1,by=1/6)),labels=1:6)))
+ niveau=variable[ETAT]
+ couleur=rev(brewer.pal(6, "RdBu"))
+ noml=levels(cut(V,quantile(V,seq(0,1,by=1/6))))
+ map("state", fill = TRUE, col=couleur[niveau]);
+ legend(-78,34,legend=noml,fill=couleur,
+ cex=1,bty="n");
+ title(titre)}
```

```
>
>
> carte(US$Frost, titre="Nombre de jours de gel par an")
```



```
> reg=lm(Murder~.-NOM-USPS-nom,data=US)
> regs=step(reg)
> carte(residuals(regs), titre="Rsidus de la rgression")
```



Choix de modèle, AIC et SIC/BIC

Le critère d'Akaike, noté souvent AIC

$$AIC = 2k - 2\log(\mathcal{L}) = 2k + n\left[\log\left(2\pi\frac{1}{n}\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\right) + 1\right]$$

dans le cas Gaussien.

Le critère de Schwarz, noté SIC (Schwarz Information Criterion), ou critère Bayésien, noté BIC

$$BIC = -2\log(\mathcal{L}) + k\ln(n) = n\ln\left(\sum_{i=1}^{n}\widehat{\varepsilon}_i^2\right) + k\ln(n).$$

```
> AIC(reg)
[1] 419.1569

> logLik(reg)
'log Lik.' -206.5784 (df=3)
```

Choix de modèle : sélection des variables

Parfois on dispose de beaucoup de variables explicatives, et on ne sait quel modèle choisir.

```
dodge=read.table("http://perso.univ-rennes1.fr/arthur.charpentier/dodge.csv",
header=TRUE,sep=",")
```

La variable `FIRE` correspond au nombre d'incendies (/1000 ménages) dans le quartier $i = 1, \cdots, 47$ de Chicago, en 1975. `X1` est la proportion d'habitations construites avant 1940, `X2` le nombre de vols commis, et `X3` le revenu médian du quartier.

Choix de modèle : sélection des variables

Les 8 modèles possibles sont les suivants

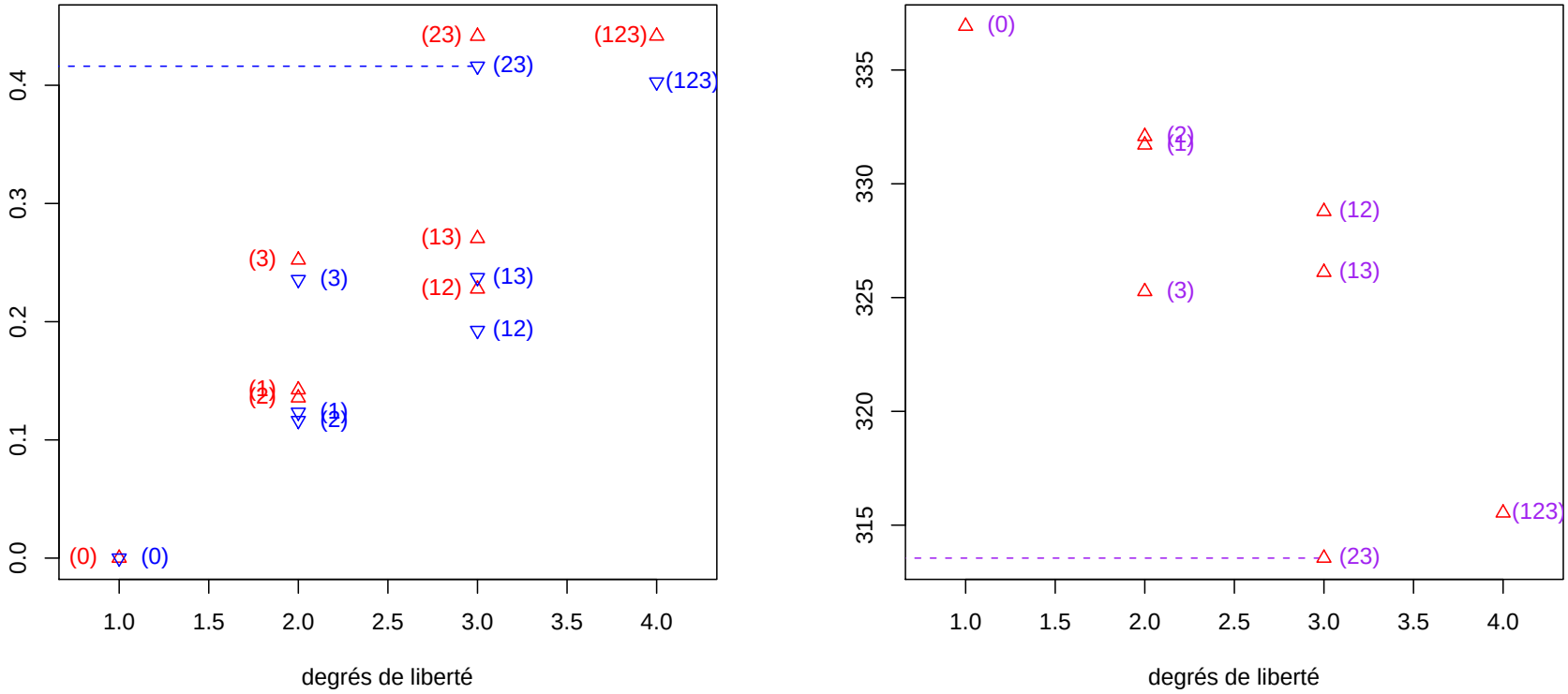
$$
\begin{aligned}
(0) \quad & Y = \beta_0 + \varepsilon \\
(1) \quad & Y = \beta_0 + \beta_1 X_1 \varepsilon \\
(2) \quad & Y = \beta_0 + \beta_2 X_2 \varepsilon \\
(3) \quad & Y = \beta_0 + \beta_3 X_3 \varepsilon \\
(12) \quad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \varepsilon \\
(13) \quad & Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 \varepsilon \\
(23) \quad & Y = \beta_0 + \beta_2 X_2 + \beta_3 X_3 \varepsilon \\
(123) \quad & Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \varepsilon
\end{aligned}
$$

Choix de modèle : sélection des variables

modèle	R^2	$\overline{R}^2$	$\log \mathcal{L}$	AIC	degrés liberté
(0)	0.0000	0.0000	-166.4638	336.9277	1
(1)	0.1423920	0.1233340	-162.8540537	331.7081074	2
(2)	0.1355551	0.1163452	-163.0406536	332.0813071	2
(3)	0.2522118	0.2355942	-159.6339128	325.2678255	2
(12)	0.2276727	0.1925669	-160.3926938	328.7853875	3
(13)	0.2703887	0.2372245	-159.0556280	326.1112560	3
(23)	0.4414889	0.4161020	-152.7755479	313.5510959	3
(123)	0.4416723	0.4027192	-152.7678305	315.5356609	4

Dans ce cas, le meilleur modèle est le modèle $Y = \beta_0 + \beta_2 X_2 + \beta_3 X3 + \varepsilon$, quel que soit le critère de choix.

Choix de modèle : sélection des variables



Une sélection automatique basée sur un minimisation du AIC peut être faite, soit backward à partir du modèle complet, soit forward à partir du modèle le plus simple.

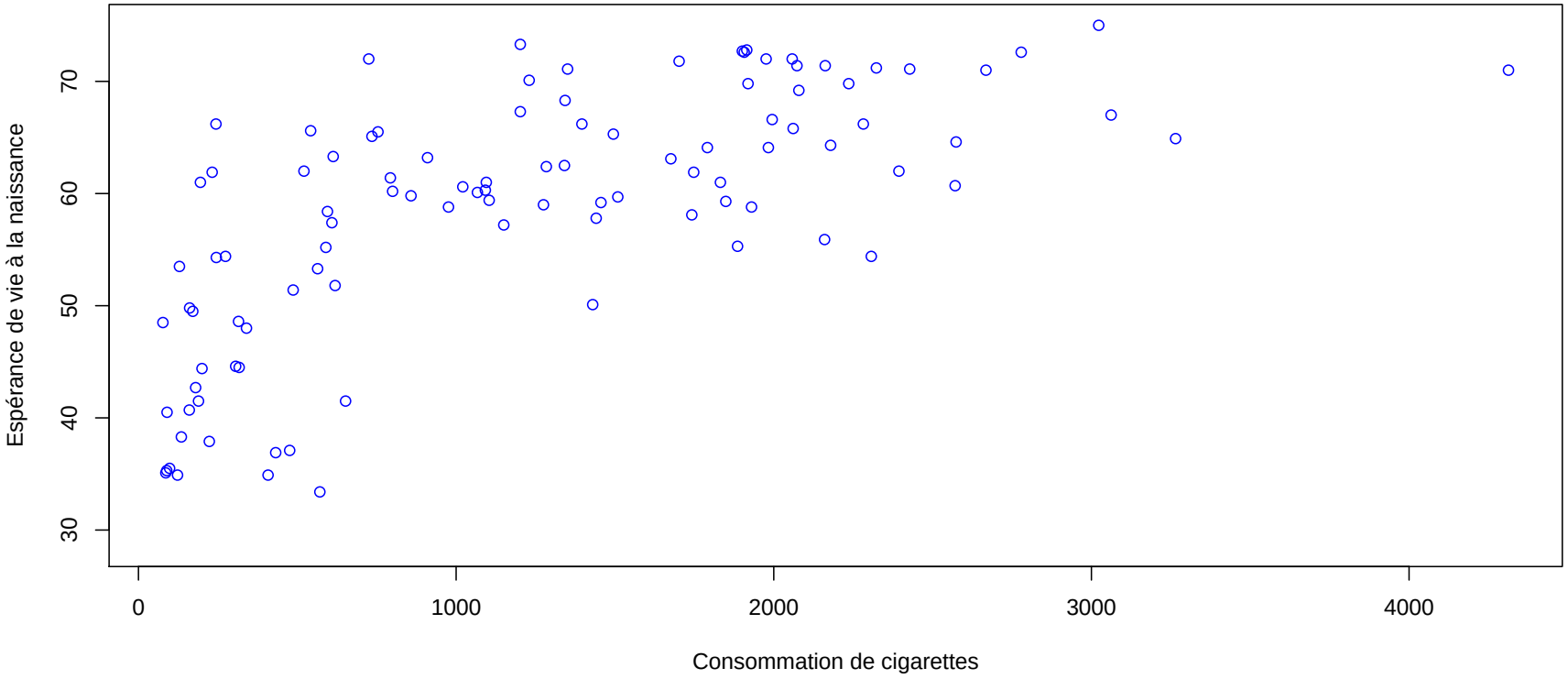
Choix de modèle : sélection des variables

```
> step(lm(Fire~.,data=D),direction = "backward")
Start:  AIC=180.16
Fire ~ X_1 + X_2 + X_3

       Df Sum of Sq     RSS     AIC
- X_1   1      0.60 1832.36  178.17
<none>              1831.75  180.16
- X_2   1    561.94 2393.70  190.73
- X_3   1    702.09 2533.84  193.41

Step:  AIC=178.17
Fire ~ X_2 + X_3

       Df Sum of Sq     RSS     AIC
<none>              1832.36  178.17
- X_2   1    620.98 2453.33  189.89
- X_3   1   1003.70 2836.06  196.70

Call:
```

```
lm(formula = Fire ~ X_2 + X_3, data = D)

Coefficients:
(Intercept)          X_2          X_3
    21.4965       0.2213      -1.5248
```

Exemple d'application : cigarettes et espérance de vie

Étudions ici le lien entre l'espérance de vie et la consommation de cigarette (par tête)



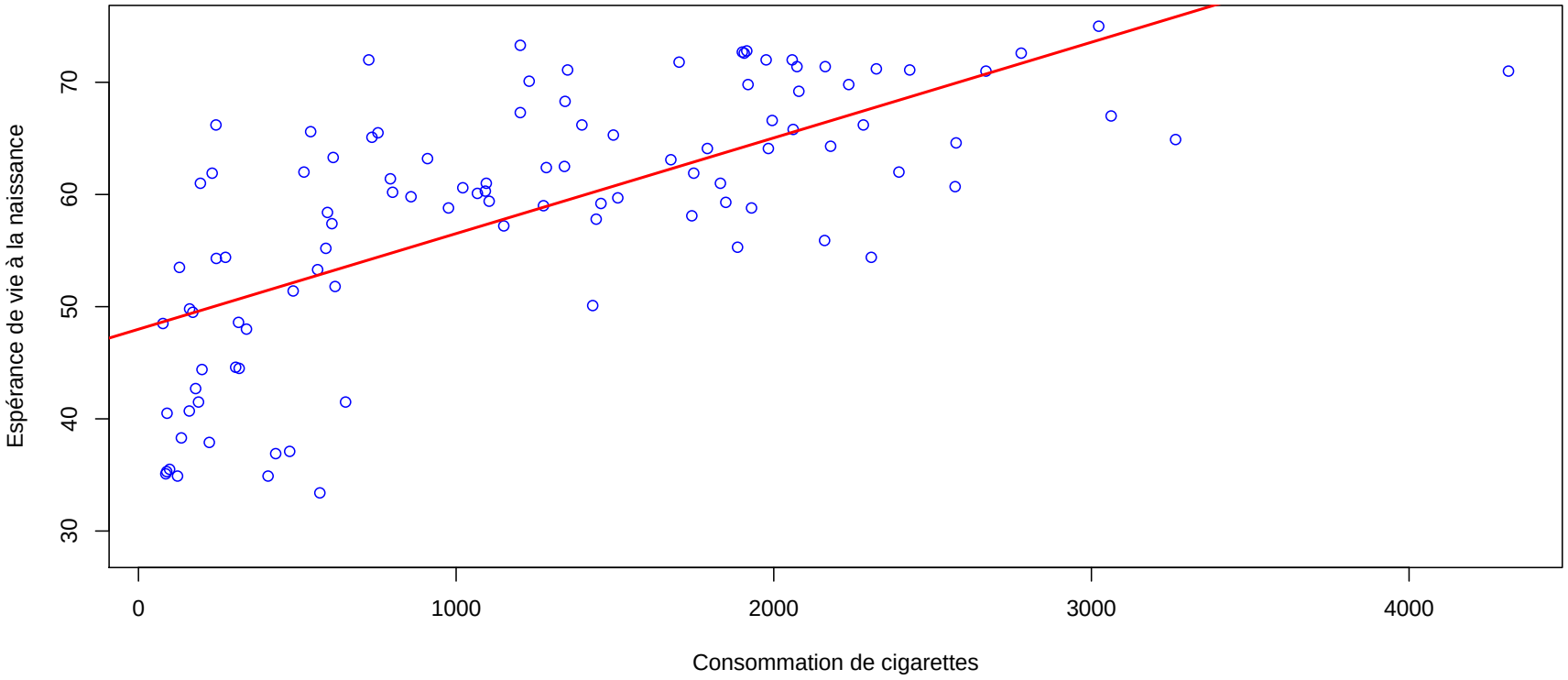
Exemple d'application : cigarettes et espérance de vie

```
> summary(lm(Y~X))

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.799e+01  1.371e+00  34.995  < 2e-16 ***
X           8.528e-03  9.007e-04   9.468 1.33e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.158 on 101 degrees of freedom
  (86 observations deleted due to missingness)
Multiple R-Squared: 0.4702,     Adjusted R-squared: 0.465
F-statistic: 89.64 on 1 and 101 DF,  p-value: 1.333e-15
```

Interprétation (rapide et sans doute falacieuse) : $\widehat{\beta}_1 \sim 0.00852$, i.e. en fumant une cigarette de plus par an, on augmente l'espérance de vie de 0.00852 années. Aussi, fumer 1 cigarette de plus (ou de moins) par jour augmente (ou diminue) l'espérance de vie de 3.11 années.

Exemple d'application : cigarettes et espérance de vie

Étudions ici le lien entre l'espérance de vie et la consommation de cigarette (par tête)



Exemple d'application : cigarettes et espérance de vie

L'interprétation est bien entendu fausse et vient du type de données utiliséées

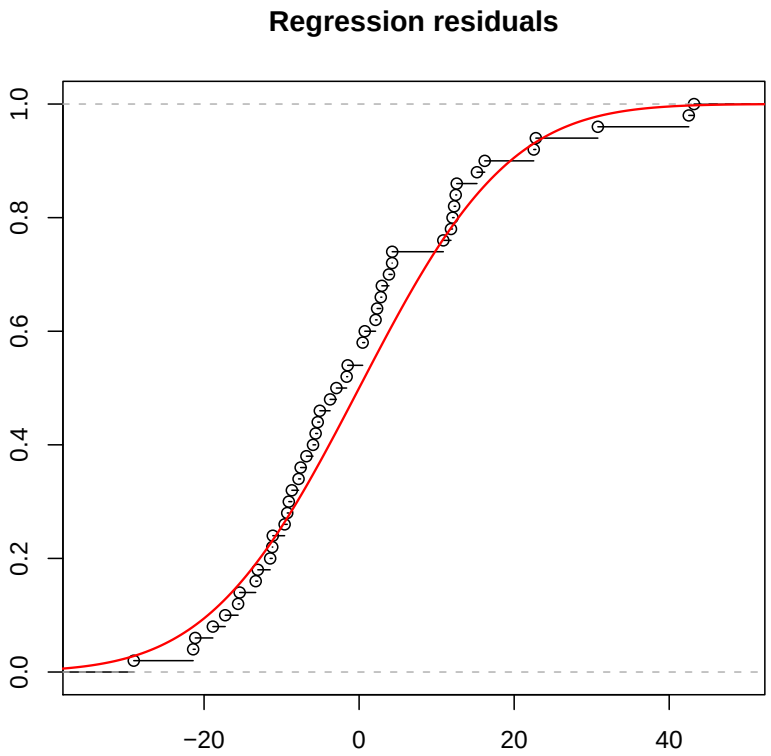
```
>  head(dl,10)
    X              Country     Continent   LE CigCon
1   1          Afghanistan          Asia 35.5     98
2   2              Albania        Europe 61.4     NA
3   3              Algeria        Africa 60.6   1021
4   4              Andorra        Europe 72.2     NA
5   5               Angola        Africa 33.4    571
6   6 Antigua and Barbuda South America 61.9     NA
7   7            Argentina South America 65.3   1495
8   8              Armenia        Europe 61.0   1095
9   9            Australia     Australia 72.6   1907
10 10              Austria        Europe 71.4   2073
```

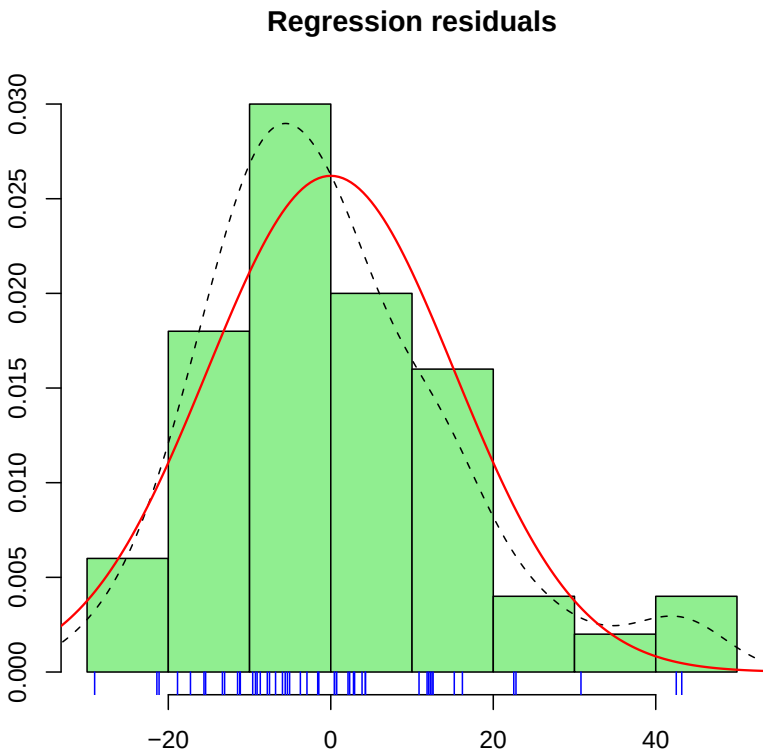
Les consommations les plus élevées de cigarettes sont observées dans les pays les plus riches, où l'espérance de vie est la plus grande.

Remarque Une régression linéaire est l'ananlyse d'une corrélation et pas d'une relation de causalité.

Tester la normalité

Tous les résultats ont étés obtenus sous l'hypothèse de normalité des résidus.





Tester la normalité

Le test de Kolmogorov Smirnov permet de tester l'ajustement d'un loi normale $\mathcal{N}(0, 15^2)$ pour les résidus,

```
> ks.test(reg$residuals, "pnorm", 0,15)

        One-sample Kolmogorov-Smirnov test
data:  E
D = 0.128, p-value = 0.3859
alternative hypothesis: two-sided

Warning message:
In ks.test(E, "pnorm", 0, 15) : cannot compute correct p-values with ties
```

Rappelons que le test de Kolmogorov Smirnov, de l'hypothèse $H_0 : F = F_\star$ (contre l'hypothèse alternative $H_0 : F \neq F_\star$) est basé sur la statistique

$$D = \sup_{x \in \mathbb{R}} \left\{ \left| F_\star(x) - \widehat{F}_n(x) \right| \right\},$$

Le théorème de Glivenko-Cantelli garantissant que

$$\sup_{x \in \mathbb{R}} \left\{ \left| F_\star(x) - \widehat{F}_n(x) \right| \right\} \xrightarrow{\text{p.s.}} 0$$

lorsque $n \to \infty$, sous H_0, i.e. si $F_\star$ est effectivement la vraie loie.

Mais formellement ce n'est pas un test de normalité : on teste ici $H_0 : F = \mathcal{N}(\mu_\star, \sigma_\star^2)$ et non pas $H_0 : F = \mathcal{N}(\star, \star)$.

Tester la normalité

```
> shapiro.test(reg$residuals)

        Shapiro-Wilk normality test

W = 0.9451, p-value = 0.02153
> ad.test(reg$residuals)

        Anderson-Darling normality test

A = 0.7941, p-value = 0.0369
```

Ces tests sont techniques, et sont détaillés dans les livres des statistiques (cf blog).

Le test de Cramér-von-Mises repose sur

$$nW^2 = n\int_{-\infty}^{\infty}[F(x) - F^\star(x)]^2 dF^\star(x)$$

soit, en pratique

$$T = nW^2 = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{2i-1}{2n} - F^\star(x_{i:n}) \right]^2,$$

où $F^\star$ est la loi théorique que l'on cherche à tester, i.e. Φ.

```
> cvm.test(reg$residuals)
        Cramer-von Mises normality test

W = 0.1257, p-value = 0.0483
> pearson.test(reg$residuals)

        Pearson chi-square normality test

P = 8.4, p-value = 0.2986
```

Le test de Jarque-Bera repose sur

$$JB = \frac{n}{6} \left(S^2 + \frac{(K-3)^2}{4} \right),$$

où n est le nombre de degrés de libertés, S est la skewness empirique, et K la kurtosis empirique,

$$S = \frac{\mu_3}{\sigma^3} = \frac{\mu_3}{\left(\sigma^2\right)^{3/2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^3}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{3/2}}$$

$$K = \frac{\mu_4}{\sigma^4} = \frac{\mu_4}{\left(\sigma^2\right)^{2}} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\right)^{2}}$$

Sous H_0 (normalité), $JB \sim \chi^2(2)$.

```
> jarque.bera.test(reg$residuals)

        Jarque Bera Test

X-squared = 8.1888, df = 2, p-value = 0.01667
```

Tester la présence d'une rupture, le test de Chow

Le test de Chow permet de tester une rupture en spécifiant explicitement la rupture. Le test de Chow est simplement un test de Ficher, où on compare des sous-modèles à un modèle global. On suppose ici

$$\boldsymbol{\beta} = \begin{cases} \boldsymbol{\beta}_1 \text{ pour } i = 1, \cdots, i_0 \\ \boldsymbol{\beta}_2 \text{ pour } i = i_0 + 1, \cdots, n \end{cases} \quad \text{et on teste} \quad \begin{cases} H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \\ H_1 : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \end{cases}$$

i_0 est ici un point entre k et $n-k$ (il faut garder suffisement d'observations pour mener le test). Chow (1960) suggère un test de la forme

$$F_{i_0} = \frac{\widehat{\eta}'\widehat{\eta} - \widehat{\varepsilon}'\widehat{\varepsilon}}{\widehat{\varepsilon}'\widehat{\varepsilon}/(n-2k)}$$

où

$$\widehat{\varepsilon}_i = Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}} \text{ pour } i = k, \cdots, n-k$$

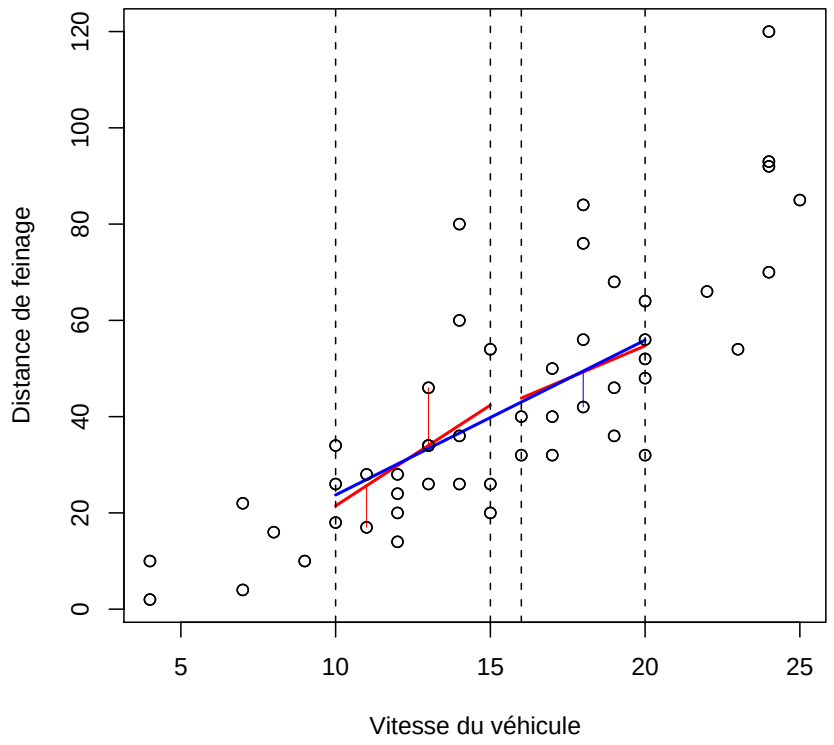
$$\widehat{\eta}_i = \begin{cases} Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_1 \text{ pour } i = k, \cdots, i_0 \\ Y_i - \boldsymbol{X}_i'\widehat{\boldsymbol{\beta}}_2 \text{ pour } i = i_0 + 1, \cdots, n-k \end{cases}$$

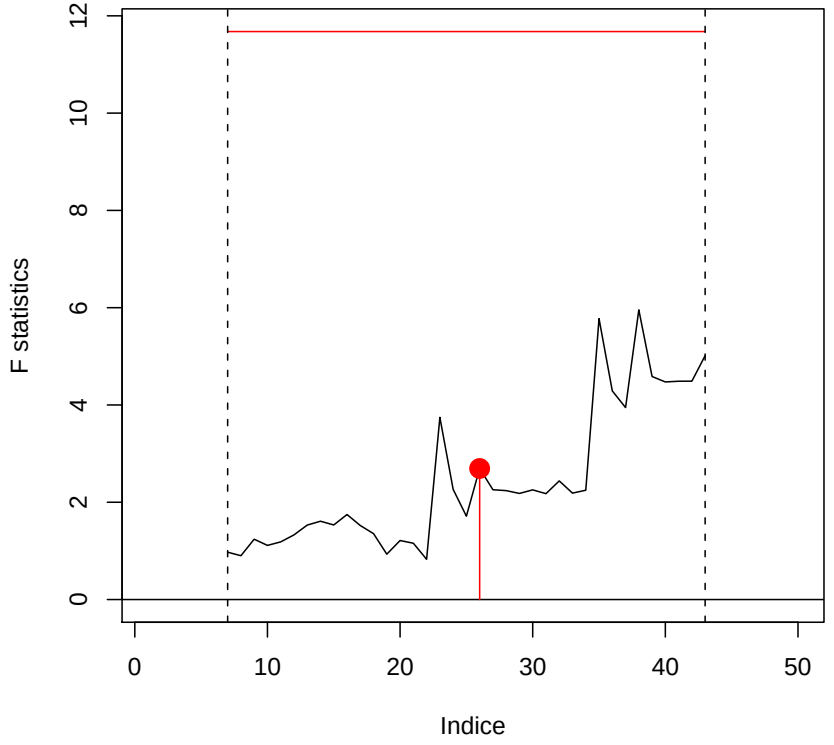
La fonction `Fstats` représente ainsi F_{i_0} pour toutes les valeurs entre k et $n-k$ (par défaut, 30% des observations sont retirées, 15% à gauche 15% à droite).

```
> reg1=lm(formula = dist ~ speed, data = cars[1:26,])
> S1=sum(reg1$residuals^2); DL1=reg1$df.residual
> reg2=lm(formula = dist ~ speed, data = cars[27:50,])
> S2=sum(reg2$residuals^2); DL2=reg2$df.residual
> reg0=lm(formula = dist ~ speed, data = cars[1:50,])
> S0=sum(reg0$residuals^2); DL0=reg0$df.residual
> ((S0-(S1+S2))/(DL0-(DL1+DL2)-1))/(S0/(DL0-1))
[1] 2.601249
```

Tester la présence d'une rupture, le test de Chow

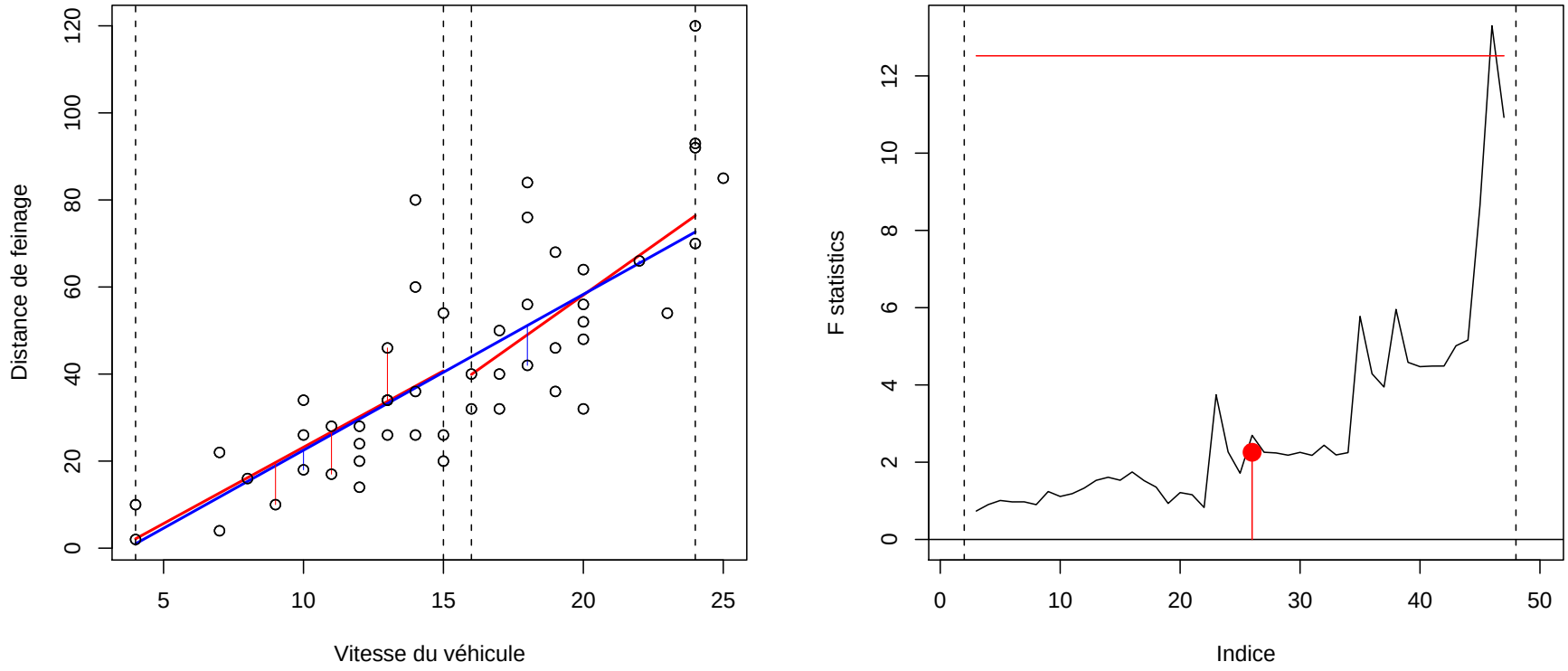
```
Fstats(dist ~ speed,data=cars,from=7/50)
```





Tester la présence d'une rupture, le test de Chow

```
Fstats(dist ~ speed,data=cars,from=2/50)
```



Tester la présence d'une rupture, le test de Chow

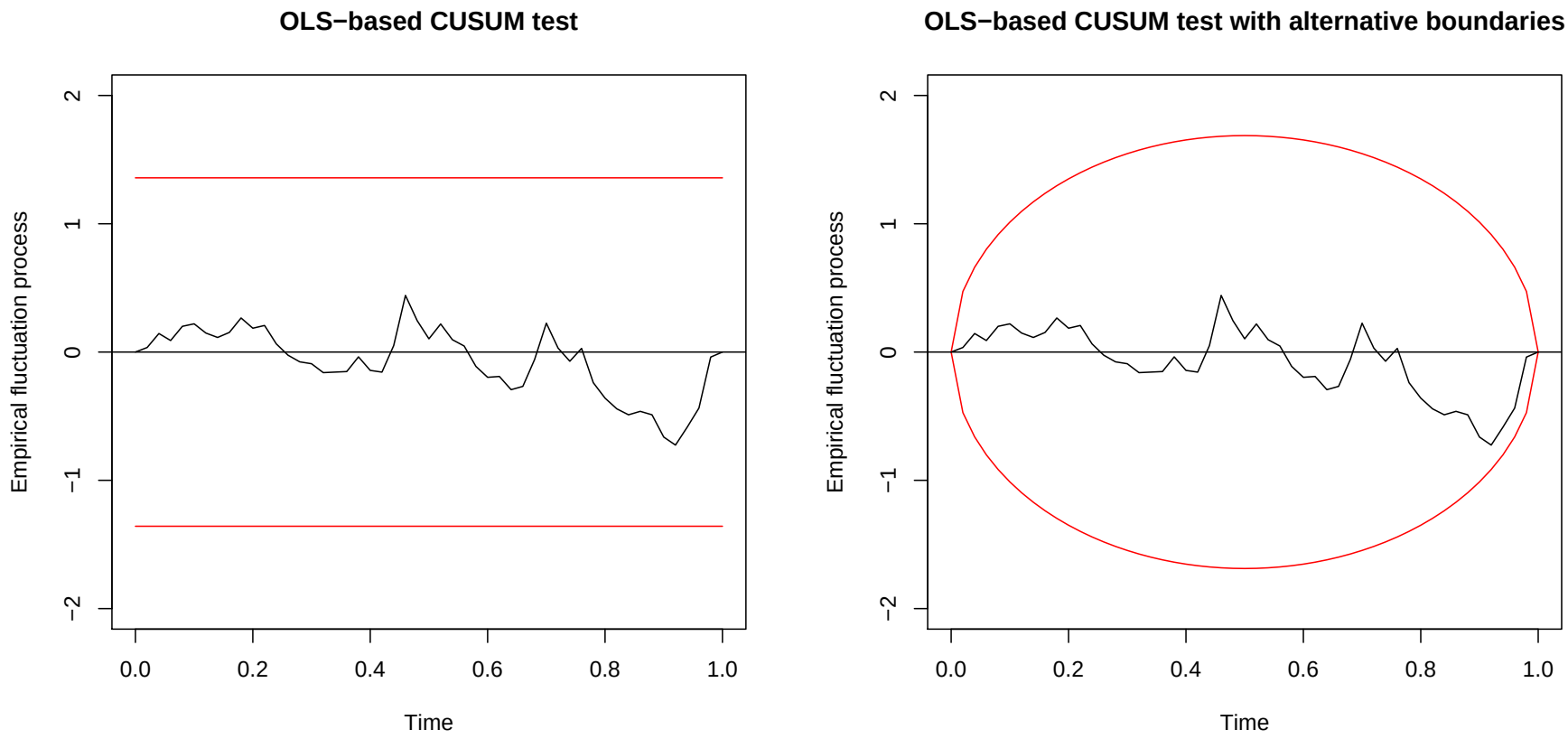
Si l'on ne sait pas à quelle date la rupture a eu lieu, des outils de type CUSUM peuvent être utilisés (cf Ploberger & Krämer (1992)). Pour $t \in [0,1]$, on pose

$$W_t = \frac{1}{\widehat{\sigma}\sqrt{n}} \sum_{i=1}^{\lfloor nt \rfloor} \widehat{\varepsilon}_i.$$

Dans ce test, si α est le niveau de confiance, les bornes généralement considérées sont $\pm\alpha$. En fait, les bornes théoriques sont plutôt de la forme $\pm\alpha\sqrt{t(1-t)}$.

Sous R, il suffit de spécifier `type=''OLS-CUSUM''` dans la fonction `efp`.

```
cusum <- efp(dist ~ speed, type = "OLS-CUSUM",data=cars)
plot(cusum,ylim=c(-2,2))
plot(cusum, alpha = 0.05, alt.boundary = TRUE,ylim=c(-2,2))
```



Les points sont ici triés par vitesse. Aussi, une rupture détectée en $t = 92\%$ signifie qu'il y a une rupture dans la modélation linaire de Y par X à partir de la $tn = 46$ème observation.

Intervalles de confiance pour Y et $\widehat{Y}$

Supposons que l'on dispose d'une nouvelle observation $\boldsymbol{X}_0$ et que l'on souhaite prédire la valeur Y_0 assoicée.

L'incertitude sur la prédiction $\widehat{Y}_0$ vient de l'erreur d'estimation sur les paramètres $\boldsymbol{\beta}$.

L'incertitude sur la réalisation Y_0 vient de l'erreur d'estimation sur les paramètres $\boldsymbol{\beta}$ et de l'erreur associée au modèle linéaire, i.e. ε_0.

Dans le cas de la régression simple, $Y_0 = \widehat{Y}_0 + \varepsilon_0 = \beta_0 + \beta_1 X_0 + \varepsilon_0$. Aussi,

- $\text{Var}(\widehat{Y}_0) = \text{Var}(\widehat{\beta}_0) + 2X_0\text{cov}(\widehat{\beta}_0, \widehat{\beta}_1) + X_0^2\text{Var}(\widehat{\beta}_1)$
- $\text{Var}(Y_0) = Var(\widehat{Y}_0) + \text{Var}(\varepsilon_0)$, si l'on suppose que le bruit est la partie non expliquée. Aussi $\text{Var}(Y_0) = \text{Var}(\widehat{\beta}_0) + 2X_0\text{cov}(\widehat{\beta}_0, \widehat{\beta}_1) + X_0^2\text{Var}(\widehat{\beta}_1) + \sigma^2$

Intervalle de confiance pour $\widehat{Y}$

L'incertitude sur la prédiction $\widehat{Y}_j$ vient de l'erreur d'estimation sur les paramètres $\boldsymbol{\beta}$.

Dans ce cas, si l'on dispose d'une nouvelle observation $\boldsymbol{X}_0$, l'intervalle de confiance pour $\widehat{Y}_0 = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y_0$ est

$$\left[\widehat{Y}_0 \pm t_{n-k}(1-\alpha/2)\widehat{\sigma}\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0}\right]$$

où $t_{n-k}(1-\alpha/2)$ est le quantile d'ordre $(1-\alpha/2)$ de la loi de Student à $n-k$ degrés de liberté.

Intervalle de confiance pour Y

L'incertitude sur la réalisation Y_j vient de l'erreur d'estimation sur les paramètres $\boldsymbol{\beta}$ et de l'erreur associée au modèle linéaire, i.e. ε_i.

Dans ce cas, si l'on dispose d'une nouvelle observation $\boldsymbol{X}_0$, l'intervalle de confiance pour $Y_0 = \widehat{Y}_0 + \varepsilon_0$ est

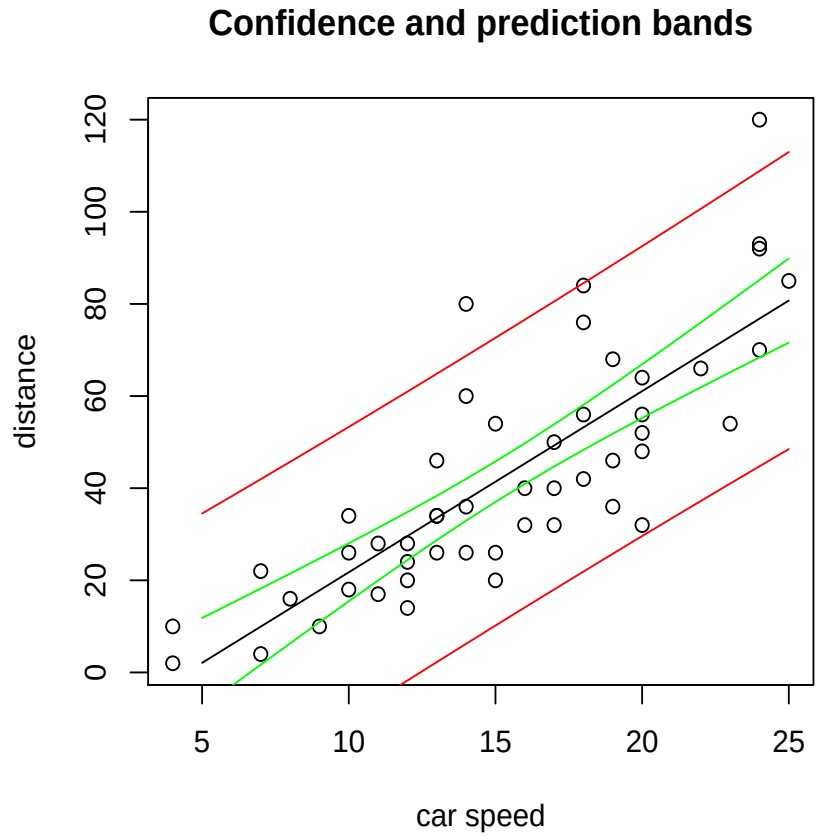
$$\left[\widehat{Y}_0 \pm t_{n-k}(1-\alpha/2)\widehat{\sigma}\sqrt{1+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}_0}\right]$$

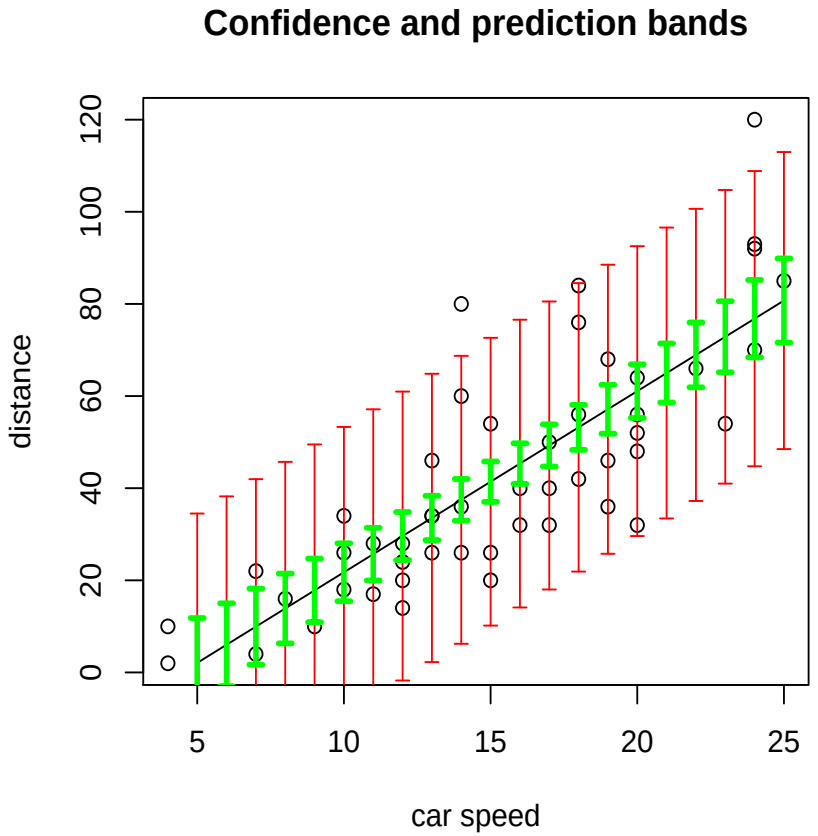
En fait, sous `R`, l'intervalle de confiance pour Y n'inclue pas l'erreur associée à l'erreur d'estimation, aussi, dans ce cas, si l'on dispose d'une nouvelle observation $\boldsymbol{X}_0$, l'intervalle de confiance pour $Y_0 = \widehat{Y}_0 + \varepsilon_0$ est

$$\left[\widehat{Y}_0 \pm t_{n-k}(1-\alpha/2)\widehat{\sigma}\right].$$

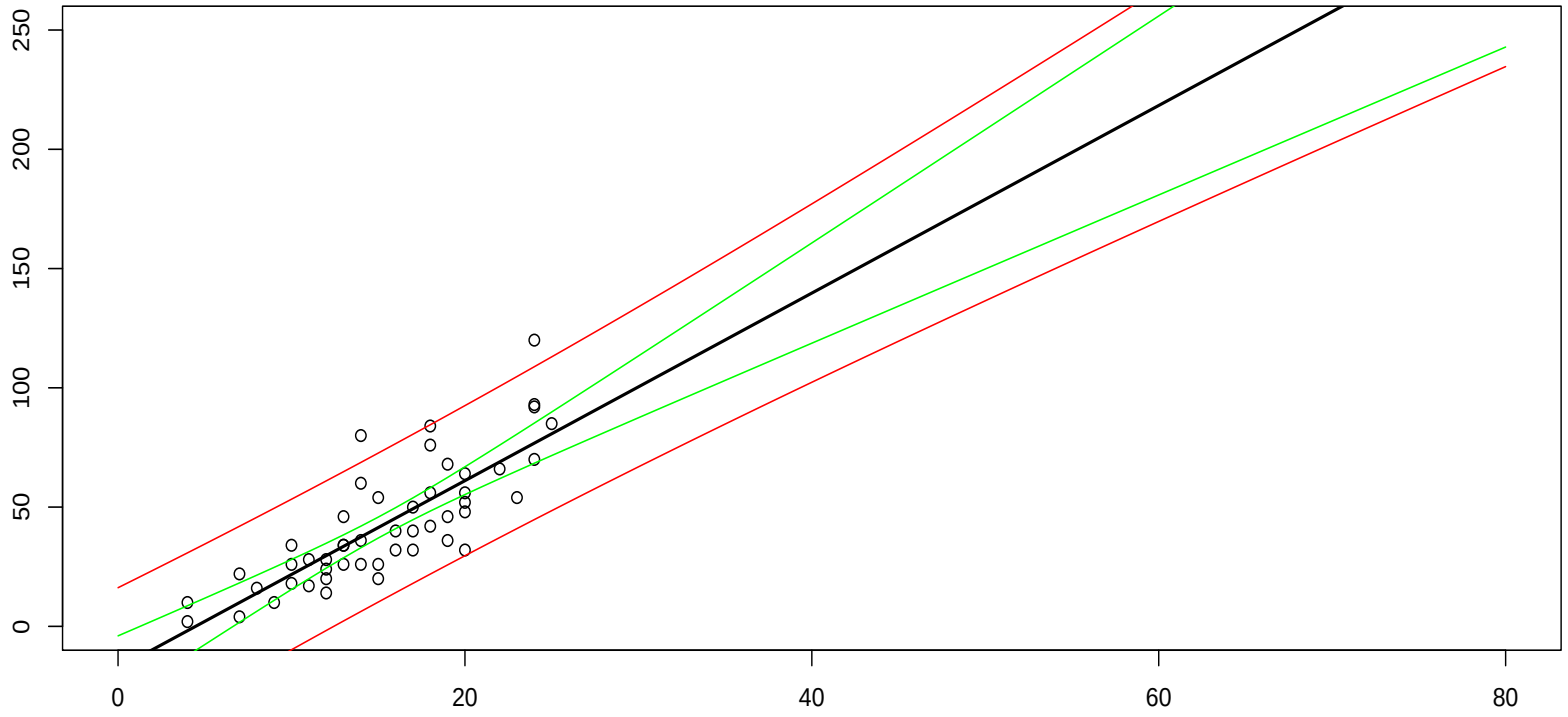
Intervalles de confiance pour Y et $\widehat{Y}$

```
pp <- predict(lm(y~x,data=D), new=data.frame(x=seq(0,30)), interval='prediction')
pc <- predict(lm(y~x,data=D), new=data.frame(x=seq(0,30)), interval='confidence')
```



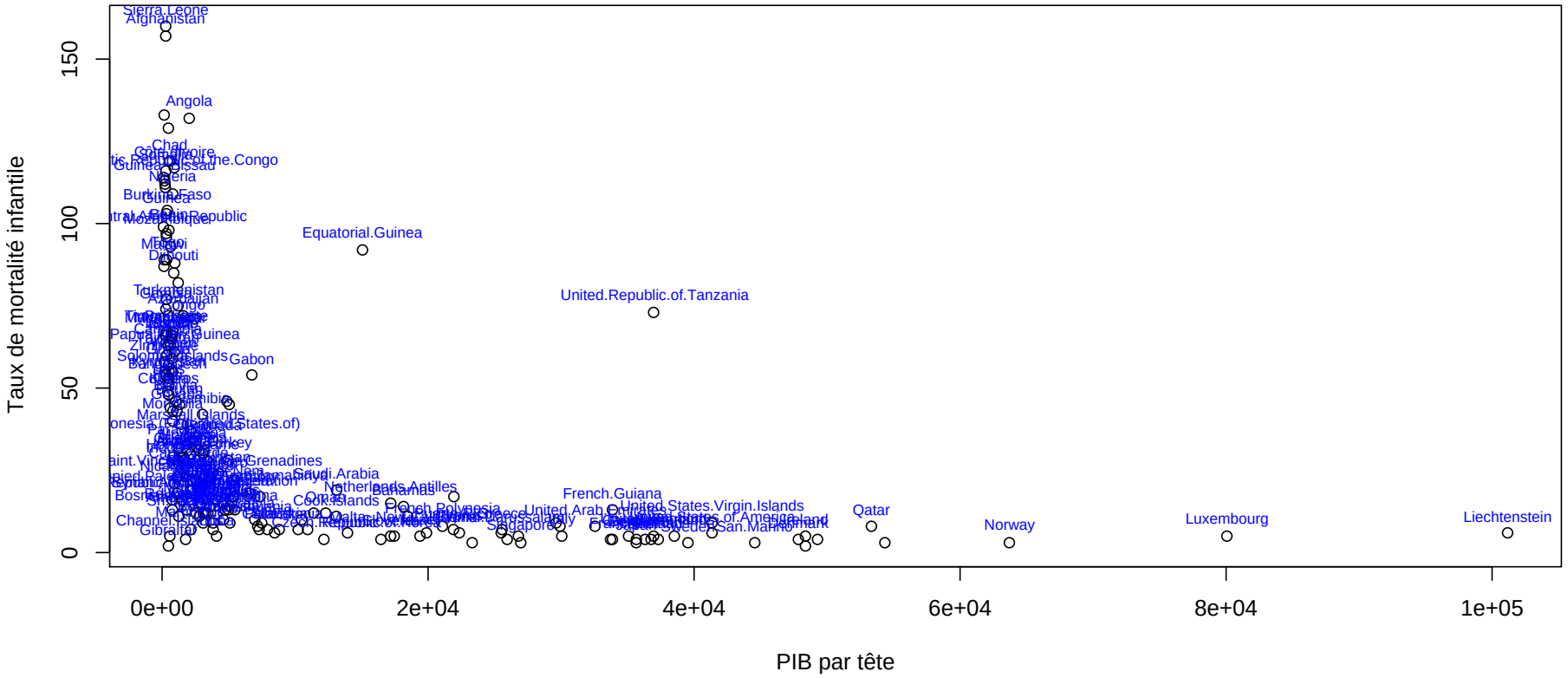


Intervalles de confiance pour Y et $\widehat{Y}$



Modèle linéaire, ou multiplicatif ?

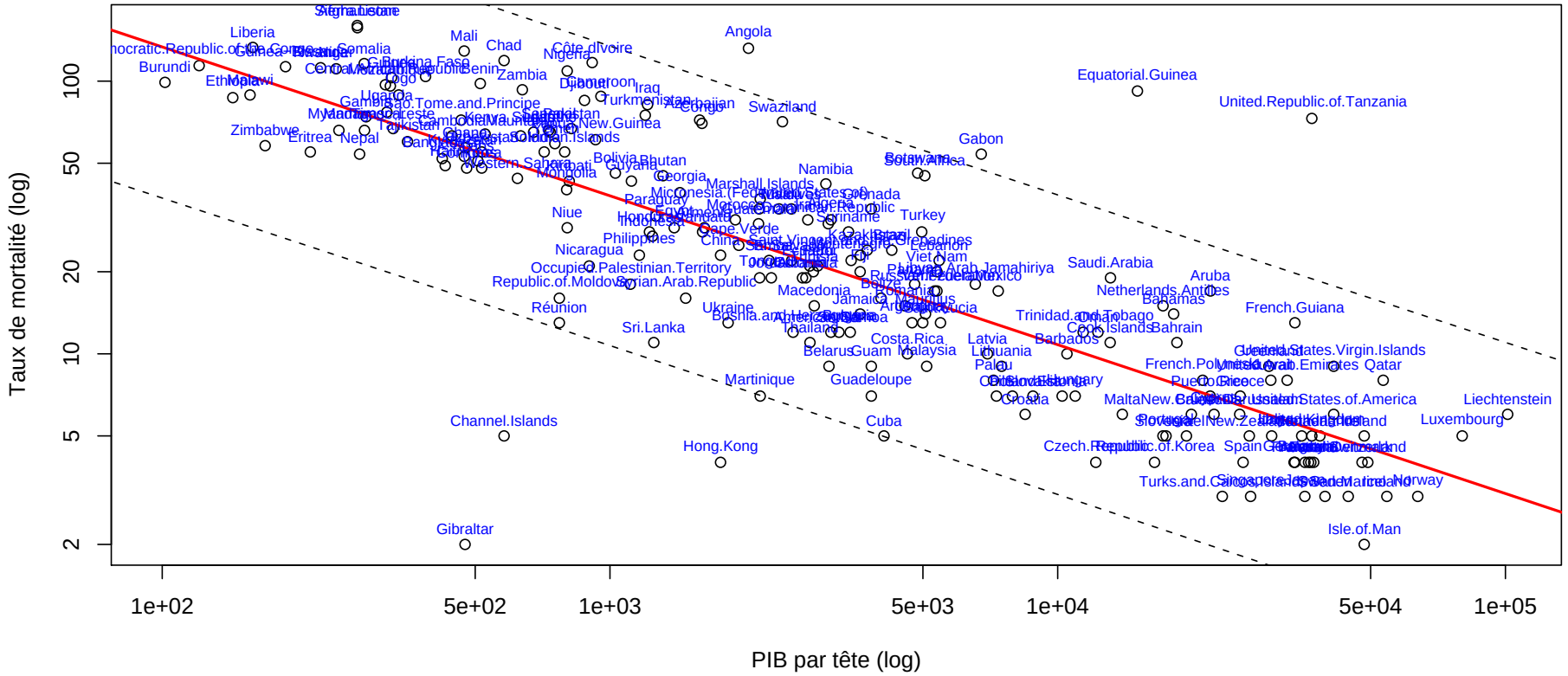
Considérons le taux de mortalité infantile comme une fonction du PIB par tête.



Visiblement le modèle linéaire ne convient pas.

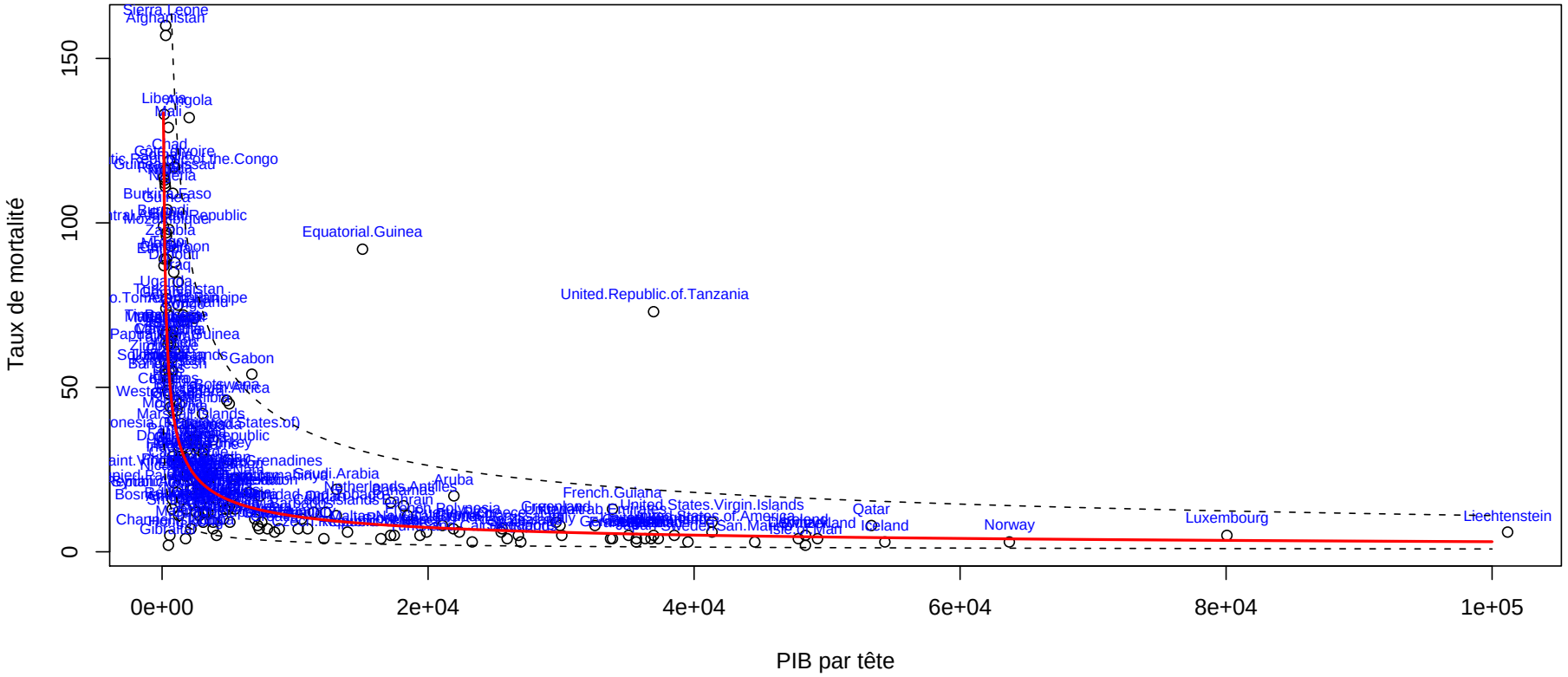
Modèle linéaire, ou multiplicatif ?

Transformation a priori : transformation logarithmique,



Modèle linéaire, ou multiplicatif ?

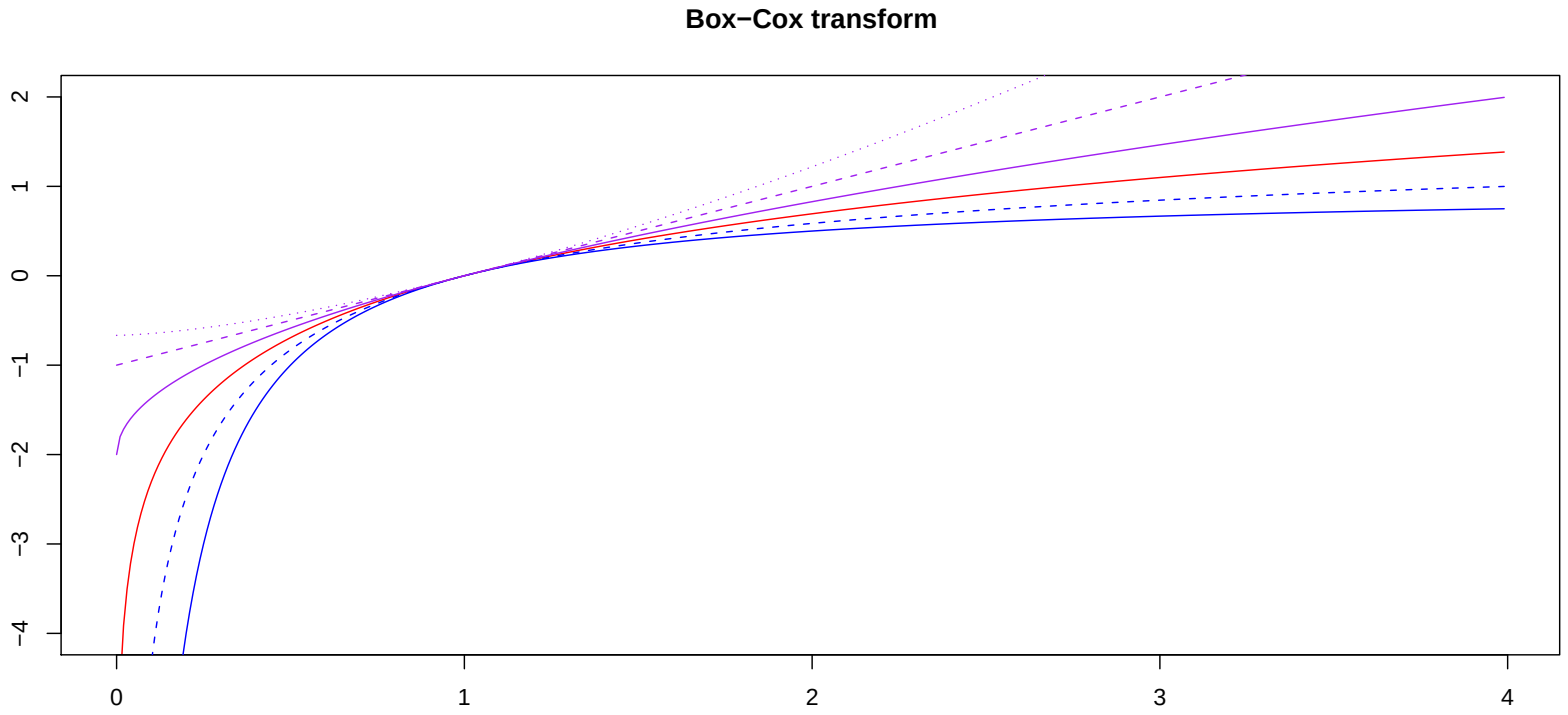
Transformation a priori : transformation logarithmique,



Modèle sur Y ou $\log Y$

La transformation paramétrique la plus classique en économétrie est la transformée de Box-Cox,

$$f(x,\lambda)=\begin{cases}\dfrac{x^{-\lambda}-1}{\lambda} \text{ pour } \lambda\neq 0\\ \log(x) \text{ pour } \lambda=0\end{cases}$$



Cette transformation ne marche toutefois que pour les valeurs positives. Une variante (proposée dans le même papier) est

$$f(x,\lambda,\mu)=\begin{cases}\dfrac{[x+\mu]^{-\lambda}-1}{\lambda} \text{ pour } \lambda\neq 0\\ \log([x+\mu]) \text{ pour } \lambda=0\end{cases}$$

En pratique, μ n'est pas considéré comme un paramètre inconnu.

L'idée est de transformer les données de telle sorte que $f(Y, \lambda^\star)$ soit approximativement normal pour un $\lambda^\star$ bien choisi.

Supposons que Y suive une loi exponentielle. La transformation de Box-Cox donne alors une loi de Weibull...

L'objectif initial de l'analyse de Box-Cox était d'estimer un coefficient $\lambda^\star$ optimal

- la première idée a été d'utiliser la méthode du maximum de vraisemblance. Grce aux propriétés asymptotiques, on peut également obtenir un intervalle de confiance approché.
- la seconde idée (proposée dès 1964) a été d'utiliser la méthode bayésienne.

Dans l'approche par maximum de vraisemblance, on suppose que

$$Y^\lambda = f(Y,\lambda) \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\mathbb{I}).$$

La vraisemblance du modèle s'écrit alors

$$\mathcal{L}(\lambda,\boldsymbol{\beta},\sigma^2|Y,\boldsymbol{X}) = (2\pi\sigma^2)^{-n/2}\exp\left(\frac{-(Y^\lambda-\boldsymbol{X}\boldsymbol{\beta})'(Y^\lambda-\boldsymbol{X}\boldsymbol{\beta})}{2\sigma^2}\right)J(\lambda,Y)$$

où $J(\lambda,Y)$ désigne le Jacobien de la transformationn $Y \mapsto Y^\lambda = f(Y,\lambda)$, i.e.

$$J(\lambda,Y) = \prod_{i=1}^{n} Y_i^{\lambda-1}.$$

Notons que - conditionnellement à λ on retrouve le modèle linéaire gaussien classique, et donc les estimations du maximum de vraisemblance pour le couple $(\boldsymbol{\beta}, \sigma^2)$ sont

$$\tilde{\boldsymbol{\beta}}_\lambda = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Y^\lambda,$$

$$\tilde{\sigma^2}_\lambda = \frac{1}{n}Y^{\lambda\prime}[\mathbb{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']Y^\lambda,$$

La fonction

$$VP : \lambda \mapsto \mathcal{L}(\lambda, \tilde{\boldsymbol{\beta}}_\lambda, \tilde{\sigma}^2_\lambda | Y, \boldsymbol{X})$$

est appelée vraisemblance profilée (*profile likelihood*). Notons que

$$\log VP(\lambda) = \text{constant } - \frac{n}{2}\log(\tilde{\sigma}^2_\lambda) + [\lambda - 1]\sum_{i=1}^{n}\log(Y_i).$$

Comme nous avons pu écrire la vraisemblance, notons qu'il est possible de faire toute sorte de tests, en particulier des tests de rapport de vraisemblance,

Pour test $H : \lambda = \lambda_0$, on utilise la statistique du rapport de vraisemblance

$$W = 2[\log VP(\widehat{\lambda}) - \log VP(\lambda_0)] \xrightarrow{\mathcal{L}} \chi^2(1)$$

Remarque Au delà de la transformée de Box-Cox, il existe aussi la transformée dite de Box-Tidwell, correpondant à une fonction puissance.

Faire une prédiction avec un modèle logarithmique

Si on retient l'idée d'un modèle linéaire sur $\log Y$, cela ne donne pas (pour l'instant) un modèle pour Y. D'après l'inégalité de Jensen, on va toujours sous estimer la valeur en prenant l'exponentielle de la valeur prédite par le modèle logarithmique.

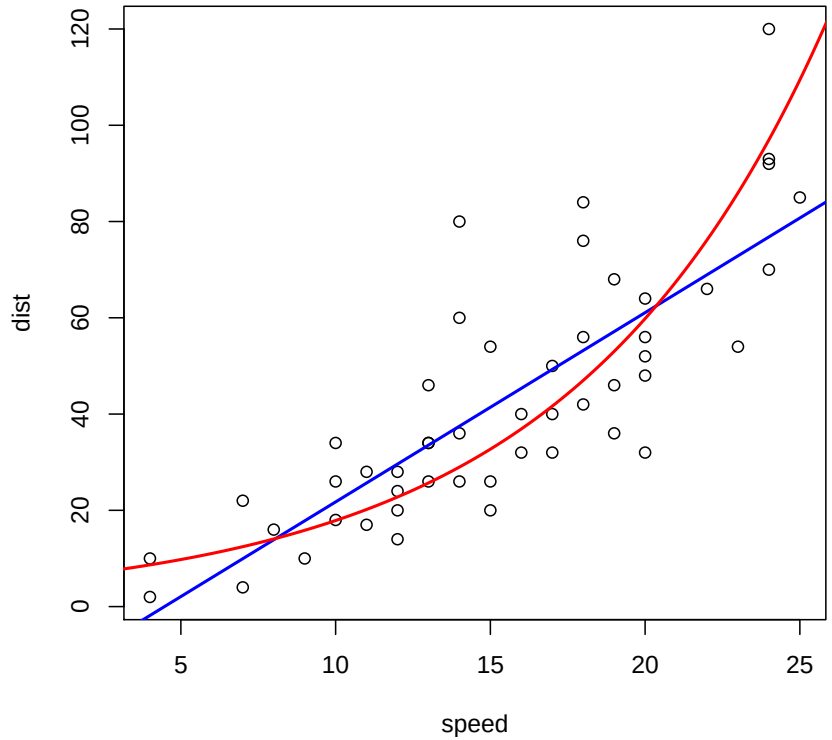
Si $Y \sim \mathcal{N}(\mu, \sigma^2)$ avec $\mathbb{E}(Y) = \mu$

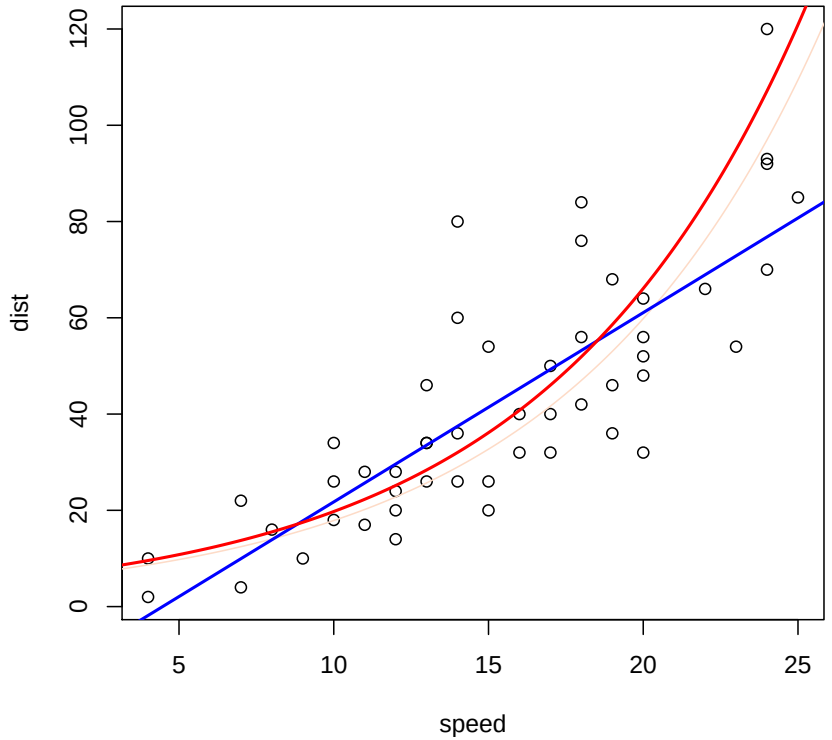
alors $\exp(Y) \sim LN(\mu, \sigma^2)$ mais $\mathbb{E}(\exp(Y)) \neq \exp(\mu)$ car

$$\mathbb{E}(\exp(Y)) = \exp\left(\mu + \frac{1}{2}\sigma^2\right).$$

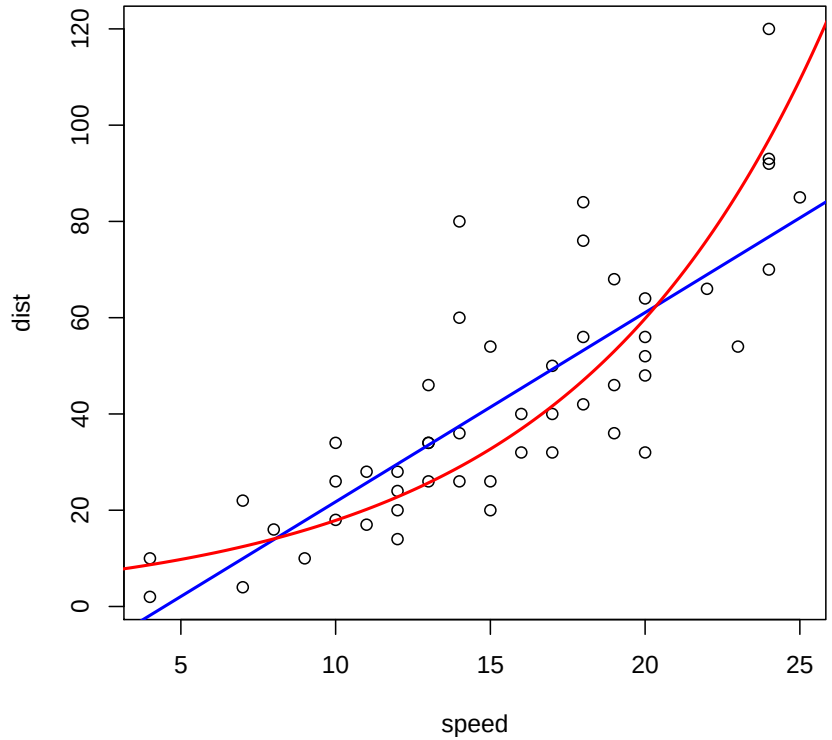
```
> reg1=lm(dist~speed,data=cars)
> reg2=lm(log(dist)~speed,data=cars)
> newcars=data.frame(speed=seq(1,26,by=.1))
```

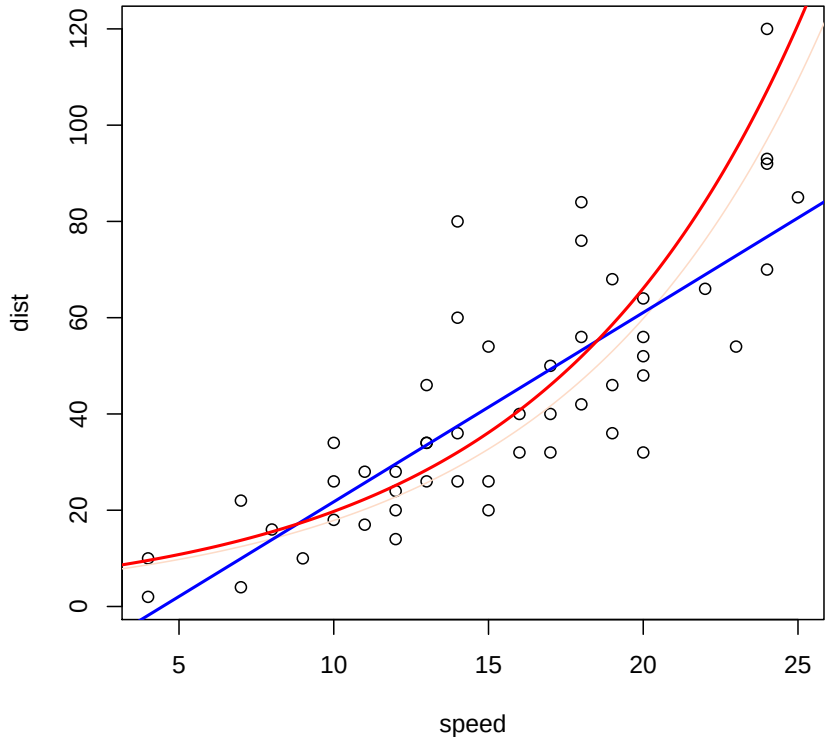
```
> p1=predict(reg1,newdata=newcars)
> p2=exp(predict(reg2,newdata=newcars))
>
```





```
> p1=predict(reg1,newdata=newcars)
> p2b=exp(predict(reg2,newdata=newcars)
+ .5*summary(reg2)$sigma^2)
```





La régression pondérée

L'idée des moindres carrés pondérés (*weighted least squares*) consiste à chercher à minimiser

$$\sum_{i=1}^{n} \omega_i [Y_i - (\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k})]^2$$

Notons $\boldsymbol{\Omega}$ la matrice diagonale $\boldsymbol{\Omega} = [\boldsymbol{\Omega}_{i,j}]$ avec $\boldsymbol{\Omega}_{i,j} = \omega_i \delta_{i=j}$. Dans ce cas, la condition du premier ordre (équations normales) s'écrit

$$\left(\boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{\Omega}\boldsymbol{Y}.$$

Si on pose $\boldsymbol{W} = \boldsymbol{\Omega}^{1/2}$ i.e. $\boldsymbol{\Omega} = \boldsymbol{W}'\boldsymbol{W}$, alors

$$\left(\boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{X}\right)\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{W}'\boldsymbol{W}\boldsymbol{Y}.$$

i.e. si $\tilde{\boldsymbol{X}}' = \boldsymbol{W}\boldsymbol{X}$ et $\tilde{\boldsymbol{Y}} = \boldsymbol{W}\boldsymbol{Y}$

$$\left(\tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}}\right)\widehat{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{Y}},$$

qui est la condition du premier ordre pour un modèle linéaire standard.

Application : prise en compte de l'hétéroscédasticité

Si l'on suppose que $Var(\varepsilon_i) = \gamma X_{i,j}^2$ où $\gamma > 0$ pour un $j = 1, \cdots, k$, alors on cherche à minimiser

$$\sum_{i=1}^{n} \frac{1}{X_{i,j}^2} [Y_i - (\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k})]^2.$$

La méthode naturelle pour estimer les coefficients consiste à considérer des moindres carrés ordinaires sur la régression

$$\frac{Y_i}{X_{i,j}} = \alpha_0 \frac{1}{X_{i,j}} + \alpha_1 \frac{X_{i,1}}{X_{i,j}} + \cdots + \alpha_k \frac{X_{i,k}}{X_{i,j}} + \frac{\varepsilon_i}{X_{i,j}}.$$

Notons que pour $j \neq i, 0$, $\widehat{\alpha}_j$ est un estimateur de β_j, $\widehat{\alpha}_j$ est un estimateur de β_0, et $\widehat{\alpha}_0$ est un estimateur de β_j.

Application : régression locale, et lissage

Rappelons qu'un espérance conditionnelle est l'espérance associée à la loi conditionnelle, i.e.

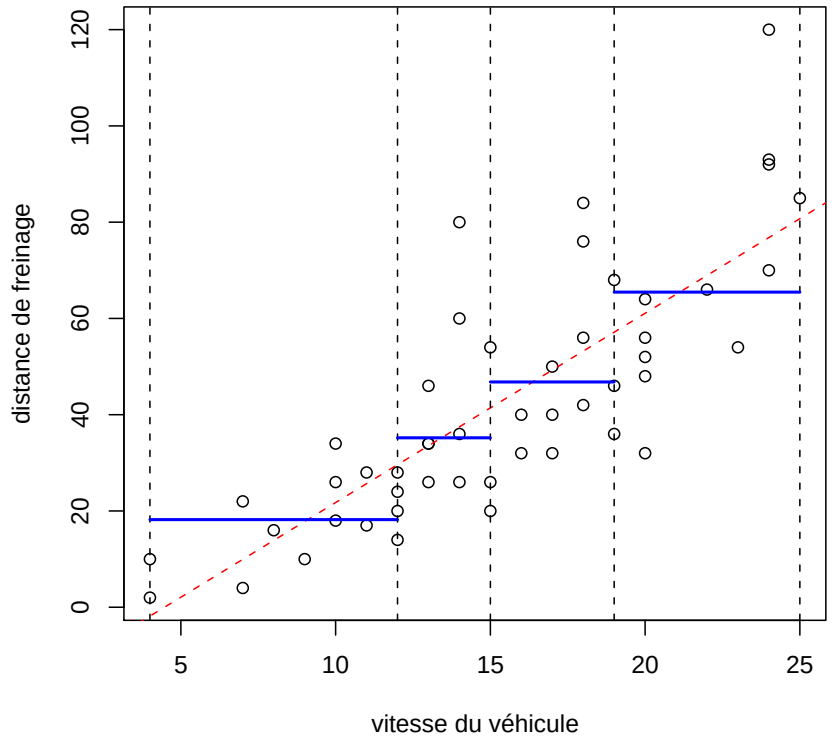
$$\varphi(x) = \mathbb{E}(Y|X=x) = \int y f_{Y|X=x}(u|x)du = \int y \frac{f_{Y,X}(u,x)}{f_X(x)}du = \frac{\int y f_{Y,X}(u,x)du}{f_X(x)}$$

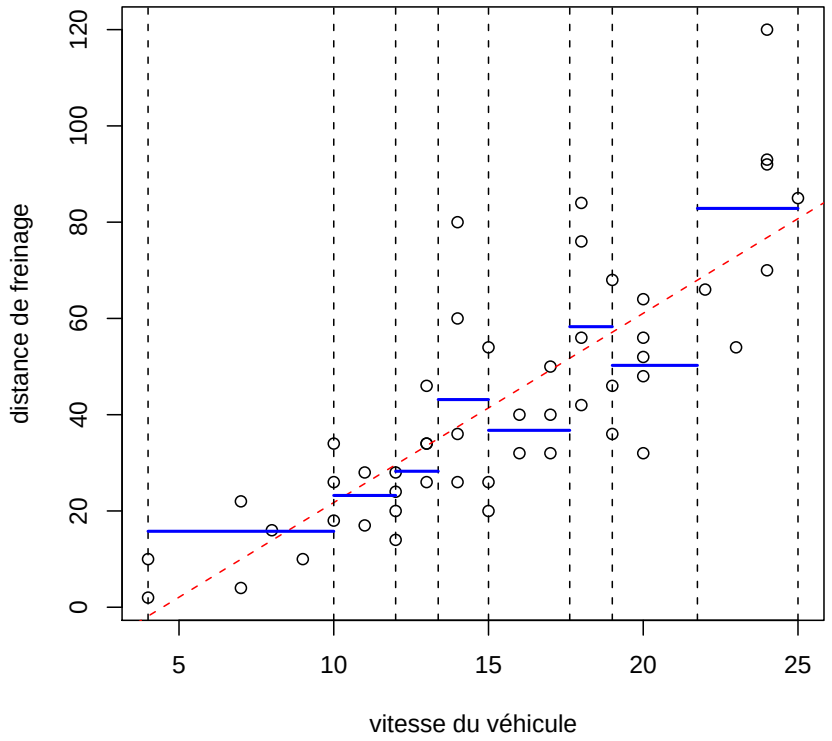
Tukey (1961) a proposé de transposer l'histogramme au à l'approximation de l'espérance conditionelle.

Soit $(B_j)_{j=1,\cdot,m}$ une partition du support de X,

$$\widehat{\varphi}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i \in B_j)}{\sum_{i=1}^{n} \mathbf{1}(X_i \in B_j)} \text{ pour tout } x \in B_j.$$

On parle de régressogramme, proposé par Tukey (1961)


distance de freinage
vitesse du véhicule
0
20
40
60
80
100
120
5
10
15
20
25

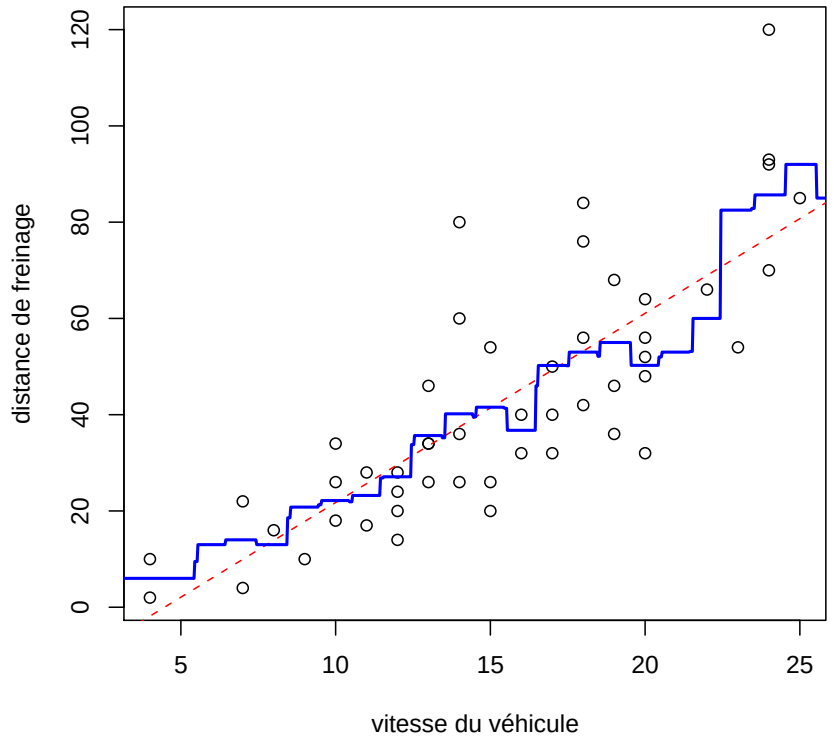

distance de freinage
vitesse du véhicule
0
20
40
60
80
100
120
5
10
15
20
25

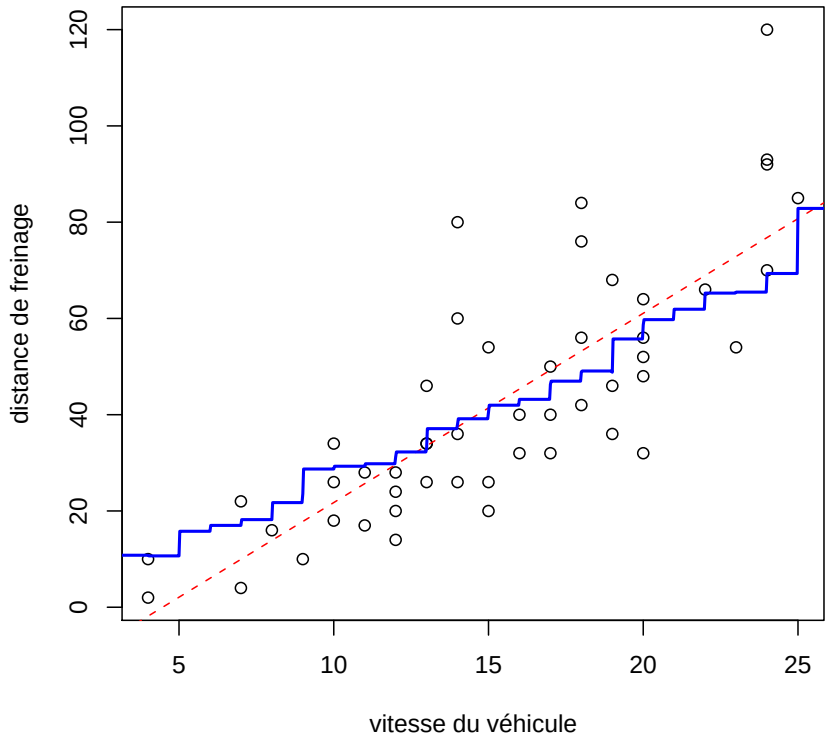
De l'histogramme au regressogramme

Naturellement, on peut considérer un régressogramme glissant,

$$\widehat{\varphi}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i \in [x_h; x+h])}{\sum_{i=1}^{n} \mathbf{1}(X_i \in [x_h; x+h])} \text{ pour tout } x,$$

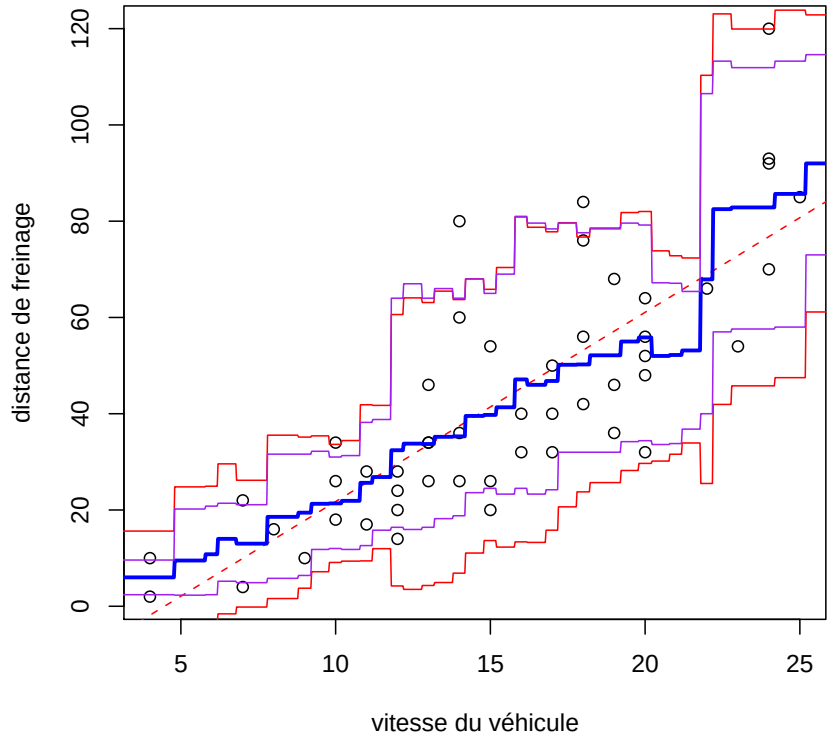
où $h > 0$.

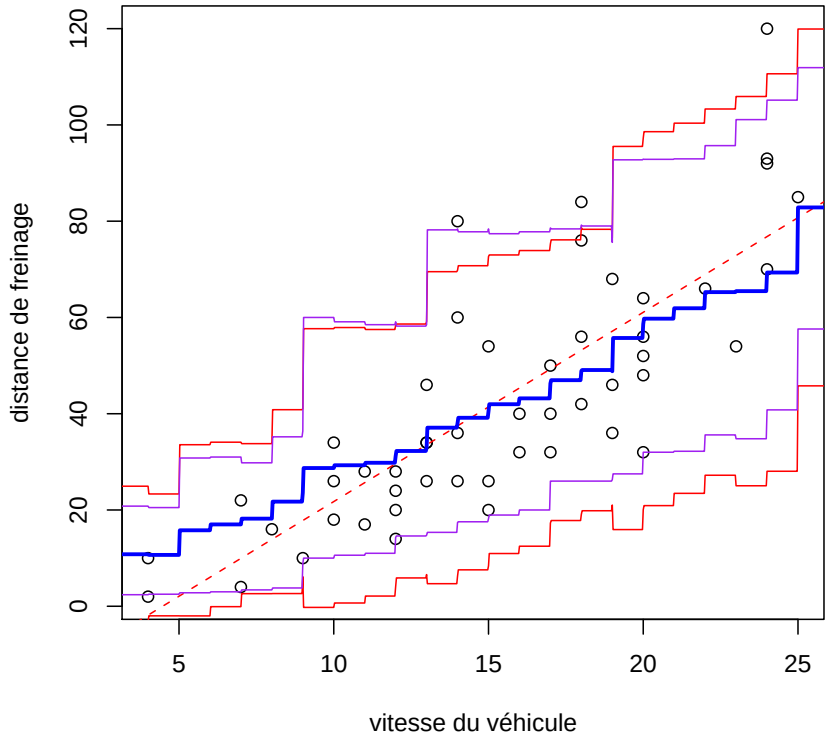

distance de freinage
vitesse du véhicule
0
20
40
60
80
100
120
5
10
15
20
25


distance de freinage
vitesse du véhicule
0
20
40
60
80
100
120
5
10
15
20
25

De l'histogramme au regressogramme

Notons qu'on peut également obtenir un intervalle de confiance, soit en utilisant un intervalle de confiance gaussien (avec l'écart-type estimé sur le voisinage, -) ou en utilisant les quantiles empiriques (-).



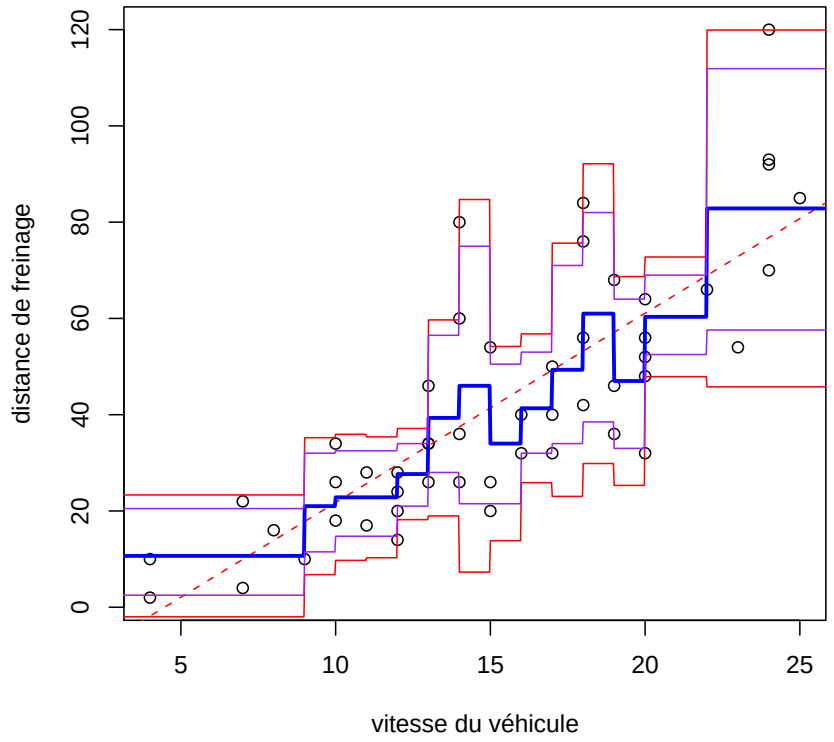


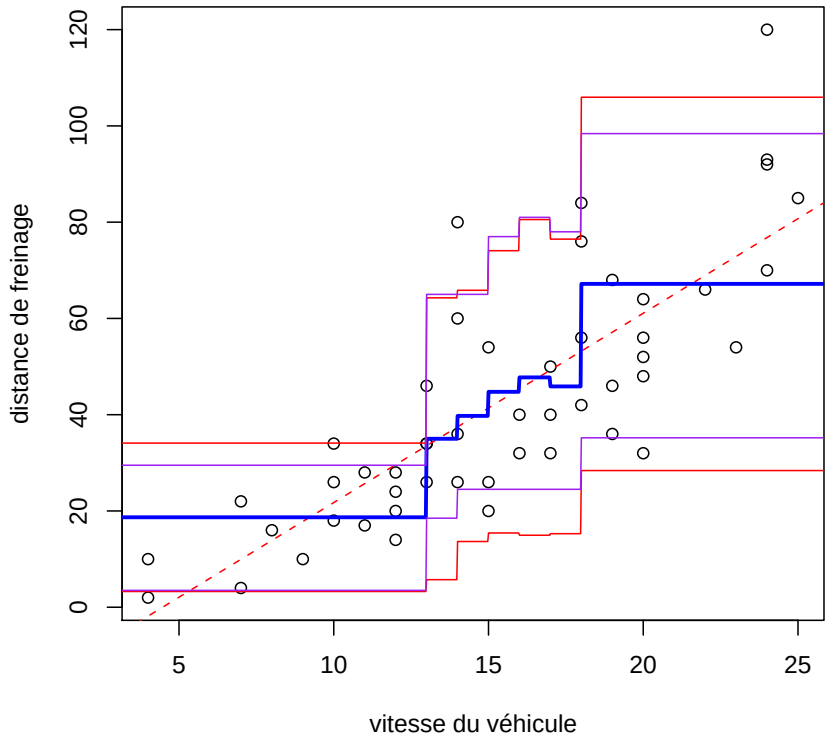
De l'histogramme au regressogramme

Nous venons de prendre la moyenne sur les voisins de x distants d'au plus $\pm h$.
Une autre idée peut être de chercher les k plus proches voisins

$$\widehat{\varphi}(x) = \frac{\sum_{i=1}^{n} Y_i \mathbf{1}(X_i \in V_x)}{\sum_{i=1}^{n} \mathbf{1}(X_i \in V_x)} \text{ pour tout } x,$$

où V_x contient les k plus proches voisins de x.


distance de freinage
0
20
40
60
80
100
120
5
10
15
20
25
vitesse du véhicule


distance de freinage
0
20
40
60
80
100
120
5
10
15
20
25
vitesse du véhicule

De l'histogramme au regressogramme

Lai (1977) a montré la convergence de cet estimateur

Proposition 1. *Si $k \to \infty$, $k/n \to 0$ alors*

$$\mathbb{E}(\widehat{\varphi}_k(x)) \sim \varphi(x) + \frac{\varphi''(x)}{8}\frac{k^2}{n^2}$$

et

$$Var(\widehat{\varphi}_k(x)) \sim \frac{2\sigma^2(x)}{k}.$$

Aussi, en faisant un compromis entre biais et variance conduit à retenir $k \approx n^{4/5}$.
Sur notre exemple, $n = 50$ et $k = 22$.

La régression locale

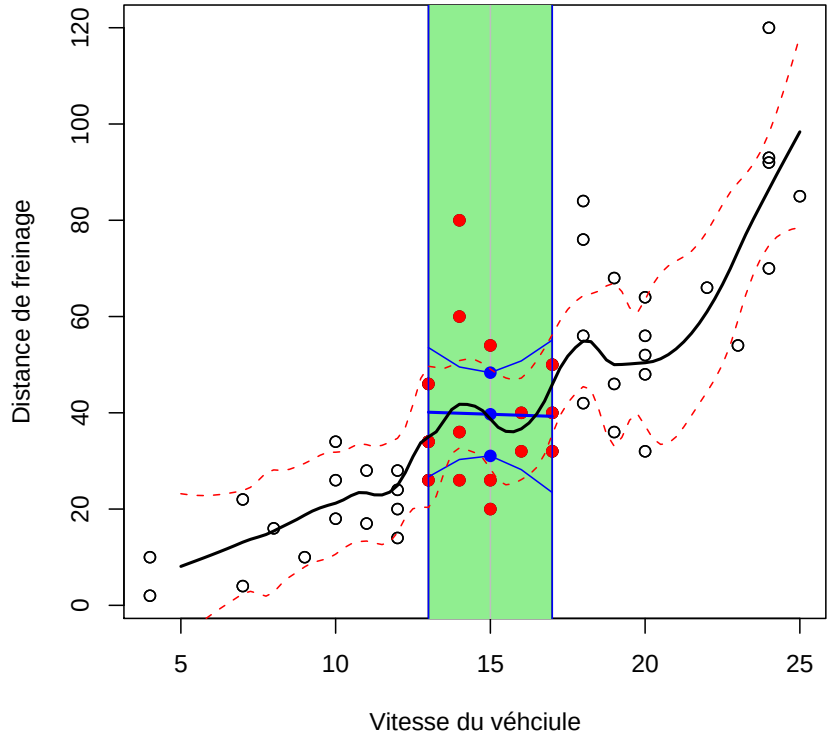
L'idée est ici simplement de considérer le voisinage en terme de nombre de voisines.

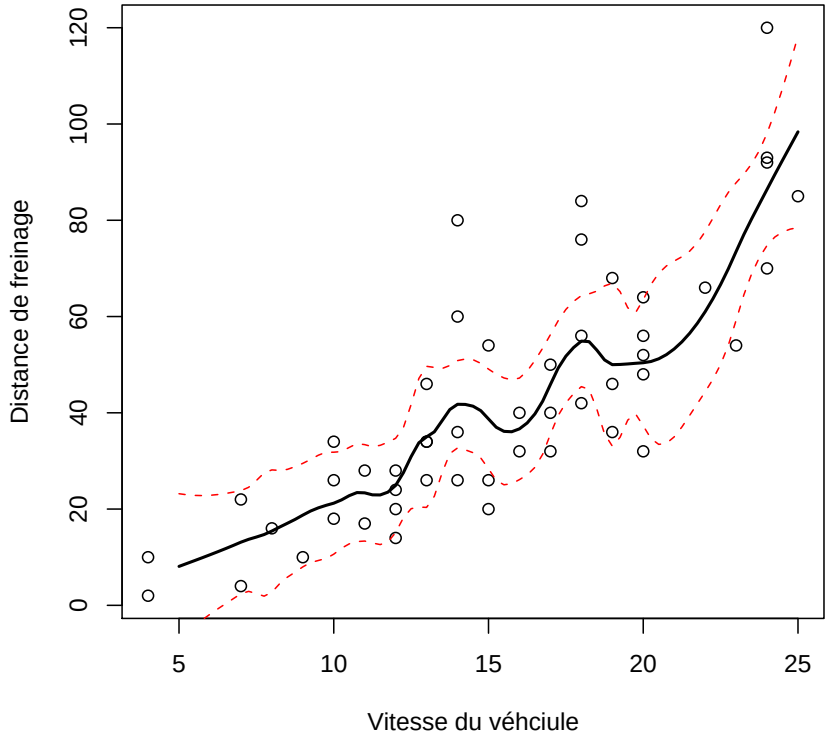
```
> loess(dist ~ speed, cars,span=0.75,degree=2)
> predict(REG, data.frame(speed = seq(5, 25, 0.25)), se = TRUE)
```

Le paramètre `span` correpond au pourcentage de points gardés pour faire l'ajustement local, et `degree` est le typoe de régression polynomiale.

La régression locale, méthode lowess

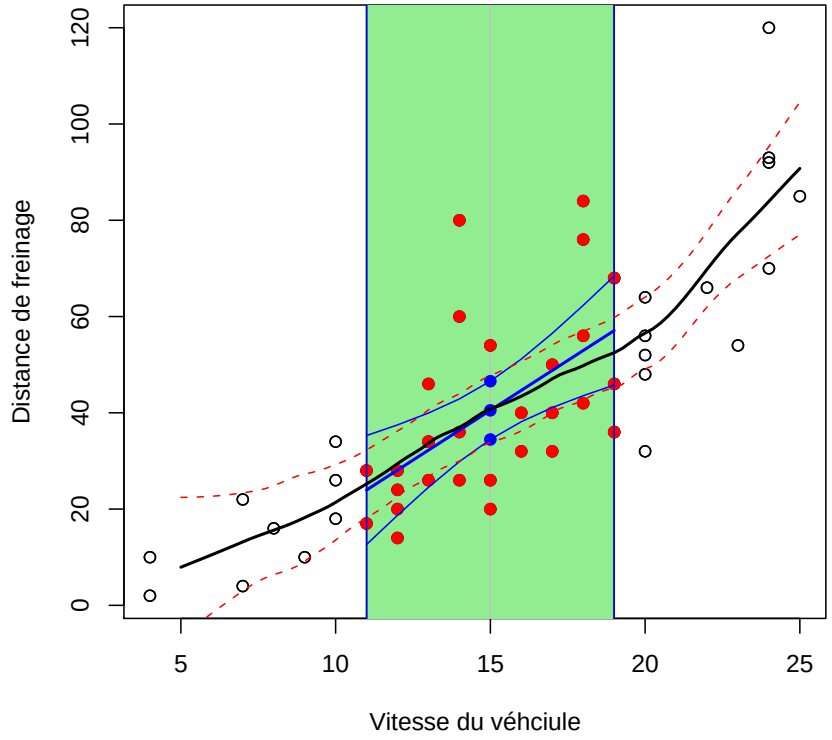
Ici ajustement local au voisinage de $x = 15$, avec 25% de points pour définir le voisinage (on garde 25% des points les plus proches, en x), et un ajustement linéaire.

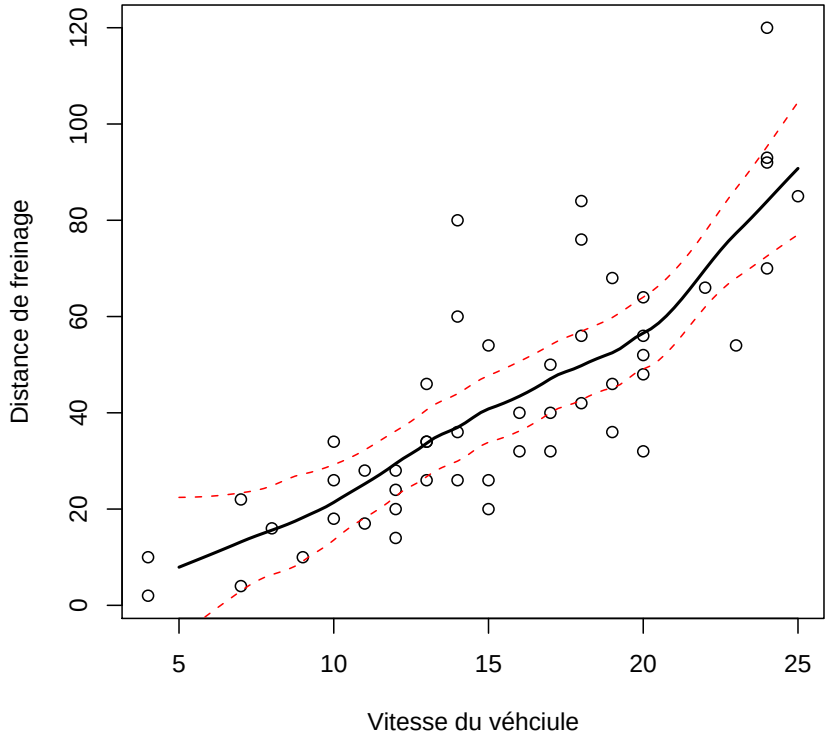




La régression locale, méthode lowess

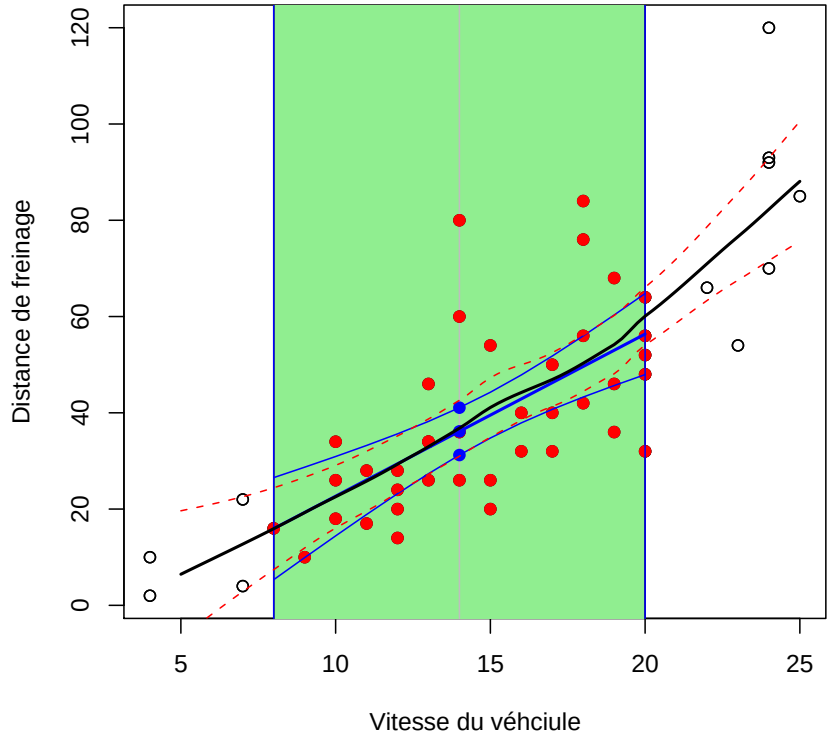
Ici ajustement local au voisinage de $x = 15$, avec 50% de points pour définir le voisinage (on garde 50% des points les plus proches, en x), et un ajustement linéaire.

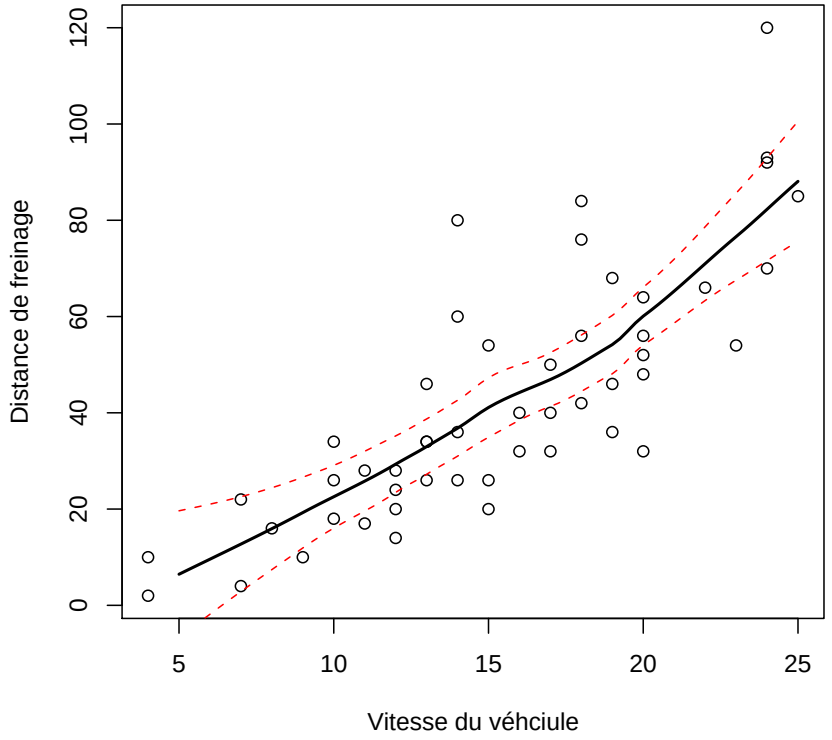




La régression locale, méthode lowess

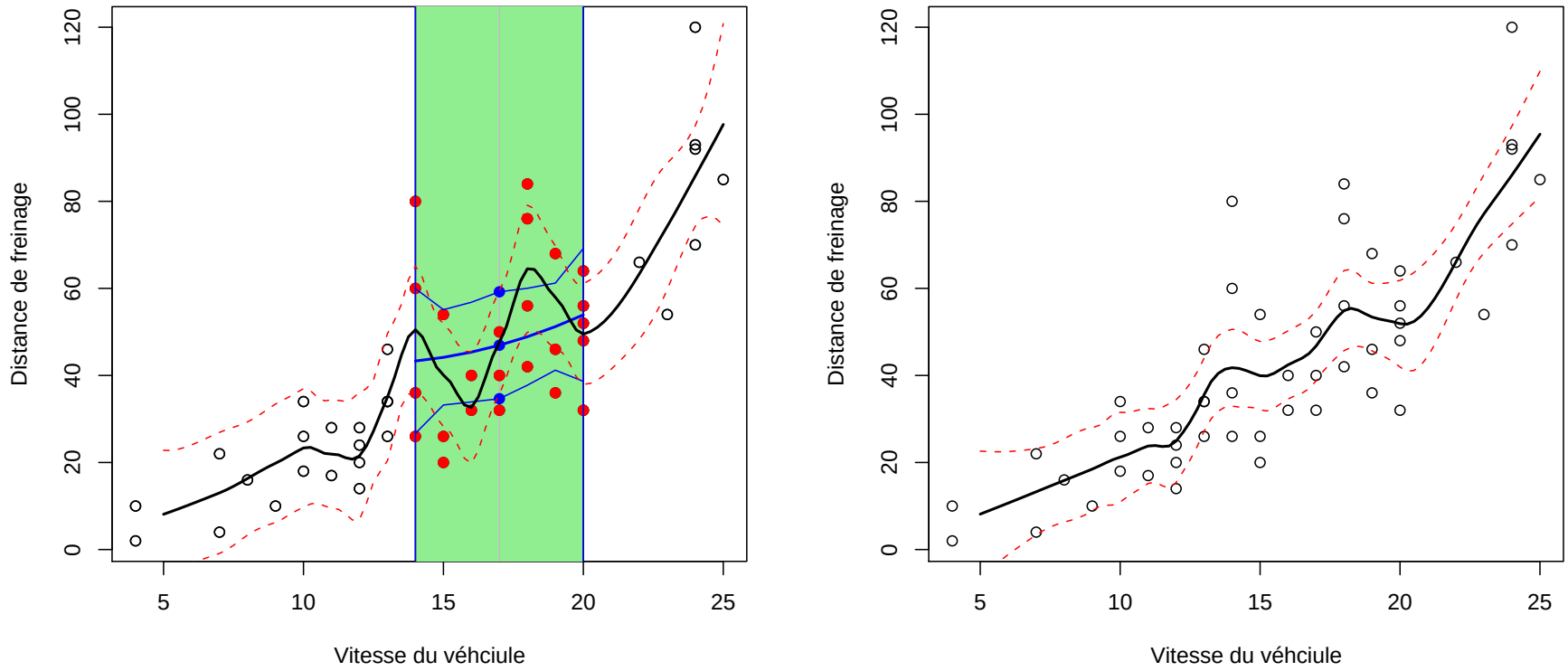
Ici ajustement local au voisinage de $x = 14$, avec 75% de points pour définir le voisinage (on garde 75% des points les plus proches, en x), et un ajustement linéaire.





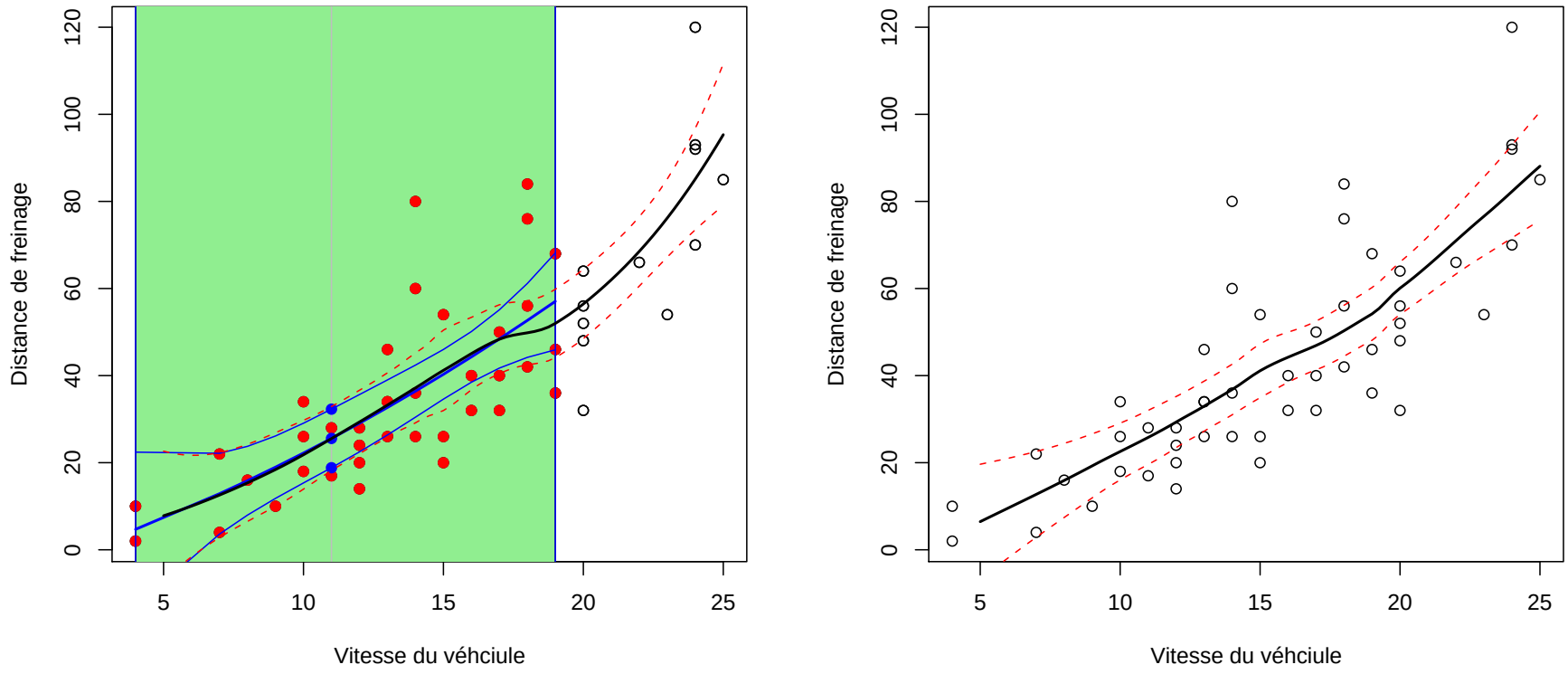
La régression locale, méthode lowess

Ici ajustement local au voisinage de $x = 17$, avec 35% de points pour définir le voisinage (on garde 35% des points les plus proches, en x), et un ajustement quadratique.



La régression locale, méthode lowess

Ici ajustement local au voisinage de $x = 11$, avec 75% de points pour définir le voisinage (on garde 35% des points les plus proches, en x), et un ajustement quadratique.



Du linéaire au nonlinéaire, une vision génŕale

Dans le cas linéaire, nous avions noté que

$$\widehat{Y} = Y - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \underbrace{\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'}_{H} Y$$

où H peut être interprétée comme une matrice de lissage. Notons que

$$\widehat{Y}_j = \sum_{i=1}^{n} [\boldsymbol{X}_j'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}']_i Y_i = \sum_{i=1}^{n} [\mathcal{H}(\boldsymbol{X}_j)]_i Y_i$$

où $\mathcal{H}(\boldsymbol{x}) = \boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Aussi, étant donnée une nouvelle observation $\boldsymbol{x} = (x_1, \cdots, x_k)$,

$$\mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x}) = \sum_{i=1}^{n} [\mathcal{H}(\boldsymbol{x})]_i Y_i = \widehat{Y}(\boldsymbol{x}).$$

Un estimateur sans biais de $\sigma^2 = Var(\varepsilon_i)$ est alors simplement

$$\widehat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2.$$

On en déduit alors simplement un intervalle de confiance pour Y sachant $\boldsymbol{X} = \boldsymbol{x}$, de la forme

$$\left[\widehat{Y}(\boldsymbol{x}) \pm z\widehat{\sigma}^2\|\mathcal{H}(\boldsymbol{x})\|^2\right], \text{ où } \|\mathcal{H}(\boldsymbol{x})\|^2 = \sum_{i=1}^{n}(\mathcal{H}(\boldsymbol{x})_i)^2$$

Definition 2. *Un prédicteur $\widehat{Y}(\boldsymbol{x})$ sera dit dit linéaire s'il pour tout $\boldsymbol{x}$, il existe un vecteur $\mathcal{S}(\boldsymbol{x}) = (\mathcal{S}(\boldsymbol{x})_1, \cdots, \mathcal{S}(\boldsymbol{x})_n)$ tel que*

$$\widehat{Y}(\boldsymbol{x}) = \sum_{i=1}^{n}\mathcal{S}(\boldsymbol{x})_i Y_i$$

Definition 3. *Si $[\widehat{Y}(\boldsymbol{X})_i]$ désine le vecteur $\widehat{Y}_1, \cdots, \widehat{Y}_n$. Il est alors possible d'écrire $[\widehat{Y}(\boldsymbol{X})_i] = SY$, où S est une matrice $n \times n$.*

Dans le cas du modèle linéaire $S = H$.

Examples de matrices de lissage S

Considérons le cas d'un moyenne glissante, i.e. on fait une moyenne locale sur les points distants (au plus d'une distance $h = 1$).

Sur les 9 observations suivantes,

X_i	1	2	3	4	5	6	7	8	9
Y_i	5	3	1	7	6	5	2	1	6

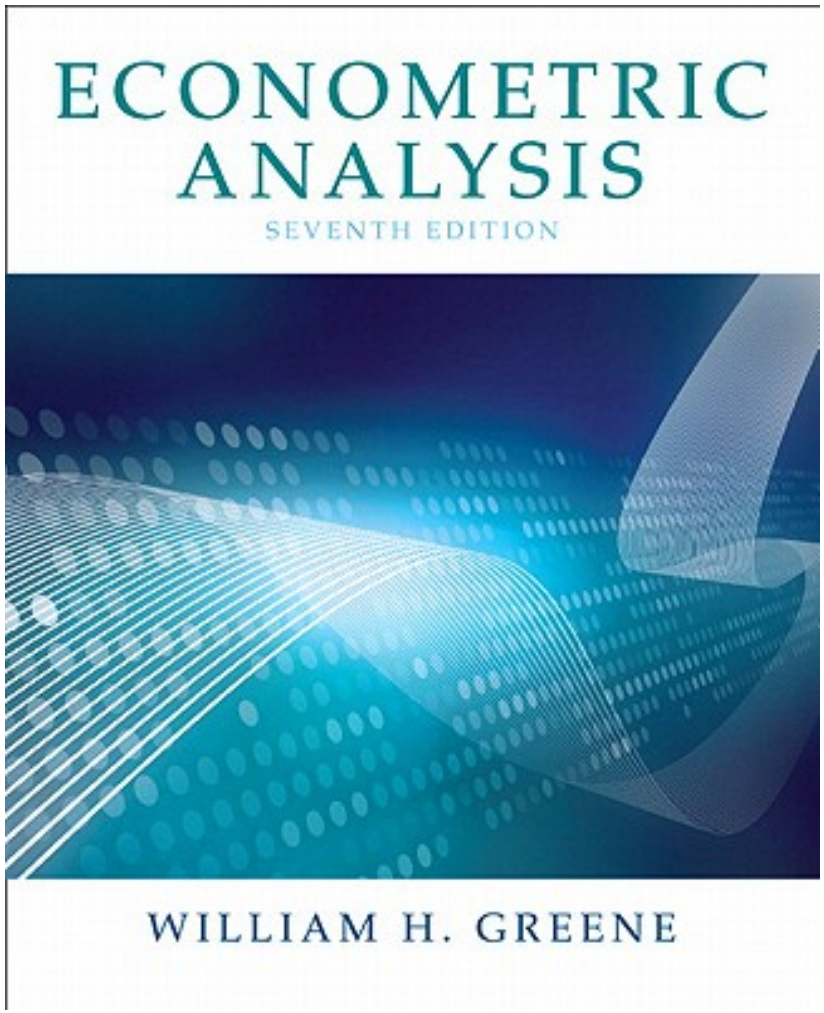
Alors

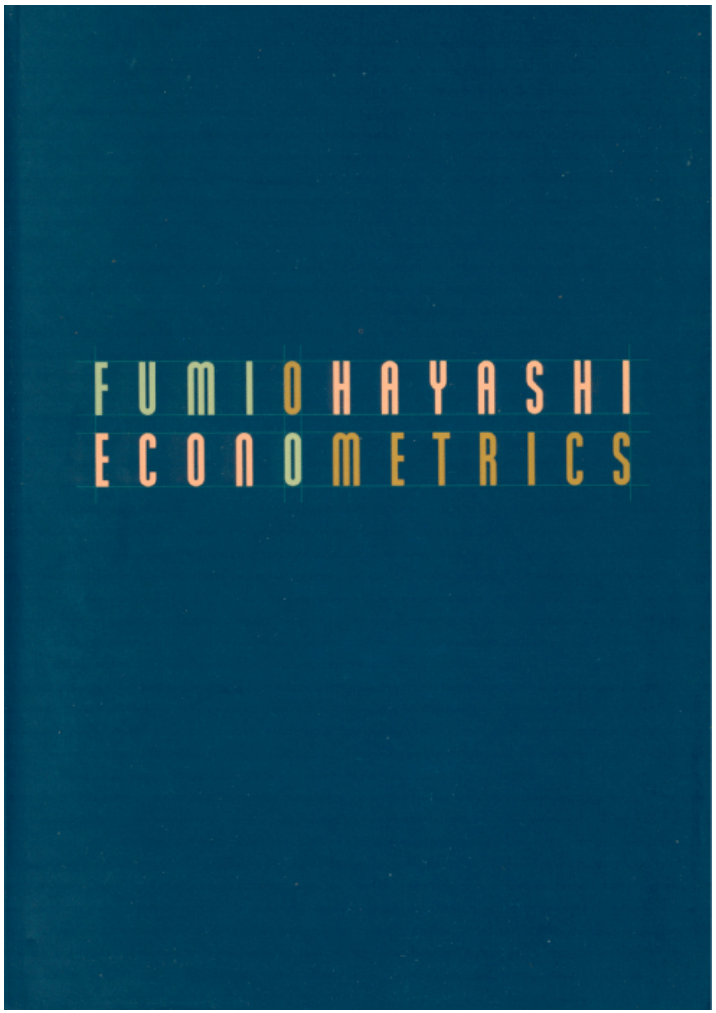
$$\widehat{Y}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n} Y_i \times \mathbf{1}(|X_i - x| \leq 1)}{\sum_{i=1}^{n} \mathbf{1}(|X_i - x| \leq 1)},$$

et

$$S_{h=1} = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1/3 & 1/3 & 1/3 & 0 & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & 1/3 & 1/3 & 1/3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 \end{pmatrix}$$

Quelques références





Quelques références