MAT7381 Chapitre 4
Régression linéaire simple

4.1 Estimateurs

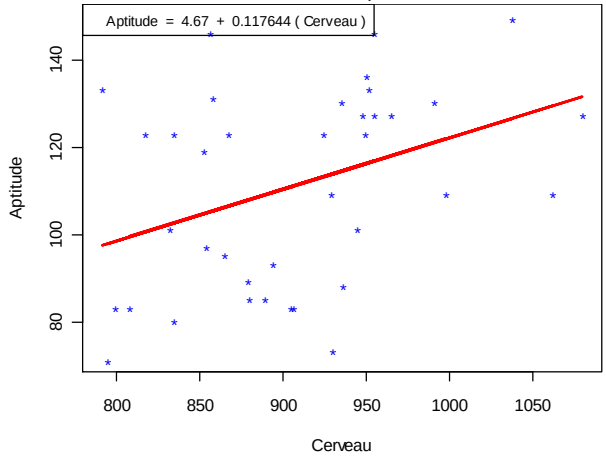
L'échantillon présenté au tableau 4.1.1 a été prélevé afin de déterminer s'il existe une relation entre l'intelligence et la grosseur du cerveau. Les variables observées sont les suivantes :

x : La grosseur du cerveau, mesurée en nombre de pixels observés par résonance magnétique
y : Le score au test d'aptitude de Wechsler, sous-test de « performance ».

Tableau 4.1.1
Grosseur du cerveau (*x*) *et aptitude mentale* (*y*) *d'un échantillon de* 38 *étudiants*

x	*y*	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
817	124	955	147	945	102	799	84	880	86
1038	150	834	124	808	84	1062	110	834	81
965	128	1080	128	889	86	794	72	948	128
952	134	924	124	906	84	867	124	949	124
929	110	856	147	791	134	858	132	894	94
991	131	879	90	955	128	950	137	930	74
854	98	865	96	832	102	998	110	936	89
905	84	852	120	935	131				

Figure 4.1.1
Graphique montrant la relation entre la grosseur du cerveau et l'aptitude mentale d'un échantillon de 38 *étudiants, et la droite des moindres carrés*



Présentées graphiquement dans la figure 4.1.1 sous la forme d'un *nuage de points*, ces données révèlent l'existence d'une certaine relation — plutôt faible, mais décelable — entre les deux variables.

Considérons le modèle linéaire classique

$$y_i = \beta_o + \beta_1 x_i + \varepsilon_i,\ i = 1, 2, \ldots, n$$

où les ε_i sont des variables aléatoires indépendantes de moyenne nulle et variance σ^2.

En termes matriciels,

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}\ ;\ \sigma^2\mathbf{I}) \tag{4.1.1}$$

où

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix},\ \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix},\ \boldsymbol{\beta} = \begin{bmatrix} \beta_o \\ \beta_1 \end{bmatrix},\ \text{et } \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Les coefficients $\hat{\beta}_o$ et $\hat{\beta}_1$ de la droite des moindres carrés $y = \hat{\beta}_o + \hat{\beta}_1 x$, minimisent

$$Q = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_o - \hat{\beta}_1 x_i)^2$$

Nous avons montré que ces coefficients sont donnés par

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{y} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} = \begin{bmatrix} 38 & 34456 \\ 34456 & 31436920 \end{bmatrix}^{-1} \begin{bmatrix} 4236 \\ 3859273 \end{bmatrix} = \begin{bmatrix} 4,6701 \\ 0,1176 \end{bmatrix}.$$

Voici une expression non matricielle pour ces estimateurs :

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{s_{xy}}{s_x^2}, \quad \text{et} \quad \hat{\beta}_o = \bar{y} - \hat{\beta}_1 \bar{x}$$

où S_{xy} est la covariance échantillonnale définie par

$$S_{xy} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Remarque Noter que la droite passe par le point $(\bar{x}\ ;\ \bar{y})$.

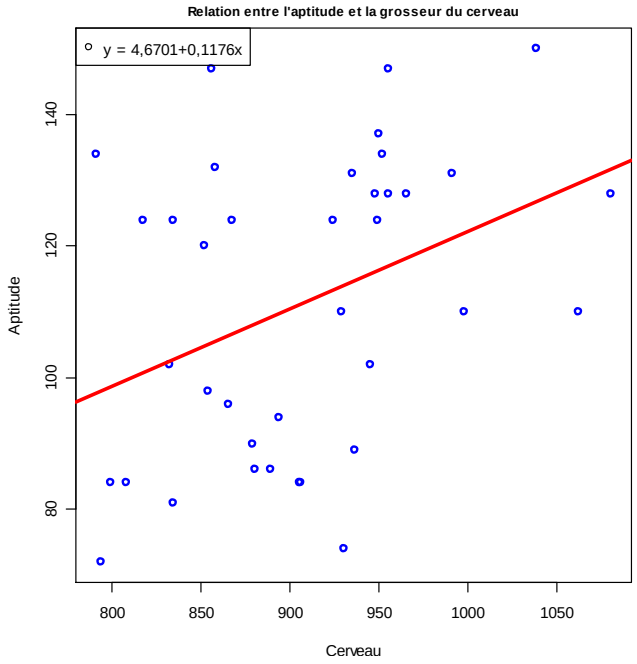
Nous avons donc $\hat{\beta}_o = 4{,}6701$ et $\hat{\beta}_1 = 0{,}1176$

$$y = 4{,}6701 + 0{,}1176x$$

Commandes R

Commandes pour tracer la droite de régression. Les commandes suivantes illustrent certaines des possibilités de R.

```
> cervInt(y~x)
> plot(y~x,ylab="Aptitude",xlab="Cerveau",type="n",main="Relation
  entre l'aptitude et la grosseur du cerveau",cex.main=.8)
> points(y~x,lwd=2,col="blue")
> abline(cervInt,lwd=3,col="red")
> legend("topleft","y = 4,6701+0,1176x",pch=1)
```



Le coefficient de corrélation
Le coefficient de corrélation est défini par

$$r = \frac{S_{xy}}{S_x S_y} \tag{1.2}$$

Le coefficient de corrélation est en fait la covariance entre les *cotes Z* des variables *X* et *Y*:

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{S_x}\right)\left(\frac{y_i - \bar{y}}{S_y}\right).$$

Dans l'exemple, le coefficient de corrélation est

$$r = 0{,}377$$

Remarques à propos du coefficient de corrélation.

1) $-1 \le r \le 1$, par l'inégalité de Cauchy-Schwartz.
2) $|r| = 1$ si et seulement si il existe des constantes *a* et *b* telles que $y_i = a + bx_i$ pour tout *i*, c'est-à-dire, si et seulement si les points du nuage se situent tous sur une même droite.
3) *r* a le même signe que $\hat{\beta}_1$ et $r = 0$ si et seulement si $\hat{\beta}_1 = 0$: $r = 0$ si et seulement si la droite des moindres carrés est horizontale.

Interprétation de σ^2 et *hypothèse d'homoscédasticité* σ^2 représente la variance des *y* correspondant à une même valeur de *x*. Donc un σ^2 petit caractérise un nuage dont les points sont rapprochés de la droite des moindres carrés. σ^2 est petite si la relation entre *x* et *y* est forte, grande sinon.

Propriétés des estimateurs

Nous savons que $\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_o \\ \hat{\beta}_1 \end{bmatrix}$ est de loi normale bivariée de moyenne $\boldsymbol{\beta}$ et de matrice de covariance

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}\begin{bmatrix} 1 & -\bar{x} \\ -\bar{x} & \sum x_i^2/n \end{bmatrix},$$

donc

$$\hat{\beta}_1 \sim \mathcal{N}(\beta_1\,;\sigma^2_{\hat{\beta}_1}) \text{ où } \sigma^2_{\hat{\beta}_1} = \frac{\sigma^2}{\sum(x_i-\bar{x})^2}$$

$$\hat{\beta}_o \sim \mathcal{N}(\beta_o\,;\sigma^2_{\hat{\beta}_o}) \text{ où } \sigma^2_{\hat{\beta}_o} = \frac{\sigma^2\sum x_i^2}{n\sum(x_i-\bar{x})^2} = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$$

$$\text{Cov}(\hat{\beta}_o\,;\,\hat{\beta}_1) = -\frac{\bar{x}\sigma^2}{\sum(x_i-\bar{x})^2}$$

Remarques

1) *Comme on doit s'y attendre, les estimateurs sont d'autant plus précis que n est grand : les formules des variances des estimateurs sont des fonctions décroissantes de n.*
2) *Les variances des estimateurs croissent avec* σ^2. *Une faible valeur de* σ^2 *rend plus précises non seulement les prédictions, mais les estimateurs aussi. Plus la dépendance entre x et y est forte, plus* σ^2 *est petit, et meilleurs sont les estimateurs.*
3) *La dispersion des x,* S_x^2 *, joue un rôle important dans la qualité des estimateurs : plus les x sont dispersés, plus les estimateurs sont efficaces.*

Estimation de σ^2***, de*** $\sigma^2_{\hat{\beta}_o}$ ***et de*** $\sigma^2_{\hat{\beta}_1}$

L'estimateur de σ^2, $\hat{\sigma}^2 = \dfrac{\boldsymbol{y}'[\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}']\boldsymbol{y}}{n-q}$ devient ici

$$\hat{\sigma}^2 = \frac{\sum(y_i-\hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n}\left[y_i-(\hat{\beta}_o+\hat{\beta}_1x_i)\right]^2}{n-2}.$$

Remarque *On peut justifier cet estimateur par analogie avec l'estimateur* S^2 *d'une variance* σ^2 *dans un échantillon aléatoire simple. On sait que* σ^2 *est la variance des* $\varepsilon_i = y_i - (\beta_o + \beta_i x_i)$. *Une estimation naturelle aurait donc été une moyenne des* ε_i^2 *, soit* $\dfrac{\sum\varepsilon_i^2}{n}$ *. Mais les* ε_i *n'étant pas connus, il faut les remplacer par* $\hat{\varepsilon}_i = y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)$ *: on remplace deux paramètres par leur estimation. C'est ce qui explique que le dénominateur est n et pas n-1. La situation est analogue à celle où l'on estime* σ^2 *à partir d'un échantillon de n variables indépendantes de* même *moyenne* μ. *On l'estime par* $S^2 = \dfrac{\sum(x_i-\bar{x})^2}{n-1}$ *alors qu'on l'aurait estimée par* $\dfrac{\sum(x_i-\mu)^2}{n}$ *si* μ *était connue.* ■

Une fois σ^2 estimée, il suffira, pour estimer les variances de $\hat{\beta}_1$ et de $\hat{\beta}_o$, de remplacer σ^2 par $\hat{\sigma}^2$ dans les formules. On obtient alors

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{\hat{\sigma}^2}{\sum(x_i-\bar{x})^2} \text{ et } \hat{\sigma}^2_{\hat{\beta}_o} = \hat{\sigma}^2\left(\frac{1}{n}+\frac{\bar{x}^2}{\sum(x_i-\bar{x})^2}\right)$$

Remarque *La formule de* $\hat{\sigma}^2 = \dfrac{\sum(y_i - \hat{y}_i)^2}{n-2}$ *est équivalente formules suivantes* :

$$\hat{\sigma}^2 = \frac{\sum(y_i - \bar{y})^2}{n-2} - \hat{\beta}_1^2 \frac{\sum(x_i - \bar{x})^2}{n-2} = \frac{\sum(y_i - \bar{y})^2}{n-2}(1 - r^2)$$

On y voit que lorsque $\beta_1 = 0$ (ou $r = 0$), $\hat{\sigma}$ est à toute fin pratique égale à S_y.

Commandes R

Voici les commandes nécessaires pour effectuer les opérations décrite dans cette section. Les données qui les illustrent sont celles du tableau 4.1.1. On suppose les vecteurs `x` et `y` déjà définie

Voici les commandes qui font tracer le graphique de la figure 4.4.4 :

```
> plot(y~x,ylab="Aptitude",xlab="Cerveau",type="n")
> points(y~x,lwd=3,col="blue")
```

Commandes **R** pour tracer la droite de régression. Les commandes suivantes illustrent certaines des possibilités de **R**.

```
> plot(y~x,ylab="Aptitude",xlab="Cerveau",type="n")
> ychapeau<-cervInt$fitted.values
> points(y~x,lwd=3,col="blue")
> lines(ychapeau~x,lwd=3,col="red")
> legend("topleft","y = 4,6701+0,1176x",pch=1)
```

Il existe, cependant, des commandes conçues spécialement pour tracer une droite de régression :

```
> plot(y~x,ylab="Aptitude",xlab="Cerveau",type="n")
> points(y~x,lwd=3,col="blue")
> abline(cervInt,lwd=3,col="red")
> legend("topleft","y = 4,6701+0,1176x",pch=1)
```

Calcul du coefficient de corrélation :

```
> cor(x,y)
[1] 0.3773496
```

La commande pour effectuer la régression est `lm()`. Nous nommons `cervInt` l'objet produit par cette commande :

```
> cervInt <-lm(y~x)
> cervInt
(Intercept)            x
    4.6701         0.1176
```

Un appel à `cervInt` ne fournit que les coefficients $\hat{\beta}_o$ et $\hat{\beta}_1$. Mais `cervInt` contient de plus amples informations, fournies par un appel à `summary(cervInt)`. Voici, en partie ce qui découle de cette commande :

```
> summary(cervInt)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.67009   43.76697   0.107   0.9156
x            0.11764    0.04812   2.445   0.0195 *

Residual standard error: 21.22 on 36 degrees of freedom
Multiple R-squared: 0.1424,     Adjusted R-squared: 0.1186
F-statistic: 5.977 on 1 and 36 DF,  p-value: 0.01952
```

La colonne `Estimate` donne les estimations $\hat{\beta}_o$ et $\hat{\beta}_1$; la colonne `Std. Error` donne les écarts-types estimés $\hat{\sigma}_{\hat{\beta}_o}$ et $\hat{\sigma}_{\hat{\beta}_1}$. Les deux dernières colonnes du tableau seront expliquées dans la prochaine section.

L'estimation $\hat{\sigma}$ est donnée par `Residual standard error: 21.22`. Le carré du coefficient de corrélation est donnée par `Multiple R-squared: 0.1424`.

La commande

```
> ls.diag(cervInt)
```

fournit de nombreux indices, précisés par un suffixe précédé du signe « $ ». Ainsi,

```
> ls.diag(cervInt)$cov.scaled
            (Intercept)           x
(Intercept)  1915.54804 -2.099509847
x              -2.09951  0.002315457
```

donne la matrice de covariance de $\hat{\boldsymbol{\beta}}$.

4.2 Intervalles de confiance et tests d'hypothèses

On voudra normalement estimer et tester des hypothèses concernant les composantes de $\boldsymbol{\beta}$. Mais on voudra également estimer certaines fonctions linéaires $\boldsymbol{\ell}'\boldsymbol{\beta}$ de $\boldsymbol{\beta}$. Nous discutons donc ce cas général qui comprend entre autres les coefficients β_o et β_1. Le résultat suivant découle de la théorie du modèle linéaire général.

Soit $\eta = \boldsymbol{\ell}'\boldsymbol{\beta}$ une fonction linéaire de $\boldsymbol{\beta}$, où $\boldsymbol{\ell}$ est un vecteur fixe. L'estimateur $\hat{\eta} = \boldsymbol{\ell}'\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\ell}'\boldsymbol{\beta}$ est sans biais ; sa variance $\sigma^2_{\hat{\eta}} = \sigma^2\boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell}$ est estimée par $\hat{\sigma}^2_{\hat{\eta}} = \hat{\sigma}^2\boldsymbol{\ell}'(\mathbf{X}'\mathbf{X})^{-1}\boldsymbol{\ell}$, et alors

$$\frac{\hat{\eta} - \eta}{\hat{\sigma}_{\hat{\eta}}} \sim t_{n-2}$$

Deux cas particuliers de cet énoncé, les paramètres β_o et β_1 eux-mêmes, donnent ceci :

$$\frac{\hat{\beta}_o - \beta_o}{\hat{\sigma}_{\hat{\beta}_o}} \sim t_{n-2} \text{ et } \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

Intervalles de confiance

Les intervalles de confiance sont données par

$$\hat{\beta}_o - t_{n-2;\alpha/2}\sigma_{\hat{\beta}_o} \le \beta_o \le \hat{\beta}_o + t_{n-2;\alpha/2}\sigma_{\hat{\beta}_o} \text{ et } \hat{\beta}_1 - t_{n-2;\alpha/2}\sigma_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{n-2;\alpha/2}\sigma_{\hat{\beta}_1}$$

où $t_{n-2;\alpha/2}$ est le point tel que, pour une variable T de loi de S*tudent* à n-2 degrés de liberté,

$$P(T > t_{n-2;\alpha/2}) = \alpha/2.$$

Il est rare qu'on ait à tester des hypothèses concernant β_o; mais on s'intéressera toujours à β_1, normalement pour tester l'hypothèse que $\beta_1 = 0$.

Tests d'hypothèses

On voudra invariablement tester l'hypothèse que la droite de régression est horizontale, c'est-à-dire, qu'il n'y a pas de relation entre les deux variables. Considérons l'hypothèse légèrement plus générale

$$H_o: \beta_1 = b$$

où b est un nombre donné. La procédure consiste à rejeter H_o si $\hat{\beta}_1$ s'écarte trop de b. On mesure l'écart entre $\hat{\beta}_1$ et b par $|T_1|$, où $T_1 = \dfrac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}$, et on rejette H_o si cet écart est supérieur au point critique $t_{n-2;\alpha/2}$.

La règle est donc

On rejette l'hypothèse que $\beta_1 = b$ *si et seulement si* $|T_1| > t_{n-2;\alpha/2}$. (4.2.1)

De même, la statistique pour tester l'hypothèse

$$H_o: \beta_o = a$$

est $|T_o|$, où $T_o = \dfrac{\hat{\beta}_o - a}{\hat{\sigma}_{\hat{\beta}_o}}$. La règle est :

On rejette l'hypothèse que $\beta_o = a$ *si et seulement si* $|T_o| > t_{n-2;\alpha/2}$. (4.2.2)

La colonne `t value`, dans le tableau suivant, fourni par la commande `summary(cervInt)`, présente les valeurs de $T_1 = \dfrac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$ et de $T_o = \dfrac{\hat{\beta}_o}{\hat{\sigma}_{\hat{\beta}_o}}$.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.67009   43.76697   0.107   0.9156
x            0.11764    0.04812   2.445   0.0195 *
```

La colonne `Pr(>|t|)` donne les valeurs p correspondantes à T_1 et T_o.

Un autre cas particulier important

La fonction $\mu_x = \beta_o + \beta_1 x$ est un cas particulier important. Elle représente la moyenne des y qui correspondent à une valeur donnée x. Son estimateur est $\hat{\mu}_x = \hat{\beta}_o + \hat{\beta}_1 x$. La variance de l'estimateur, qui est donnée par

$$\sigma^2_{\hat{\mu}_x} = \sigma^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right], \text{ estimée par } \hat{\sigma}^2_{\hat{\mu}_x} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum(x_i-\overline{x})^2}\right] \quad (4.2.4)$$

et la statistique

$$T = \frac{\hat{\mu}_x - \mu_x}{\hat{\sigma}_{\hat{\mu}_x}}$$

suit une loi de Student à n-2 degrés de liberté : $T \sim t_{n-2}$.

On peut alors obtenir un intervalle de confiance pour μ_x par la formule usuelle

$$\hat{\mu}_x - t_{n-2;\alpha/2}\,\hat{\sigma}_{\hat{\mu}_x} \le \mu_x \le \hat{\mu}_x + t_{n-2;\alpha/2}\,\hat{\sigma}_{\hat{\mu}_x} \quad (4.2.5)$$

On peut également tester l'hypothèse que cette moyenne prend une certaine valeur donnée μ_o,

$$H_o: \mu_x = \mu_o$$

La règle est

$$\textit{On rejette l'hypothèse que } \mu_x = \mu_o \text{ si et seulement si } \left| \frac{\hat{\mu}_x - \mu_o}{\hat{\sigma}_{\hat{\mu}_x}} \right| > t_{n-2;\alpha/2} \qquad (4.2.6)$$

Limites de prédiction

L'intervalle de confiance pour μ_x est une affirmation concernant la *moyenne* des *y* qui correspondent à une valeur donnée *x*. Ce n'est pas une prédiction concernant une valeur future de *y*. La *prédiction* $\hat{y}_x$ de la prochaine valeur de *y* correspondant à *x* (qui ne diffère pas de l'estimation de μ_x),

$$\hat{y}_x = \hat{\beta}_o + \hat{\beta}_1 x \ (= \hat{\mu}_x)$$

suit une loi normale, et l'écart $y - \hat{y}_x$ a pour variance

$$\sigma^2_{y-\hat{y}_x} = \sigma^2 + \sigma^2_{\hat{\mu}_x} = \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right) \Rightarrow \sigma_{y-\hat{y}_x} = \sigma\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \qquad (4.2.7)$$

d'où

$$\frac{y-\hat{y}_x}{\sigma_{y-\hat{y}_x}} \sim \mathcal{N}(0\,;\,1)$$

On estime $\sigma_{y-\hat{y}_x}$ par

$$\hat{\sigma}_{y-\hat{y}_x} = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2_{\hat{\mu}_x}} = \hat{\sigma}\sqrt{1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}} \qquad (4.2.8)$$

et par suite,

$$\frac{y-\hat{y}_x}{\sigma_{y-\hat{y}_x}} \sim t_{n-2}$$

Les *limites de prédiction* à 100(1 - α)% sont

$$\hat{y}_x - t_{n-2;\alpha/2}\,\hat{\sigma}_{y-\hat{y}_x} \le y_x \le \hat{y}_x + t_{n-2;\alpha/2}\,\hat{\sigma}_{y-\hat{y}_x} \qquad (4.2.9)$$

Là on peut affirmer avec 100(1-α)% de sécurité que la prochaine observation y_x se situera entre les deux bornes.

Commandes R .

Voici comment obtenir un intervalle de confiance à 95 % pour les moyennes μ_{800}, μ_{900}, μ_{1000} :

```
> x0<-c(800,900,1000)
> predict(cervInt,data.frame(x=x0),interval="confidence",level=.95)
        fit       lwr      upr
1  98.78517  86.24628 111.3241
2 110.54956 103.53863 117.5605
3 122.31394 110.84400 133.7839
```

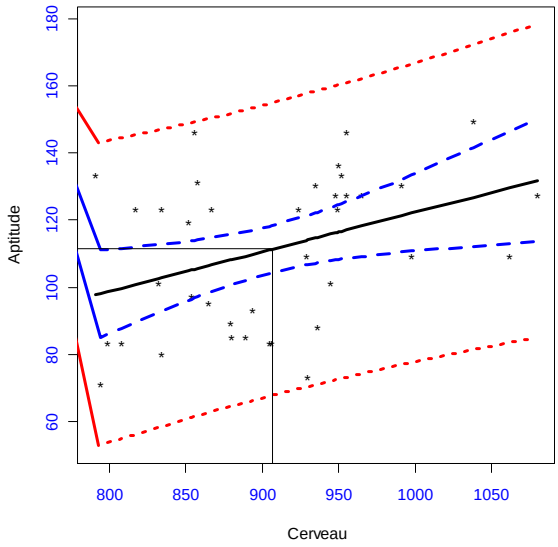
Et voici comment obtenir des limites de prédiction à 95 % pour y_{800}, y_{900}, y_{1000} :

```
> predict(cervInt,data.frame(x=x0),interval="prediction",level=.95)
        fit       lwr      upr
```

```
1  98.78517 53.96757 143.6028
2 110.54956 66.95430 154.1448
3 122.31394 77.78358 166.8443
```

Le programme suivant permet d'obtenir un graphique de la droite des moindres carrés, les intervalles de confiance et les limites de prédiction. On définit d'abord une matrice **A** dont les colonnes sont les valeurs de *x* et de *y*, dans cet ordre.

```
function (A,niveau)
{x<-A[,1]
y<-A[,2]
y<-y[order(x)]
x<-sort(x)
a<-lm(y~x)
xbar<-mean(x)
ybar<-mean(y)
yhat<-a$fitted.values
ic<-predict(a,data.frame(x=x),level=niveau,interval="confidence")
pred<-predict(a,data.frame(x=x),level=niveau,interval="prediction")
par(mar=c(4,4,1,0))
plot(y~x,pch="*",xlab=colnames(A)[1],ylab=colnames(A)[2],col.axis="blue",xlim=c(min(x)
,max(x)),ylim=c(min(pred[,2]),max(pred[,3])))
lines(ic[,1]~x,col="black",lwd=3,lty=1)
lines(ic[,2]~x,col="blue",lwd=3,lty=2)
lines(ic[,3]~x,col="blue",lwd=3,lty=2)
lines(pred[,2]~x,col="red",lwd=3,lty=3)
lines(pred[,3]~x,col="red",lwd=3,lty=3)
lines(c(xbar,xbar),c(0,ybar))
lines(c(0,xbar),c(ybar,ybar))
}
```



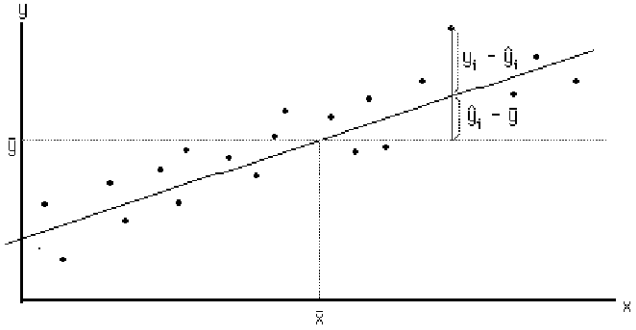
4.3 Analyse de variance et test de l'hypothèse H_1 par la loi de Fisher

La somme des carrés $\sum_i (y_i - \bar{y})^2$, que nous appelons « somme des carrés *totale* » et désignons par SCT est une mesure de la dispersion totale des *y* dans l'échantillon, indépendamment des *x*. Cette somme de carrés peut être décomposée en deux parties. La première, $\sum_i (\hat{y}_i - \bar{y})^2$, appelée « somme des carrés *expliquée* » et notée SCE, est la partie de la dispersion des *y* qui est attribuable à la dispersion des *x*, donc

"expliquée" par *x*. La deuxième, $\sum_i (y_i - \hat{y}_i)^2$, appelée "somme des carrés résiduelle" et désignée par SCR, est la partie de la dispersion totale des *y* que l'on ne peut *pas* attribuer aux variations des *x*. On peut démontrer que

$$\sum_i (y_i - \bar{y})^2 = \sum_i (\hat{y}_i - \bar{y})^2 + \sum_i (y_i - \hat{y}_i)^2 \qquad (4.3.1)$$

$$\text{SCT} = \text{SCE} + \text{SCR}$$



Graphiquement, SCE est la somme des carrés des distances verticales entre les points sur la droite des moindres carrés $\hat{y} = \hat{\beta}_o + \hat{\beta}_1 x$ et les points sur la droite horizontale $y = \bar{y}$. Cette somme de carrés a tendance à être petite si la droite des moindres carrés s'approche d'une droite horizontale, c'est-à-dire, si les données ne témoignent pas d'une forte dépendance entre *y* et *x*. SCR est la somme des distances verticales entre les points du nuage et la droite des moindres carrés. Cette somme de carrés a tendance à être petite si les points sont rapprochés de la droite des moindres carrés, cas où la dépendance entre *y* et *x* est forte.

Remarques

1. SCR *et* $\hat{\sigma}^2$ *sont liés par la relation suivante :*

$$\hat{\sigma}^2 = \frac{\text{SCR}}{n-2} \qquad (4.3.2)$$

Donc SCR *petit signifier que les* y_i *ont tendance à être peu dispersés par rapport à leur moyenne* $\beta_o + \beta_1 x_i$*, ce qui se manifeste dans l'échantillon par un nuage de points rapproché de la droite des moindres carrés. Nous avons aussi la relation suivante entre* $\hat{\beta}_1$ *et* SCR*:*

$$\text{SCR} = \sum (y_i - \bar{y})^2 - \hat{\beta}_1^2 \sum (x_i - \bar{x})^2 \qquad (4.3.3)$$

2. *Le carré du coefficient de corrélation est lié à ces termes par l'équation*

$$r^2 = 1 - \frac{\text{SCR}}{\text{SCT}} \qquad (4.3.4)$$

C'est ce qui justifie l'énoncé suivant : $100r^2$ *est le pourcentage de la variation totale qui est* expliquée *par la variable indépendante.*

3. SCE *et* $\hat{\beta}_1$ *sont liés par l'équation suivante* :

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

Donc SCE *petit signifie que* $|\hat{\beta}_1|$ *est petit, et par conséquent que la droite est près d'être horizontale.*

La décomposition est traditionnellement présentée sous la forme d'une table appelée table d'analyse de variance*, dans laquelle on indique aussi les « moyennes » de carrés, c'est-à-dire, les sommes de carrés divisées par le nombre de degrés de liberté.*

Table d'analyse de variance

Source	*Somme des carrés*	*d.ℓ.*	*Moyenne des carrés*	*Espérances*
Régression	SCE = $\sum(\hat{y}_i - \bar{y})^2$	1	MCE = SCE/1	E(MCE)= $\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
Résiduelle	SCR = $\sum(y_i - \hat{y}_i)^2$	n - 2	MCR = $\frac{\text{SCR}}{n-2} = \hat{\sigma}^2_{y.x}$	E(MCR) = σ^2
Total	SCT = $\sum(y_i - \bar{y})^2$	n - 1	MCT = $\frac{\text{SCT}}{n-1} = s_y^2$	E(MCT)= $\sigma^2 + \beta_1^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Nous avons présenté l'espérance des moyennes des carrés (qui ne font habituellement pas partie d'une table d'analyse de variance) car elles motivent le choix d'une certaine statistique F qui peut aussi bien servir à tester l'hypothèse $H_1 : \beta_1 = 0$ contre une alternative bilatérale.

Le test décrit ci-dessus pour cette hypothèse est basé sur la statistique $T = \hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}$ et la région critique est $|T| > t_{n\text{-}2;\alpha/2}$. Nous montrerons que

$$|T| > t_{n\text{-}2;\alpha/2} \Leftrightarrow \left(\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}\right)^2 > F_{1;n\text{-}2;\alpha} \Leftrightarrow \frac{\text{MCE}}{\text{MCR}} > F_{1;n\text{-}2;\alpha.}$$

En effet, il est évident que $|T| > t_{n\text{-}2;\alpha/2} \Leftrightarrow T^2 > t^2_{n-1;\alpha/2}$. Or $T^2 = \left(\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}\right)^2$ étant, sous H_o, une variable de loi $t_{n\text{-}2}$, sont carré est une statistique de loi F à 1 et n-2 degrés de liberté, ce qui entraîne que $t^2_{n-1;\alpha/2} = F_{1;n\text{-}2;\alpha}$.

Nous avons donc la première équivalence, $|T| > t_{n\text{-}2;\alpha/2} \Leftrightarrow \left(\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}\right)^2 > F_{1;n\text{-}2;\alpha}$. La deuxième équivalence, $\left(\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}\right)^2 > F_{1;n\text{-}2;\alpha} \Leftrightarrow \frac{\text{MCE}}{\text{MCR}} > F_{1;n\text{-}2;\alpha}$ découle du fait que $\left(\hat{\beta}_1 / \hat{\sigma}_{\hat{\beta}_1}\right)^2 = \frac{\text{MCE}}{\text{MCR}}$, ce qu'on peut démontrer par de simples manipulations algébriques. On peut alors justifier la région critique $\frac{\text{MCE}}{\text{MCR}} > F_{1;n\text{-}2;\alpha}$ en notant que l'espérance du numérateur, $\sigma^2 + \beta_1^2 \sum(x_i - \bar{x})^2$ est supérieure à celle du dénominateur (qui est σ^2) et ne lui est égal que si H_o est vraie.

Commande R :

Une seule commande R suffit à obtenir la table d'analyse de variance :

```
> anova(cervInt)
Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x          1  2690.4 2690.45  5.9773 0.01952 *
Residuals 36 16204.1  450.11
```

Selon cette table, MCE = 2690,45, MCR = 450,11, et F = MCE/MCR = 5,9773. Pour une variable F de loi $\mathcal{F}_{1;36}$, $P(F > 5{,}9773) = 0{,}01952$, ce qui mène au rejet de l'hypothèse que $\beta_1 = 0$ si notre seuil est, par exemple, $\alpha = 0{,}05$.

4.4 Techniques diagnostiques

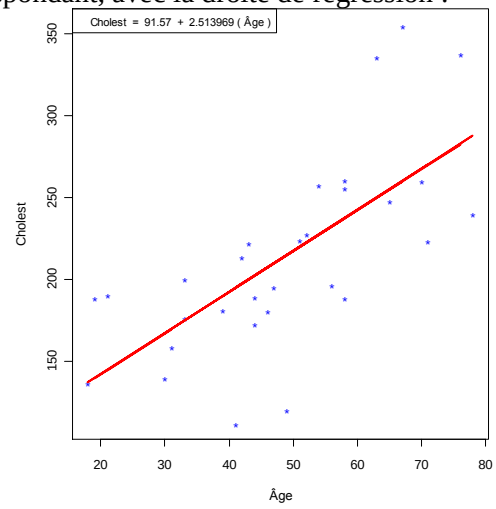
Il existe plusieurs indices qui aident à juger dans quelle mesure les données sont conformes aux hypothèses du modèle. Nous illustrons les analyses faites à partir des données du tableau suivant, qui présente l'age (`Age`) et le taux de cholestérol total (`Cholest`) de 30 sujets. Il s'agit d'une régression linéaire simple, mais la plupart des techniques présentées ici s'appliquent également à la régression multiple.

Tableau 4.4.1

Relation entre le cholestérol et l'age

Age	*Cholest*	*Age*	*Cholest*	*Age*	*Cholest*	*Age*	*Cholest*	*Age*	*Cholest*
46	181	41	112	44	190	70	261	49	121
71	224	43	223	58	262	76	339	33	201
51	225	47	196	78	241	54	259	33	177
30	140	42	214	65	249	44	173	19	189
56	197	39	182	18	137	63	337	31	159
52	228	58	189	58	257	67	356	21	191

Voici le nuage de points correspondant, avec la droite de régression :




```
> reglin<-lm(Cholest~Âge)
> summary(reglin)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  91.5716    25.6526   3.570  0.00131 **
Âge           2.5140     0.5014   5.013 2.67e-05 ***

Residual standard error: 44.14 on 28 degrees of freedom
Multiple R-squared: 0.473,      Adjusted R-squared: 0.4542
F-statistic: 25.13 on 1 and 28 DF,  p-value: 2.673e-05
```

Les questions auxquelles nous tentons de répondre avec les techniques diagnostiques sont:

- La relation est-elle linéaire ?
- L'hypothèse d'homoscédasticité est-elle vérifiée?
- Les ε_i sont-elles distribuées selon une loi normale?

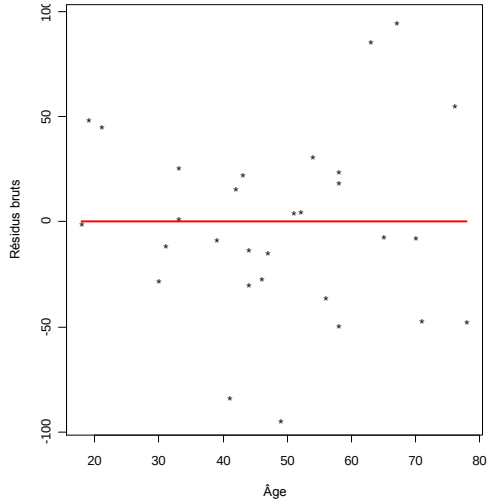
Résidus

Ce sont les *résidus* qui aideront à y répondre, c'est-à-dire, les quantités

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_o + \hat{\beta}_1 x_i)$$

Ces résidus sont présentés dans le nuage de points ci-dessus en fonction de *x*. Le but de certaines des méthodes graphiques présentées ici est de mettre en relief des anomalies qui sont moins évidentes dans un simple nuage de points.

Examen visuel des résidus en fonction de Âge:



Résidus centrés-réduits :
La matrice

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \qquad (4.4.1)$$

joue un rôle important dans l'analyse des résidus, entre autres parce que

$$V(\hat{\boldsymbol{y}}) = \sigma^2\mathbf{H} \text{ et } V(\hat{\varepsilon}_i) = \sigma^2(\mathbf{I}-\mathbf{H}).$$

Donc la variance de $\hat{\varepsilon}_i$ est $\sigma^2(1-h_{ii})$, et cette variance est estimée par $\hat{\sigma}^2(1-h_{ii})$.

Proposition Les *composantes h_{ii} satisfont les inégalités*

$$0 \le h_{ii} \le 1$$

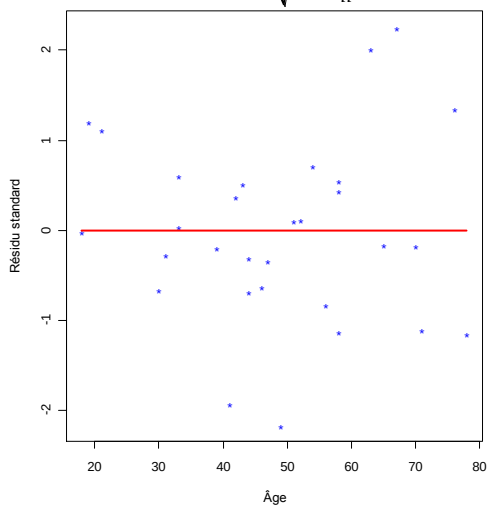
Si l'une des colonnes de **X** *est une colonne de* 1, *alors* $h_{ii} \ge 1/n$.

Démonstration

Selon le théorème 1.3.7, les valeurs d'une forme quadratique $\boldsymbol{x}'\mathbf{A}\boldsymbol{x}$ se situent entre la plus petite et la plus grande valeur propre de **A**, lorsque $\boldsymbol{x}'\boldsymbol{x} = 1$. **H** étant idempotente, ses valeurs propres sont égales à 0 ou 1. Puisque $h_{ii} = \boldsymbol{e}_i'\mathbf{H}\boldsymbol{e}_i$ on en déduit que $0 \le h_{ii} \le 1$. Par ailleurs, si l'une des colonnes de **X** est $\boldsymbol{e}$, une colonne de 1, disons $\mathbf{X} = [\boldsymbol{e} \,|\mathbf{X}_1]$, alors selon le théorème 1.5.4, $\mathbf{H} = \frac{ee'}{n} + \mathbf{X}_{1.0}(\mathbf{X}_{1.0}'\mathbf{X}_{1.0})^{-1}\mathbf{X}_{1.0}'$, où $\mathbf{X}_{1.0} = \mathbf{C}\mathbf{X}_1$, $\mathbf{C} = \mathbf{I} - \boldsymbol{e}\boldsymbol{e}'/n$. Alors chaque élément de la diagonale est $\frac{1}{n}$ + (un élément de la diagonale de $\mathbf{X}_{1.0}(\mathbf{X}_{1.0}'\mathbf{X}_{1.0})^{-1}\mathbf{X}_{1.0}'$), qui est supérieur à 0. ■

Les résidus *centrés-réduits* sont donc

$$\hat{\varepsilon}_{i(cr)} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} \tag{4.4.2}$$

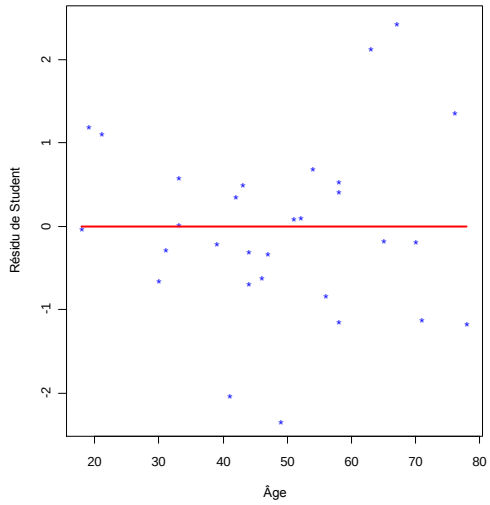


Résidus de Student

Une quantité préférée à un résidu centré-réduit est le *résidu de Student* t_i. Il est défini par

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\sqrt{\hat{\mathrm{V}}(y_i - \hat{y}_{i(i)})}} = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+h_{ii(i)}}}, \tag{4.4.3}$$

où $\hat{y}_{i(i)}$ est la prédiction de y_i à partir d'une régression faite avec toutes les données sauf la i^e et $\hat{\sigma}_{(i)}$ est l'écart-type estimé à partir de toutes les données sauf la i^e, $h_{ii(i)} = \boldsymbol{x}_i'(\mathbf{X}_{(i)}'\mathbf{X}_{(i)})^{-1}\boldsymbol{x}_i$, $\mathbf{X}_{(i)}$ étant la matrice $\mathbf{X}$ sans sa i^e ligne et $\boldsymbol{x}_i'$ est la i^e ligne de $\mathbf{X}$.



Sous les hypothèses du modèle,

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+h_{ii(i)}}} \sim t_{n\text{-}q\text{-}1} \tag{4.4.4}$$

Le numérateur $y_i - y_{i(i)}$ est de moyenne nulle et d'écart-type $\sigma^2(1+h_{ii(i)})$. Donc $\dfrac{y_i - \hat{y}_{i(i)}}{\sigma\sqrt{1+h_{ii(i)}}}$ est une normale centrée-réduite. La statistique $\dfrac{(n-1-q)\hat{\sigma}^2_{i(i)}}{\sigma^2}$ est de loi χ^2 à (n-1-q) degrés de liberté, indépendante de $\hat{y}_{i(i)}$ (dans un modèle linéaire, l'estimateur de la variance est indépendant de l'estimateur de la moyenne) et y_i est indépendante de $\hat{y}_{i(i)}$ et de $\hat{\sigma}^2_{i(i)}$. Donc $\dfrac{y_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{1+h_{ii(i)}}} = \dfrac{(y_i - \hat{y}_{i(i)})/\sigma\sqrt{1+h_{ii(i)}}}{\sqrt{\hat{\sigma}^2_{i(i)}/\sigma^2}} \sim t_{n\text{-}1\text{-}q}$.

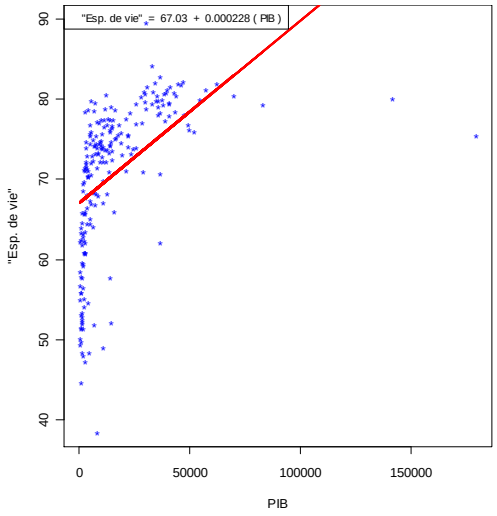
Remarque *Le calcul de t_i peut sembler très onéreux, mais en fait on peut l'obtenir à partir de données déjà calculées. On peut montrer que*

$$(n-q-1)\hat{\sigma}^2_{(i)} = (n-q)\hat{\sigma}^2 - \frac{(y_i - \hat{y}_i)^2}{1-h_{ii}} \quad ; \quad h_{ii(i)} = \frac{h_{ii}}{1-h_{ii}} \tag{4.4.5}$$

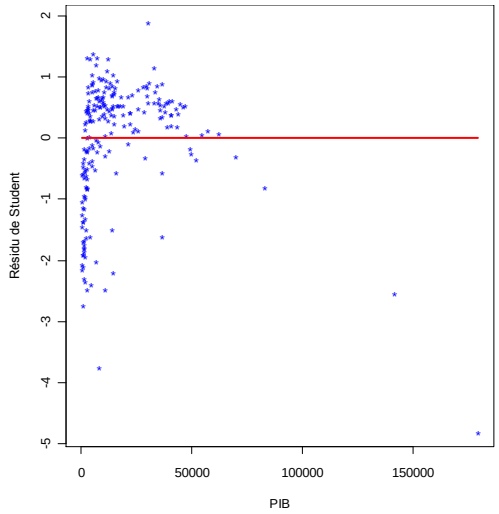
où q est le nombre de colonnes de **X**.

Exemple

Considérons la relation entre le pib d'un pays et l'espérance de vie de ses habitants pour un ensemble de 218 pays. La dépendance est très nettement non linéaire, comme le montre le nuage de points :



Le graphique des résidus de Student (en fonction du PIB) est encore plus éloquent :



Mesures d'influence

La quantité h_{ii} (le i^e élément de la diagonale de **H**) est une mesure de *l'influence* d'une observation, une mesure basée uniquement sur la valeur de *X*. On remarque que la variance Var($\hat{y}_i$) = $\sigma^2 h_{ii}$. Une observation a beaucoup d'influence si h_{ii} est importante. La moyenne des h_{ii} est q/n, où q est le nombre de colonnes de **X**, soit $q = 2$ dans une régression linéaire simple; une pratique conventionnelle consiste à signaler les observations pour lesquels h_{ii} est *trois fois* supérieure à cette moyenne, c'est-à-dire, si $h_{ii} > 6/n$.

La statistique D de Cook

Une autre mesure d'influence est la statistique *D* de Cook, définie par

$$D_i = \frac{(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})'(\hat{\mathbf{y}} - \hat{\mathbf{y}}_{(i)})}{q\hat{\sigma}^2} = \frac{(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)})}{q\hat{\sigma}^2} \qquad (4.4.6)$$

où $\hat{\mathbf{y}}_{(i)}$ est le vecteur des prédictions de tous les y_i faites avec toutes les données sauf la i^e; et $\hat{\boldsymbol{\beta}}_{(i)}$ est l'estimation de $\boldsymbol{\beta}$ faite à partir de toutes les données sauf la i^e. Dans une régression simple, la valeur de q est 2. On peut montrer que

$$D_i = \frac{\hat{\varepsilon}_i^2 h_{ii}}{q\hat{\sigma}^2(1-h_{ii}^2)} \qquad (4.4.7)$$

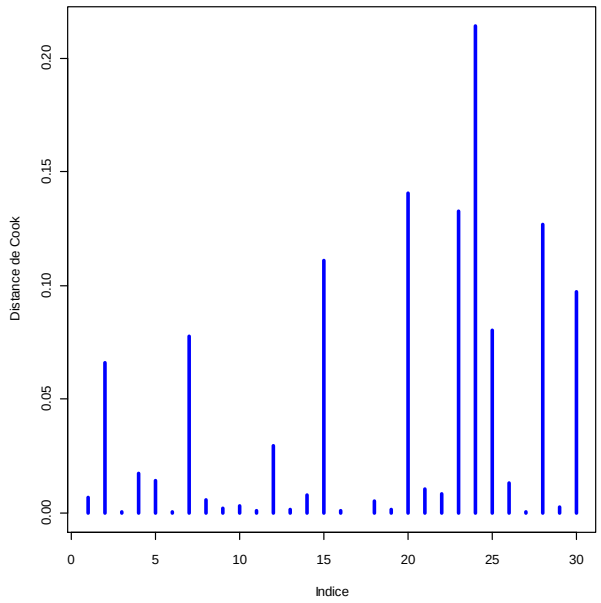
Cook suggère qu'on compare la valeur de D_i au point critique d'une statistique $\mathcal{F}_{q;n-q;\alpha}$.

Pour obtenir les distances de Cook (arrondies à 5 décimales), on fait

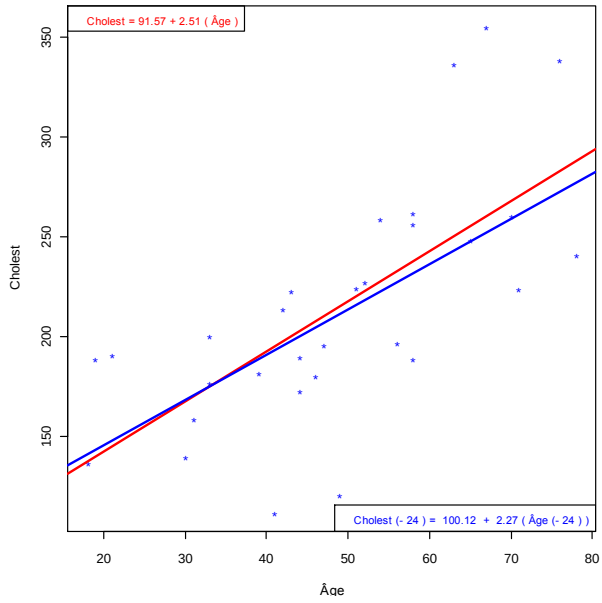
```
> round(ls.diag(reglin)$cooks,5)
      1       2       3       4       5       6       7       8       9      10
0.00646 0.06580 0.00026 0.01710 0.01409 0.00031 0.07756 0.00563 0.00174 0.00306
     11      12      13      14      15      16      17      18      19      20
0.00074 0.02950 0.00148 0.00764 0.11094 0.00072 0.00000 0.00485 0.00124 0.14065
     21      22      23      24      25      26      27      28      29      30
0.01031 0.00847 0.13256 0.21429 0.08052 0.01327 0.00012 0.12686 0.00241 0.09719
```

Voici une présentation graphique des distances de Cook :

```
> cook<-ls.diag(reglin)$cooks
> plot(cook,type="h",lwd=3,col="blue",xlab="Indice",ylab="Distance de Cook")
```



La donnée 24 est la plus influente de toutes. Pour se faire une idée de son influence, on compare la droite de régression à celle qu'on aurait obtenue en l'absence de cette donnée. Le graphique suivant montre que l'influence de cette donnée n'est somme toute pas si importante :



Les quantités DFFITS

Les quantités DFFITS_i sont une autre façon d'évaluer l'influence d'une observation. Elles sont définies par

$$\text{DFFITS}_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}\sqrt{h_{ii}}},$$

où $\hat{y}_{i(i)}$ est la prédiction de y_i faite à partir de toutes les données sauf la i^e. Le dénominateur représente une estimation de l'écart-type $\sigma_{\hat{y}_i} = \sigma\sqrt{h_{ii}}$ de $\hat{y}_i$, avec σ estimé par $\hat{\sigma}_{(i)}$.

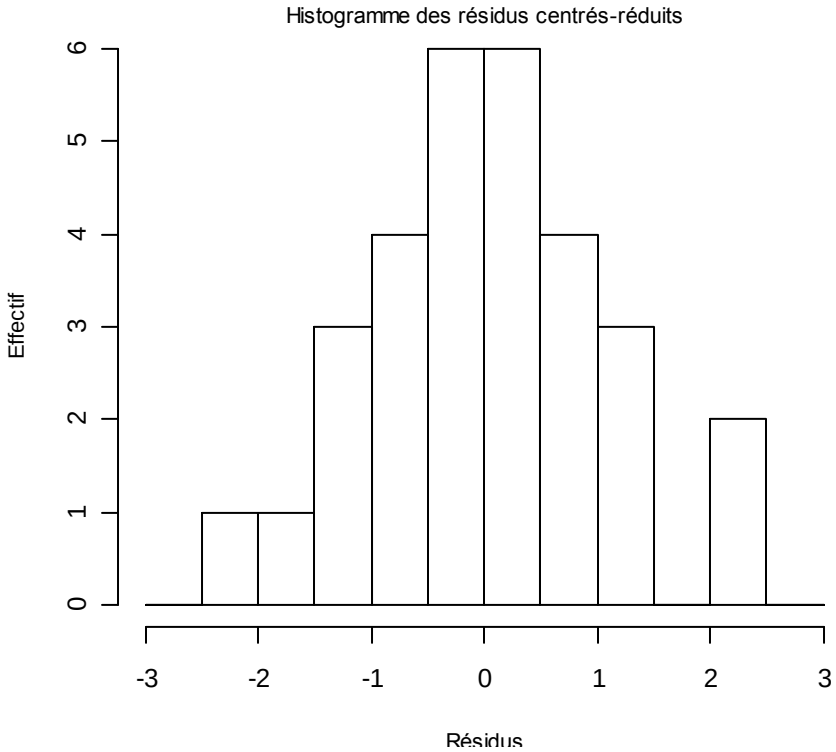
Il serait plus raisonnable de diviser l'écart $\hat{y}_i - \hat{y}_{i(i)}$ par une estimation de l'écart-type de $\hat{y}_i - \hat{y}_{i(i)}$ plutôt que par celui de $\hat{y}_i$. Si on estime σ par $\hat{\sigma}_{(i)}$, l'écart-type estimé de $\hat{y}_i - \hat{y}_{i(i)}$ est $\hat{\sigma}_{(i)}h_{ii}/\sqrt{(1-h_{ii})}$. Les statistiques modifiées deviennent alors

$$\text{DFFITS}_i^* = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\hat{\sigma}_{(i)}h_{ii}/\sqrt{(1-h_{ii})}}.$$

Normalité des résidus

L'histogramme des résidus

Une façon naturelle de vérifier si les ε_i sont de loi normale est d'examiner l'histogramme des résidus centrés-réduits $\hat{\varepsilon}_{i(cr)}$. En voici un histogramme :



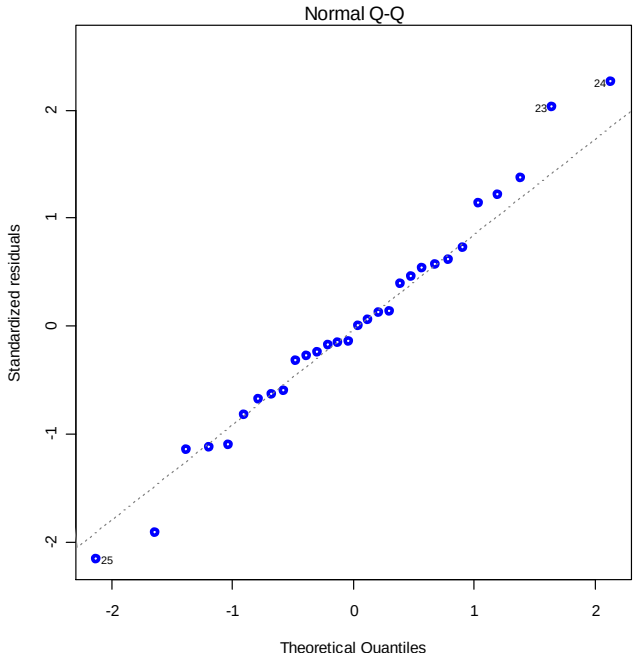
On ne peut pas vraiment conclure à la *non*-normalité des $\hat{\varepsilon}_{i(cr)}$.

Les graphiques Q-Q

Une technique graphique qui est probablement plus efficace est le graphique que nous appellerons le graphique Q-Q (quantile-quantile). Dans un échantillon aléatoire simple, $Y_1, \ldots, Y_n$, on considère les statistiques d'ordre, disons $Y_{(1)} \leq \ldots \leq Y_{(n)}$. Ces valeurs devraient être proches de leur espérance: les points $(Y_{(i)}\,;\, \mathrm{E}(Y_{(i)}))$ devraient s'aligner le long de la droite $y = x$ si l'espérance $\mathrm{E}(Y_{(i)})$ est calculée sous une hypothèse correcte. En particulier, sous l'hypothèse que la population est de loi normale, la quantité $\mathrm{E}(Y_{(i)})$ doit être l'espérance de la i^e statistique d'ordre d'un échantillon d'une population $\mathcal{N}(\mu\,;\,\sigma^2)$. Si les valeurs de μ et de σ ne sont pas connues, et ne font pas partie de l'hypothèse à tester, on ne peut pas

calculer ces espérances. Mais les statistiques $Z_{(i)} = \dfrac{Y_{(i)} - \mu}{\sigma}$ sont les statistiques d'ordre d'un échantillon d'une population $\mathcal{N}(0 ; 1)$, et l'espérance de la i^e statistique d'ordre d'une $\mathcal{N}(0 ; 1)$ peut être calculée. On l'approche par $\Phi^{-1}[(i\text{-}3/8)/(n + ¼)]$, où Φ est la fonction de répartition d'une variable de loi $\mathcal{N}(0 ; 1)$. On a donc, sous l'hypothèse d'une population normale, $\mathrm{E}\left(\dfrac{Y_{(i)} - \mu}{\sigma}\right) = \Phi^{-1}\left(\dfrac{i-3/8}{n+1/4}\right) \Rightarrow \mathrm{E}(Y_{(i)}) = \mu + \sigma \Phi^{-1}\left(\dfrac{i-3/8}{n+1/4}\right)$. Donc le nuage de points $\left(\Phi^{-1}\left(\dfrac{i-3/8}{n+1/4}\right) ; Y_{(i)} \right)$ devrait se situer proche d'une droite.

Habituellement, on considère les résidus centrés réduits $\hat{\varepsilon}_{i(cr)}$ comme un échantillon aléatoire simple (bien qu'en fait ils soient corrélés), et on trace le nuage de points $\left(t_{(i)};\ \Phi^{-1}\left(\dfrac{i-3/8}{n+1/4}\right) ; \hat{\varepsilon}_{i(cr)}\right)$. Voici le nuage pour l'exemple :



L'alignement n'étant pas très éloigné d'une droite, on n'a pas de raison de douter de la normalité des *y*.

Commandes R

Nous illustrons les commandes qui permettent de calculer les diverses mesures décrites ci-dessus avec les données du tableau 4.4.1.

Éléments de la diagonale de la matrice **H** (à 5 décimales) :

```
> round(ls.diag(reglin)$hat,5)
 [1] 0.03418 0.09827 0.03410 0.07782 0.04046 0.03485 0.04072 0.03733 0.03365 0.03890
[11] 0.04514 0.04482 0.03602 0.04482 0.14513 0.06818 0.15390 0.04482 0.09261 0.13045
[21] 0.03714 0.03602 0.06022 0.07718 0.03336 0.06460 0.06460 0.14614 0.07315 0.13140
```

Résidus standardisés (à 5 décimales) :

```
> round(ls.diag(reglin)$std.res,5)
 [1] -0.60427 -1.09890  0.12023 -0.63672 -0.81761  0.13148 -1.91153  0.53861 -0.31636
[10]  0.38917 -0.17657 -1.12145 -0.28117  0.57063 -1.14326 -0.14033  0.00436  0.45473
```

```
[19] -0.15576  1.36936  0.73125 -0.67342  2.03417  2.26370 -2.16026  0.61995  0.05780
[28]  1.21753 -0.24718  1.13355
```

Résidus de Student (à 5 décimales) :

```
> round(ls.diag(reglin)$stud.res,5)
 [1] -0.59729 -1.10315  0.11809 -0.62982 -0.81264  0.12916 -2.01302  0.53167 -0.31122
[10]  0.38320 -0.17349 -1.12684 -0.27650  0.56363 -1.14982 -0.13785  0.00428  0.44820
[19] -0.15302  1.39210  0.72503 -0.66670  2.16378  2.45932 -2.32381  0.61300  0.05676
[28]  1.22856 -0.24299  1.13958
```

Distances de Cook (arrondies à 5 décimales) :

```
> round(ls.diag(reglin)$cooks,5)
      1       2       3       4       5       6       7       8       9      10
0.00646 0.06580 0.00026 0.01710 0.01409 0.00031 0.07756 0.00563 0.00174 0.00306
     11      12      13      14      15      16      17      18      19      20
0.00074 0.02950 0.00148 0.00764 0.11094 0.00072 0.00000 0.00485 0.00124 0.14065
     21      22      23      24      25      26      27      28      29      30
0.01031 0.00847 0.13256 0.21429 0.08052 0.01327 0.00012 0.12686 0.00241 0.09719
```

Pour une représentaion graphique des distances de Cook :

```
> cook<-ls.diag(reglin)$cooks
> plot(cook,type="h",lwd=3,col="blue",xlab="Indice",ylab="Distance de Cook")
```

Les quantités DFFITS :

```
> round(ls.diag(reglin)$dfits,5)
 [1] -0.11237 -0.36418  0.02219 -0.18295 -0.16688  0.02454 -0.41475  0.10470 -0.05808
[10]  0.07709 -0.03772 -0.24408 -0.05345  0.12209 -0.47375 -0.03729  0.00183  0.09708
[19] -0.04889  0.53919  0.14240 -0.12888  0.54772  0.71123 -0.43168  0.16110  0.01492
[28]  0.50826 -0.06827  0.44323
```

Les graphiques Q-Q

La commande plot.lm(reglin) fournit plusieurs graphiques. Le deuxième est le graphique Q-Q.

```
> plot.lm(reglin,which=2,col="blue",lwd=3)
```

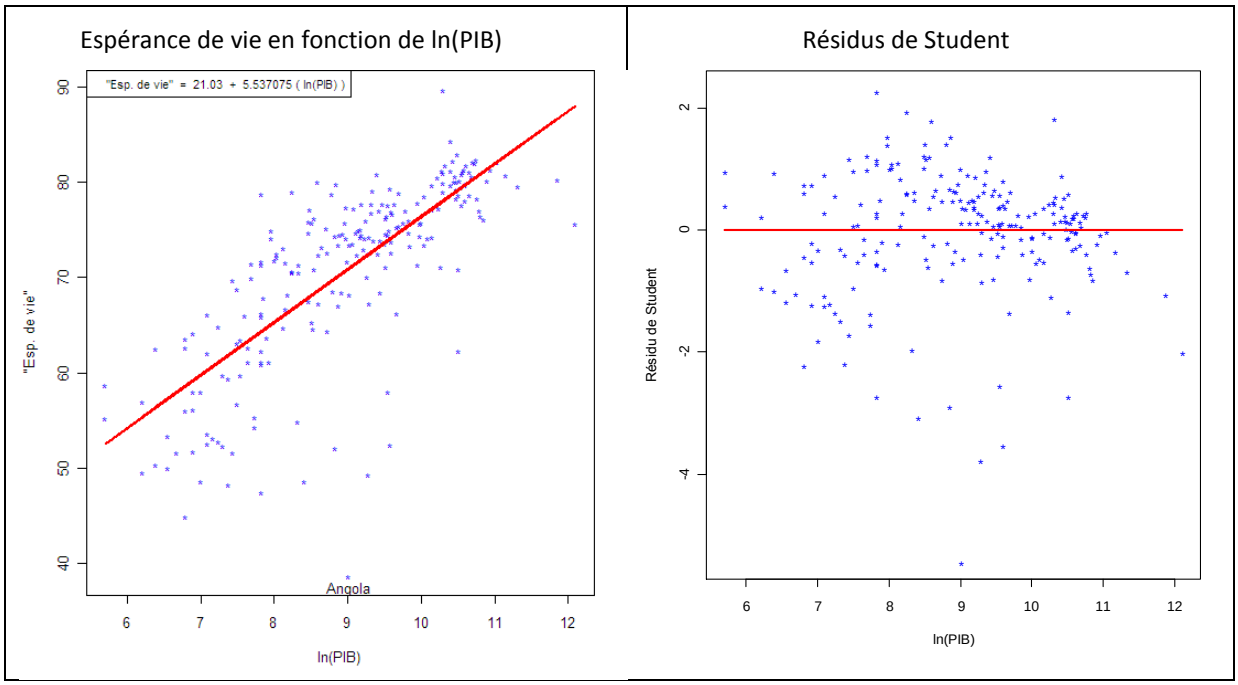
4.5 Transformations de variables

Les graphiques ou les techniques diagnostiques peuvent révéler des violations des hypothèse de la régression linéaire : hétéroscédasticité, par exemple, ou absence de normalité, ou encore non linéarité. On peut parfois pallier ces difficultés par une transformation des variables. Nous discutons ici quelques situations dans lesquelles une transformation de la variable dépendante peut avoir l'effet voulu.

Exemple *Une relation non linéaire*

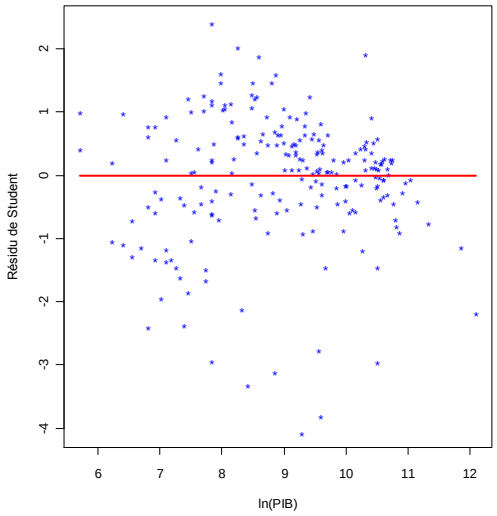
Dans l'exemple de la section 4.4, la relation entre le PIB et l'espérance de vie est fortement non linéaire.

Dans le contexte, il est raisonnable de penser que l'espérance de vie croît plus rapidement lorsque le PIB est petit. On pourrait alors considérer un taux d'accroissement *inversement proportionnel* au PIB. En général, si $\mu_y(x) = E(y|x)$, on suppose que $\frac{d\mu_y(x)}{dx} = \frac{\beta_1}{x}$, ce qui entraîne que $\mu_y(x) = \beta_o + \beta_1 \ln x$. Dans le cas présent, `espvie` = $\beta_o + \beta_1 \ln$ (`pib`). On examine donc la relation entre `espvie` et une variable construite, `lpib` = ln (`pib`). Le nuage de points suivant et le graphique des résidus de Student montrent en effet qu'une relation linéaire s'ajuste assez bien aux données :



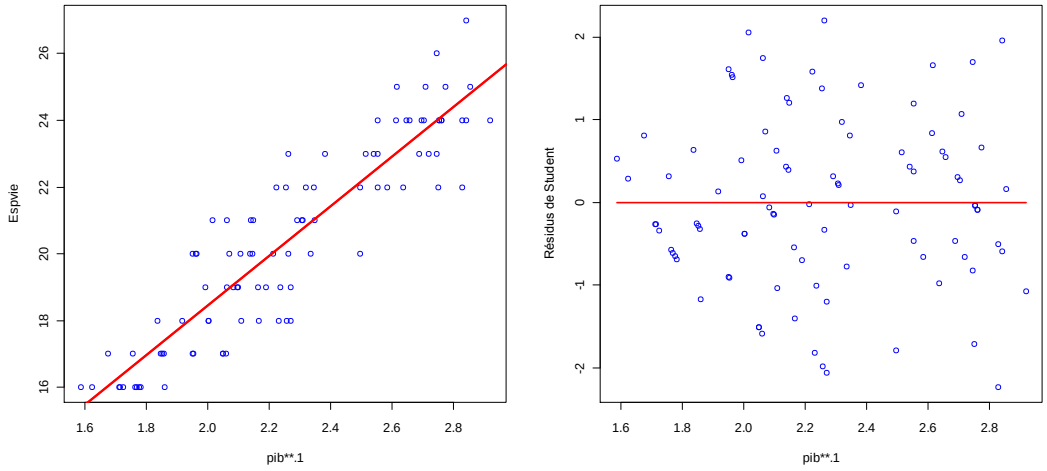
Le graphique des résidus ne contredit pas l'hypothèse de linéarité :

Si la non linéarité ne semble plus faire problème, l'hétéroscédasticité, par contre, subsiste, et, en plus, une certaine donnée, la 6^{e} (l'Angola) semble tout à fait aberrante. Si on suppose que cette observation est erronée, ou qu'elle n'appartient pas vraiment à la population visée, on peut l'omettre. On obtient alors le graphique des résidus de Student suivant :



Plusieurs pays demeurent aberrants.

Un autre modèle, adéquat mais mal motivé : $\texttt{espvie} = \beta_o + \beta_1(\texttt{pib}^{0,1})$. Illustré dans les graphiques suivants (après élimination d'une autre donnée aberrante.)



■

Le cas ou* $\mathbf{Var}(\varepsilon_i) = d_i\sigma^2$*, les* d_i *connus

Cette situation se produit lorsque chaque y_i est en fait une moyenne $\bar{y}_i$ de n_i observations. Supposons, par exemple, qu'on souhaite déterminer la relation entre la moyenne d'un élève au Cégep et sa moyenne au cours de la première session à l'université. Un grand échantillon d'élèves est prélevé, mais les élèves se trouvent groupés et seules les moyennes des groupes sont retenues : les moyennes $\bar{y}_i$ (moyenne à l'université) et $\bar{x}_i$ (moyenne au Cégep). On ne peut plus supposer que les variances des $\bar{y}_i$ sont égales à moins que les tailles des classes soient les mêmes. Le modèle $\bar{y}_i = \beta_o + \beta_1 \bar{x}_i + \varepsilon_i$, serait encore applicable, mais si les tailles des classes sont $n_1, n_2, \ldots, n_k$, alors $\mathrm{Var}(\bar{y}_i) = \mathrm{Var}(\varepsilon_i) = \sigma^2/n_i$, ce qui viole l'hypothèse d'homoscédasticité. Un modèle approprié est

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \mathrm{V}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{D}^{-1},\ \text{où}\ \mathbf{D} = \begin{bmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & n_k \end{bmatrix}$$

On peut alors transformer les données :

$$\boldsymbol{z} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\delta},\ \mathrm{V}(\boldsymbol{\delta}) = \sigma^2\mathbf{I}$$

où

$$\boldsymbol{z} = \mathbf{D}^{1/2}\boldsymbol{y},\ \mathbf{Z} = \mathbf{D}^{1/2}\mathbf{X};\ \boldsymbol{\delta} = \mathbf{D}^{1/2}\boldsymbol{\varepsilon}$$

Dans le cas présent, les estimations des paramètres sont

$$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\boldsymbol{z} = (\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{z}\ ,\ \text{et}\ \hat{\sigma}^2 = \frac{\boldsymbol{z}'(I - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}')\boldsymbol{z}}{n-2} = \frac{\boldsymbol{y}'[\mathbf{D} - \mathbf{D}^{1/2}\mathbf{X}(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}]\boldsymbol{y}}{n-1}$$

Exemple

Considérons donc un groupe réparti en 20 classes de tailles différentes pour lesquelles nous observons les moyennes $\bar{y}_i$ et $\bar{x}_i$. Voici les données, où

n_i = Nombre d'élèves dans la classe i

$\bar{y}_i$ = Moyenne des résultats à l'université

$\bar{x}_i$ = Moyenne des résultats au Cégep.

n_i	$\overline{y}_i$	$\overline{x}_i$	n_i	$\overline{y}_i$	$\overline{x}_i$
20	3,832	5,303	44	3,892	5,500
45	3,938	5,521	55	3,76	5,285
60	3,937	5,493	76	3,882	5,459
15	3,866	5,512	23	3,574	5,205
43	3,717	5,234	12	3,804	5,347
67	3,805	5,364	45	3,68	5,181
44	3,929	5,506	23	3,753	5,236
23	3,942	5,518	44	3,67	5,271
44	3,916	5,598	28	4,029	5,544
43	3,758	5,345	40	3,679	5,269

Déterminons les valeurs transformées :

```
      z                z
   17.14          4.47 23.72
   26.42          6.71 37.04
   30.50          7.75 42.55
   14.97          3.87 21.35
   24.37          6.56 34.32
   31.15          8.19 43.91
   26.06          6.63 36.52
   18.91          4.80 26.46
   25.98          6.63 37.13
   24.64          6.56 35.05
   25.82          6.63 36.48
   27.88          7.42 39.19
   33.84          8.72 47.59
   17.14          4.80 24.96
   13.18          3.46 18.52
   24.69          6.71 34.76
   18.00          4.80 25.11
   24.34          6.63 34.96
   21.32          5.29 29.34
   23.27          6.32 33.32
```

$\hat{\beta} = \begin{bmatrix} -0,5281154 \\ 0,8070843 \end{bmatrix}$; $\hat{\sigma}^2 = 0{,}07404025$ ■

Transformation de la variable dépendante : la méthode de Box-Cox

La méthode de Box-Cox est basée sur la supposition qu'il existe une transformation de la forme $u = u(y\,;\lambda) = \dfrac{y^\lambda - 1}{\lambda}$, telle que la régression de $\boldsymbol{u}$ sur $\boldsymbol{x}$ est linéaire et que u_i est normale, de moyenne $\beta_o+\beta_1 x_i$ et de variance σ^2. La transformation $u(y\,;\lambda)$ est définie pour tout $\lambda > 0$, mais on étend sa définition pour qu'elle inclue la transformation logarithmique : on ajoute donc $u(y\,;0) = \log(\boldsymbol{y})$ $(= \lim_{\lambda\to 0} \dfrac{y^\lambda - 1}{\lambda})$. Le paramètre λ est estimé par la méthode du maximum de vraisemblance.

Soit f la fonction de densité conjointe des u_i. On a alors

$$f(u_1;\ \ldots\ ;u_n) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} e^{(-1/2\sigma^2)\sum_{i=1}^n (u_i-\beta_o-\beta_1 x_i)^2}$$

La densité conjointe des observations brutes (les y_i) est donc $f[u(y_1;\lambda);\ \ldots\ ;u(y_n;\lambda)]J(\boldsymbol{y}\,;\lambda)$, où le Jacobien $J(\boldsymbol{y}\,;\lambda) = \left(\prod_{i=1}^n y_i\right)^{\lambda-1} = G^{n(\lambda-1)}$ où $G = \left(\prod_{i=1}^n y_i\right)^{1/n}$. La fonction de vraisemblance est

$L(\sigma^2 ; \beta_o ; \beta_1; \lambda) = f[u(y_1;\lambda); \ \ldots ; u(y_n;\lambda)]J(\boldsymbol{y} ; \lambda) = \dfrac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} e^{(-1/2\sigma^2)\sum_{i=1}^{n}(u(y_i;\lambda)-\beta_o-\beta_1 x_i)^2} G^{n(\lambda-1)}$.

Lorsqu'on maximise L par rapport à σ^2, β_o et β_1 pour un λ fixe, on obtient les estimateurs usuels, soit

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(u_i - \hat{\beta}_o - \hat{\beta}x_i)^2 \ , \ \hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(u_i - \bar{u})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ et } \hat{\beta}_o = \bar{u} - \hat{\beta}_1\bar{x} \ \ u_i = u(y_i;\lambda)$$

et $\ln L(\hat{\sigma}^2;\hat{\beta}_o;\hat{\beta}_1;\lambda) = \ -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\hat{\sigma}^2 - \frac{n}{2} + \frac{n}{2}(\lambda-1)\ln G^2 \ = \text{Constante} - \frac{n}{2}\ln\hat{\sigma}^2 + \frac{n}{2}(\lambda-1)\ln G^2$.

$= \text{Constante} - \frac{n}{2}\ln\left(\frac{\hat{\sigma}}{G^{\lambda-1}}\right)^2$. Or $\left(\frac{\hat{\sigma}}{G^{\lambda-1}}\right)^2$ est la somme des carrés résiduelle SCR de la régression de $\boldsymbol{z}(\lambda) = \frac{u(\lambda)}{G^{\lambda-1}}$ sur $\boldsymbol{x}$, une somme que nous devons minimiser afin de maximiser la vraisemblance. Donc la méthode consiste à calculer le vecteur $\boldsymbol{z}(\lambda)$ pour différentes valeurs de λ, déterminer la régression de $\boldsymbol{z}(\lambda)$ sur $\boldsymbol{x}$, puis calculer la somme de carrés résiduelle SCR(λ). Nous retenons la valeur de λ qui minimise SCR(λ).

Exemple

Le tableau suivant contient des données sur la concentration d'une certaine substance dans le sang en fonction du temps écoulé depuis l'injection de la substance. Les variables sont

Conc : La concentration de la substance en question dans le sang
Heures : Le nombre d'heures après l'injection

Voici les données

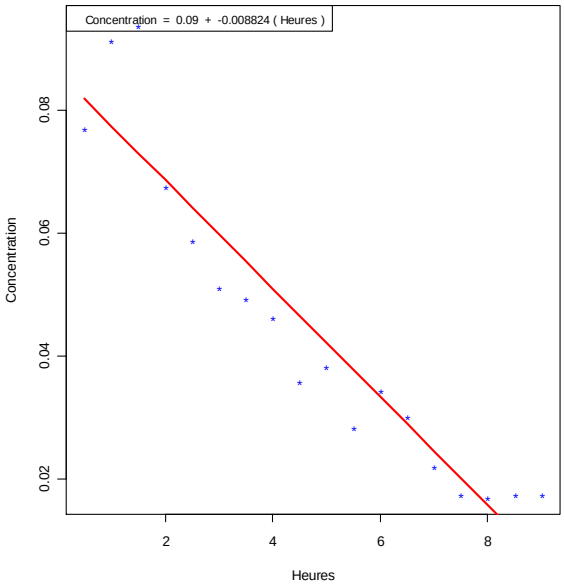
Heures	*Conc*	*Heures*	*Conc*	*Heures*	*Conc*	*Heures*	*Conc*	*Heures*	*Conc*
0,5	0,0773	2,5	0,0591	4,5	0,0360	6,5	0,0305	8,0	0,0173
1,0	0,0916	3,0	0,0513	5,0	0,0384	7,0	0,0222	8,5	0,0178
1,5	0,0941	3,5	0,0497	5,5	0,0286	7,5	0,0177	9,0	0,0177
2,0	0,0679	4,0	0,0465	6,0	0,0346				

Voici l'analyse par R :

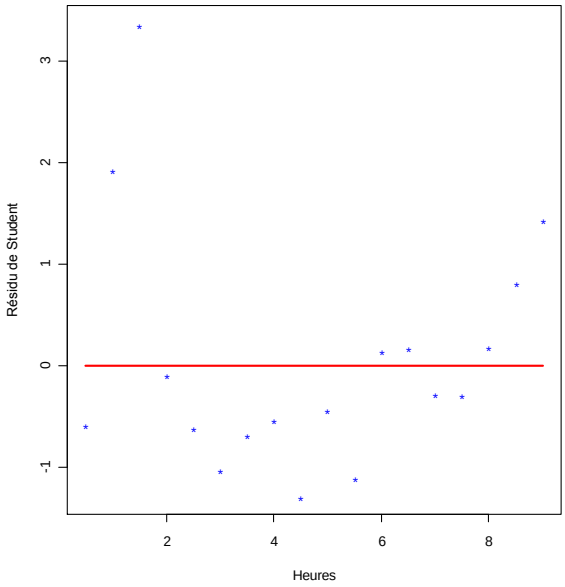
```
> summary(lm(Conc~Heures))
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.086265   0.004265   20.23 8.04e-13 ***
Heures      -0.008824   0.000788  -11.20 5.56e-09 ***

Residual standard error: 0.008672 on 16 degrees of freedom
Multiple R-Squared: 0.8869,     Adjusted R-squared: 0.8798
F-statistic: 125.4 on 1 and 16 DF,  p-value: 5.563e-09
```

L'analyse montre que le nombre d'heures est un prédicteur utile, mais là n'est pas la question. La question est de savoir si le modèle d'une droite est adéquat. Les graphiques suivants semblent indiquer que non :

Concentration en fonction du nombre d'heures



Résidus de Student en fonction du nombre d'heures

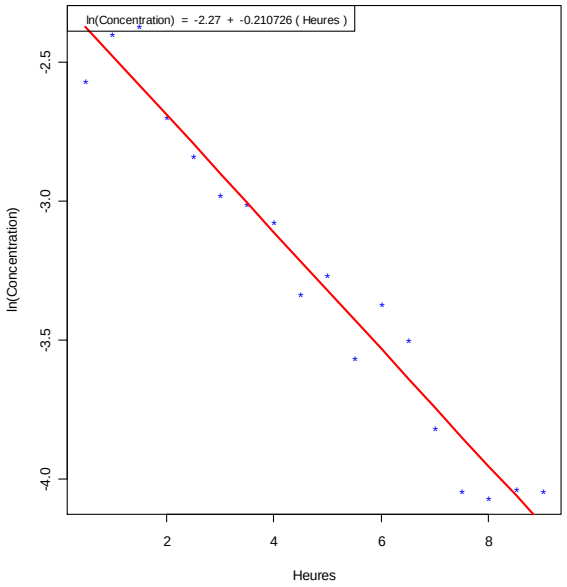


Il y a certaines raisons théoriques pour croire que c'est le *logarithme* de la concentration qui devrait décroître linéairement avec le temps. Ceci découle de l'hypothèse que le taux de décroissance de la substance dans le sang est proportionnelle à la quantité existante. Voici l'analyse faite avec le logarithme de la concentration (variable `Logconc`) comme variable dépendante :

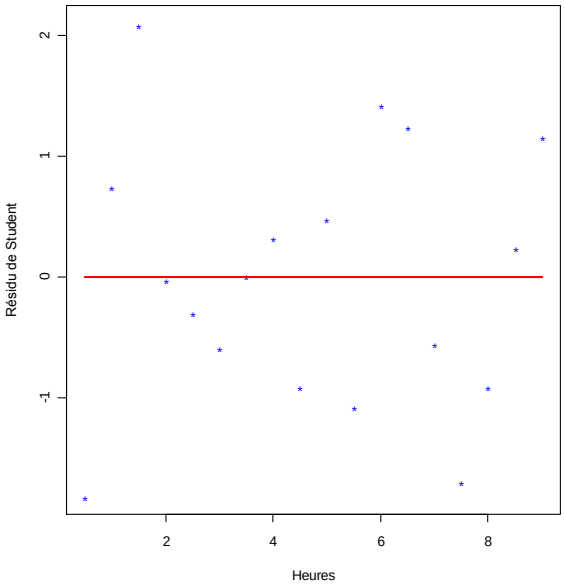
```
> summary(lm(logconc~Heures))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.26716    0.06127  -37.00  < 2e-16 ***
Heures      -0.21073    0.01132  -18.61 2.89e-12 ***

Residual standard error: 0.1246 on 16 degrees of freedom
Multiple R-Squared: 0.9559,     Adjusted R-squared: 0.9531
F-statistic: 346.5 on 1 and 16 DF,  p-value: 2.887e-12
```

Le coefficient r^2 (`R-Squared`) est passé de 91,1% à 96,3%, une amélioration faible mais sensible. Le graphique des résidus montre également que le modèle s'ajuste mieux aux données.

ln(Concentration) en fonction du nombre d'heures

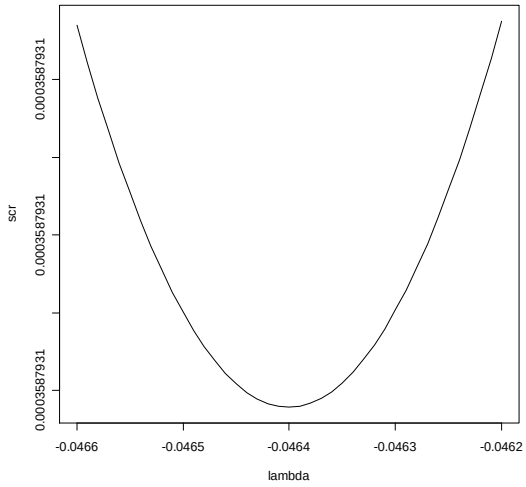


Résidus de Student en fonction du nombre d'heures



Est-ce qu'une application de la transformation de Box-Cox aurait mené à la même transformation? Vérifions.

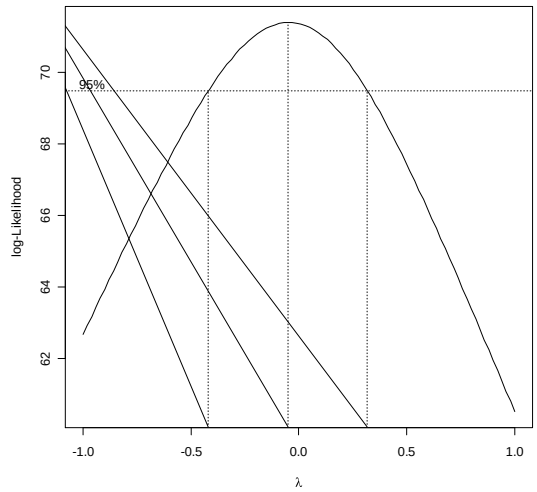
On a G = $\left(\prod_{i=1}^{n} y_i\right)^{1/n} = \left(\prod_{i=1}^{18} y_i\right)^{1/18}$ = 0,03807827. Nous calculons le vecteur $\boldsymbol{z}(\lambda) = \dfrac{y^{\lambda} - 1}{\lambda G^{\lambda - 1}}$ pour différentes valeurs de λ, puis calculons la somme de carrés résiduelle SCR(λ) correspondant à chaque valeur de λ. Voici un graphique montrant la relation entre λ et SCR(λ).



Nous devons retenir la valeur $\hat{\lambda}$ de λ qui minimise SCR(λ). On voit qu'une valeur approchée est $\hat{\lambda}$ = -0,05. Une fois la valeur $\hat{\lambda}$ déterminée, nous pouvons ne pas tenir compte des constantes dans la définition de ***z*** et définir ***z*** plus simplement comme $z = y^{\hat{\lambda}} - 1$ ou $z = y^{\hat{\lambda}}$. Il se trouve qu'avec $\hat{\lambda}$ = -0,05, la transformation est très proche de la transformation logarithmique (la corrélation entre `logconc` et ***z*** est de l'ordre de 0,999). On peut faire faire cette analyse par R. On commence par télécharger les procédures MASS, puis on donne la commande (le paramètre **`lambda=seq(-1,1,.1)`** donne l'ensemble

des valeurs de λ pour lesquelles on veut effectuer les calculs). Ce qui est donné, c'est le logarithme de la fonction de vraisemblance; donc ce que nous recherchons, c'est la valeur de λ qui *maximise* ce logarithme:

```
> boxcox(lm(y~x),lambda=seq(-1,1,.1))
```



Pour obtenir le détail numérique on fait plutôt

```
> boxcox(lm(y~x),plotit=F,lambda=seq(-.1,0,.01))
$x
 [1] -0.10 -0.09 -0.08 -0.07 -0.06 -0.05 -0.04 -0.03 -0.02 -0.01  0.00
$y
 [1] 71.351 71.366 71.377 71.386 71.392 71.395 71.394 71.391 71.384 71.374 71.361
```

RÉSUMÉ

1 Modèle de régression linéaire simple : y_i est une variable aléatoire de moyenne $\beta_0 + \beta_1 x_i$ et de variance $\sigma^2_{y.x}$.

2 Les estimateurs de β_1 et β_0 sont $\hat{\beta}_1 = \dfrac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sum(x_i-\bar{x})^2} = \dfrac{s_{xy}}{s_x^2}$ et $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$.

3 $\hat{\beta}_0 \sim \mathcal{N}(\beta_0 ; \sigma^2_{\hat{\beta}_o})$ et $\hat{\beta}_1 \sim \mathcal{N}(\beta_1 ; \sigma^2_{\hat{\beta}_1})$ où $\sigma^2_{\hat{\beta}_\circ} = \sigma^2\left(\dfrac{1}{n} + \dfrac{\sigma^2}{\sum(x_i-\bar{x})^2}\right)$ et $\sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2}{\sum(x_i-\bar{x})^2}$.

4 On estime $\sigma^2_{\hat{\beta}_o}$ et $\sigma^2_{\hat{\beta}_1}$ en remplaçant σ^2 par son estimateur, qui est

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^n \left(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)\right)^2}{n-2}.$$

et alors

$$\frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma}_{\hat{\beta}_0}} \sim t_{n-2} \text{ et } \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\hat{\beta}_1}} \sim t_{n-2}$$

5 Les intervalles de confiance pour β_0 et β_1 sont donnés par

$$\hat{\beta}_0 - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_\circ} \le \beta_0 \le \hat{\beta}_0 + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_\circ} \text{ et } \hat{\beta}_1 - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1} \le \beta_1 \le \hat{\beta}_1 + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\beta}_1}$$

6 Un intervalle de confiance pour $\mu_X = \beta_0 + \beta_1 x$ est donné par

$$\hat{\mu}_x - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_x} \le \mu_x \le \hat{\mu}_x + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_x} \quad \hat{\sigma}^2_{\hat{\mu}_x} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].$$

7 Limites de prédiction:

$$\hat{y}_x - t_{n-2;\alpha/2}\hat{\sigma}_{y-\hat{y}_x} \le y_x \le \hat{y}_x + t_{n-2;\alpha/2}\hat{\sigma}_{y-\hat{y}_x}, \text{ où } \hat{\sigma}_{y-\hat{y}_x} = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}.$$

La région critique pour tester les hypothèses H_o: $\beta_1 = b$ et H_o : $\beta_0 = a$, respectivement, sont

$$\left|\frac{\hat{\beta}_1 - b}{\hat{\sigma}_{\hat{\beta}_1}}\right| > t_{n-2;\alpha/2} \text{ et } \left|\frac{\hat{\beta}_0 - a}{\hat{\sigma}_{\hat{\beta}_0}}\right| > t_{n-2;\alpha/2}$$

8 La statistique $\hat{\mu}_x = \hat{\beta}_o + \hat{\beta}_1 x \sim \mathcal{N}(\mu_x ; \sigma^2_{\hat{\mu}_x})$ où $\sigma^2_{\hat{\mu}_x} = \sigma^2\left[\dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right]$. La variance $\sigma^2_{\hat{\mu}_x}$ est estimée par

$$\hat{\sigma}^2_{\hat{\mu}_x} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right].$$

9 $\dfrac{\hat{\mu}_x - \mu_x}{\hat{\sigma}_{\hat{\mu}_x}} \sim t_{n-2}$, ce qui donne l'intervalle de confiance $\hat{\mu}_x - t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_x} \le \mu_x \le \hat{\mu}_x + t_{n-2;\alpha/2}\hat{\sigma}_{\hat{\mu}_x}$

10 Si y_x est une nouvelle observations correspondant à une valeur données x, alors $\dfrac{y - \hat{y}_x}{\hat{\sigma}_{\hat{y}_x}} \sim t_{n-2}$,

où $\hat{\sigma}_{y-\hat{y}_x} = \sqrt{\hat{\sigma}^2 + \hat{\sigma}^2_{\hat{\mu}_x}} = \hat{\sigma}\sqrt{1 + \dfrac{1}{n} + \dfrac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}}$.

Nous obtenons alors les limites de prédiction $\hat{y}_x - t_{n-2;\alpha/2}\hat{\sigma}_{y-\hat{y}_x} \le y_x \le \hat{y}_x + t_{n-2;\alpha/2}\hat{\sigma}_{y-\hat{y}_x}$

11 Décomposition de la somme des carrés totale (SCT):

$$\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2 = \text{SCE} + \text{SCR}.$$

Table d'analyse de variance

Source	*Somme des carrés*	*d.ℓ.*	*Moyenne des carrés*	*Espérances*
Régression	SCE = $\sum(\hat{y}_i - \bar{y})^2$	1	MCE = SCE/1	E(MCE)= $\sigma^2 + \beta_1^2 \sum_{i=1}^{n}(x_i - \bar{x})^2$
Résiduelle	SCR = $\sum(y_i - \hat{y}_i)^2$	n - 2	MCR = SCR/(n-2) = $\hat{\sigma}^2$	E(MCR) = σ^2
Total	SCT = $\sum(y_i - \bar{y})^2$	n - 1	MCT = SCT/(n-1) = s_y^2	E(MCT)= $\sigma^2 + \beta_1^2 \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$

Annexe

Tableau 4.A.1
Espérance de vie, PIB et mortalité infantile de plusieurs pays

	Pays	PIB	EspVie	MortInf	Décès		Pays	PIB	EspVie	MortInf	Décès
1	Afghanistan	900	45.02	149.20	17.39	110	Lebanon	14400	75.01	15.85	6.54
2	Albania	8000	77.41	14.61	6.15	111	Lesotho	1700	51.63	55.04	15.19
3	Algeria	7300	74.50	25.81	4.69	112	Liberia	500	57.00	74.52	10.62
4	American_Samoa	8000	74.21	9.66	4.10	113	Libya	14000	77.65	20.09	3.40
5	Andorra	46700	82.43	3.80	6.35	114	Liechtenstein	141100	80.31	4.15	7.61
6	Angola	8200	38.76	175.90	23.40	115	Lithuania	16000	75.34	6.27	11.33
7	Anguilla	12200	80.87	3.47	4.44	116	Luxembourg	82600	79.61	4.44	8.48
8	Antigua_and_Barbuda	16400	75.48	14.63	5.73	117	Macau	33000	84.41	3.18	3.72
9	Argentina	14700	76.95	10.81	7.38	118	Macedonia	9700	75.14	8.54	8.91
10	Armenia	5700	73.23	18.85	8.46	119	Madagascar	900	63.63	51.45	7.79
11	Aruba	21800	75.72	12.92	7.84	120	Malawi	800	51.70	81.04	13.22
12	Australia	41000	81.81	4.61	6.88	121	Malaysia	14700	73.79	15.02	4.93
13	Austria	40400	79.78	4.32	10.14	122	Maldives	6900	74.45	27.45	3.71
14	Azerbaijan	10900	67.36	51.08	8.25	123	Mali	1200	52.61	111.35	14.29
15	Bahamas_The	28700	71.18	13.49	6.88	124	Malta	25600	79.72	3.69	8.60
16	Bahrain	40300	78.15	10.43	2.61	125	Marshall_Islands	2500	71.76	23.74	4.38
17	Bangladesh	1700	69.75	50.73	5.75	126	Mauritania	2100	61.14	60.42	8.83
18	Barbados	21800	74.34	11.86	8.39	127	Mauritius	14000	74.48	11.52	6.68
19	Belarus	13600	71.20	6.25	13.77	128	Mexico	13900	76.47	17.29	4.86
20	Belgium	37800	79.51	4.33	10.57	129	Micronesia	2200	71.52	24.34	4.35
21	Belize	8400	68.23	21.95	5.87	130	Moldova	2500	71.37	12.43	10.74
22	Benin	1500	59.84	61.56	9.00	131	Monaco	30000	89.73	1.79	8.28
23	Bermuda	69900	80.71	2.47	7.57	132	Mongolia	3600	68.31	37.26	6.04
24	Bhutan	5500	67.30	44.48	7.12	133	Montserrat	3400	73.16	15.23	7.20
25	Bolivia	4800	67.57	42.16	6.85	134	Morocco	4800	75.90	27.53	4.75
26	Bosnia_and_Herzegovina	6600	78.81	8.67	8.80	135	Mozambique	1000	51.78	78.95	13.00
27	Botswana	14000	58.05	11.14	10.57	136	Namibia	6900	52.19	45.59	12.95
28	Brazil	10800	72.53	21.17	6.36	137	Nauru	5000	65.35	8.66	6.11
29	British_Virgin_Islands	38500	77.63	13.63	4.49	138	Nepal	1200	66.16	44.54	6.81
30	Brunei	51600	76.17	11.51	3.35	139	Netherlands	40300	79.68	4.59	8.85
31	Bulgaria	13500	73.59	16.68	14.32	140	New_Caledonia	15000	76.75	5.71	5.28
32	Burkina_Faso	1200	53.70	81.40	12.74	141	New_Zealand	27700	80.59	4.78	7.15
33	Burma	1400	64.88	49.23	8.16	142	Nicaragua	3000	71.90	22.64	5.03
34	Burundi	300	58.78	61.82	9.61	143	Niger	700	53.40	112.22	14.11
35	Cambodia	2100	62.67	55.49	8.07	144	Nigeria	2500	47.56	91.54	16.06
36	Cameroon	2300	54.39	60.91	11.83	145	Northern_Mariana_Islands	12500	77.08	5.79	3.28
37	Canada	39400	81.38	4.92	7.98	146	Norway	54600	80.20	3.52	9.24
38	Cape_Verde	3800	70.70	26.94	6.34	147	Oman	25600	74.22	15.47	3.45
39	Cayman_Islands	43800	80.68	6.63	5.10	148	Pakistan	2500	65.99	63.26	6.92
40	Central_African_Republic	700	50.07	99.38	15.01	149	Palau	8100	71.78	12.43	7.87
41	Chad	1600	48.33	95.31	15.47	150	Panama	13000	77.79	11.64	4.65
42	Chile	15400	77.70	7.34	5.97	151	Papua_New_Guinea	2500	66.24	43.29	6.58
43	China	7600	74.68	16.06	7.03	152	Paraguay	5200	76.19	23.02	4.57

Tableau 4.A1 (suite)

	Pays	PIB	EspVie	MortInf	Décès		Pays	PIB	EspVie	MortInf	Décès
44	Colombia	9800	74.55	16.39	5.26	153	Peru	9200	72.47	22.18	5.93
45	Comoros	1000	64.20	62.63	7.23	154	Philippines	3500	71.66	19.34	5.02
46	Congo_Democratic_Republic_of_the	300	55.33	78.43	11.06	155	Poland	18800	76.05	6.54	10.17
47	Congo_Republic_of_the	4100	54.91	76.05	11.49	156	Portugal	23000	78.54	4.66	10.80
48	Cook_Islands	9100	74.70	15.81	7.37	157	Puerto_Rico	16300	78.92	8.07	7.95
49	Costa_Rica	11300	77.72	9.45	4.33	158	Qatar	179000	75.70	12.05	2.43
50	Cote_d'Ivoire	1800	56.78	64.78	10.16	159	Romania	11600	73.98	11.02	11.81
51	Croatia	17400	75.79	6.16	11.91	160	Russia	15900	66.29	10.08	16.04
52	Cuba	9900	77.70	4.90	7.47	161	Rwanda	1100	58.02	64.04	9.88
53	Cyprus	21000	77.82	9.38	6.45	162	St_Helena_Ascen._&_Tristan_da_Cunha	2500	78.76	16.38	6.88
54	Czech_Republic	25600	77.19	3.73	10.86	163	Saint_Kitts_and_Nevis	13700	74.60	9.66	7.10
55	Denmark	36600	78.63	4.24	10.19	164	Saint_Lucia	11200	76.84	12.72	7.00
56	Djibouti	2800	61.14	54.94	8.23	165	Saint_Pierre_and_Miquelon	7000	79.87	7.47	8.83
57	Dominica	10400	75.98	12.78	8.06	166	St_Vincent_&_Grenadines	10300	74.15	14.27	6.98
58	Dominican_Republic	8900	77.31	22.22	4.35	167	Samoa	5500	72.40	22.74	5.34
59	Ecuador	7800	75.73	19.65	5.00	168	San_Marino	36200	83.01	4.72	7.89
60	Egypt	6200	72.66	25.20	4.82	169	Sao_Tome_&_Principe	1800	63.11	53.21	8.18
61	El_Salvador	7200	73.44	20.30	5.62	170	Saudi_Arabia	24200	74.11	16.16	3.33
62	Equatorial_Guinea	36600	62.37	77.30	9.03	171	Senegal	1900	59.78	56.42	9.26
63	Eritrea	600	62.52	41.33	8.08	172	Serbia	10900	74.32	6.52	13.85
64	Estonia	19100	73.33	7.06	13.55	173	Seychelles	23200	73.52	11.66	6.91
65	Ethiopia	1000	56.19	77.12	11.04	174	Sierra_Leone	900	56.13	78.38	11.73
66	Faroe_Islands	32900	79.72	6.06	8.67	175	Singapore	62100	82.14	2.32	4.95
67	Finland	35400	79.27	3.43	10.24	176	Slovakia	22000	75.83	6.59	9.60
68	France	33100	81.19	3.29	8.76	177	Slovenia	28200	77.30	4.17	10.87
69	French_Polynesia	18000	77.10	7.27	4.87	178	Solomon_Islands	2900	74.18	17.82	3.93
70	Gabon	14500	52.49	49.95	13.00	179	Somalia	600	50.40	105.56	14.87
71	Gambia_The	1900	63.51	71.67	7.65	180	South_Africa	10700	49.33	43.20	17.09
72	Georgia	4900	77.12	15.17	9.92	181	Spain	29400	81.17	3.39	8.80
73	Germany	35700	80.07	3.54	10.92	182	Sri_Lanka	5000	75.73	9.70	5.92
74	Ghana	2500	61.00	48.55	8.75	183	Sudan	2300	55.42	68.07	11.00
75	Gibraltar	43000	78.68	6.69	8.18	184	Suriname	9700	74.22	17.61	5.54
76	Greece	29600	79.92	5.00	10.70	185	Swaziland	4500	48.66	63.09	14.60
77	Greenland	36500	70.96	10.05	8.12	186	Sweden	39100	81.07	2.74	10.20
78	Grenada	10200	73.04	11.43	7.94	187	Switzerland	42600	81.07	4.08	8.72
79	Guatemala	5200	70.88	26.02	4.98	188	Syria	4800	74.69	15.62	3.68
80	Guernsey	44600	82.16	3.55	8.44	189	Taiwan	35700	78.32	5.18	7.00
81	Guinea	1000	58.11	61.03	10.45	190	Tajikistan	2000	66.03	38.54	6.60
82	Guinea-Bissau	1100	48.70	96.23	15.27	191	Tanzania	1400	52.85	66.93	12.09
83	Guyana	7200	67.08	36.76	7.20	192	Thailand	8700	73.60	16.39	7.29
84	Haiti	1200	62.17	54.02	8.21	193	Timor-Leste	2600	67.95	38.01	5.89
85	Honduras	4200	70.61	20.44	5.02	194	Togo	900	62.71	51.48	7.96
86	Hong_Kong	45900	82.04	2.90	7.07	195	Tonga	6100	75.16	13.65	4.90
87	Hungary	18800	74.79	5.31	12.68	196	Trinidad_and_Tobago	21200	71.37	27.69	8.29

Tableau 4.A1 (suite)

	Pays	PIB	EspVie	MortInf	Décès		Pays	PIB	EspVie	MortInf	Décès
88	Iceland	38300	80.90	3.20	6.96	197	Tunisia	9400	75.01	25.92	5.83
89	India	3500	66.80	47.57	7.48	198	Turkey	12300	72.50	23.94	6.10
90	Indonesia	4200	71.33	27.95	6.26	199	Turkmenistan	7500	68.52	42.34	6.24
91	Iran	10600	70.06	42.26	5.94	200	Turks_&_Caicos_Islands	11500	79.11	11.97	2.99
92	Iraq	3800	70.55	41.68	4.82	201	Tuvalu	3400	64.75	34.52	9.20
93	Ireland	37300	80.19	3.85	6.34	202	Uganda	1300	53.24	62.47	11.71
94	Isle_of_Man	35000	80.64	4.32	9.92	203	Ukraine	6700	68.58	8.54	15.74
95	Israel	29800	80.96	4.12	5.47	204	United_Arab_Emirates	49600	76.51	11.94	2.06
96	Italy	30500	81.77	3.38	9.84	205	United_Kingdom	34800	80.05	4.62	9.33
97	Jamaica	8300	73.45	14.60	6.54	206	United_States	47200	78.37	6.06	8.38
98	Japan	34000	82.25	2.78	10.09	207	Uruguay	13700	76.21	9.69	9.58
99	Jersey	57000	81.38	3.98	7.52	208	Uzbekistan	3100	72.51	21.92	5.29
100	Jordan	5400	80.05	16.42	2.69	209	Vanuatu	5100	64.70	46.85	7.43
101	Kazakhstan	12700	68.51	24.15	9.38	210	Venezuela	12700	73.93	20.62	5.17
102	Kenya	1600	59.48	52.29	8.93	211	Vietnam	3100	72.18	20.90	5.96
103	Kiribati	6200	64.39	38.89	7.40	212	Virgin_Islands	14500	79.33	7.24	7.17
104	Korea_North	1800	68.89	27.11	9.08	213	Wallis_and_Futuna	3800	78.98	4.67	4.68
105	Korea_South	30000	79.05	4.16	6.26	214	West_Bank	2900	75.01	14.92	3.58
106	Kuwait	48900	77.09	8.07	2.11	215	Western_Sahara	2500	61.13	60.44	8.96
107	Kyrgyzstan	2200	70.04	29.27	6.79	216	Yemen	2700	63.74	55.11	7.02
108	Laos	2500	62.39	59.46	8.13	217	Zambia	1500	52.36	66.60	12.61
109	Latvia	14700	72.68	8.42	13.60	218	Zimbabwe	500	49.64	29.50	13.58