# EXEMPLE QUI ILLUSTRE LA RÉGRESSION MULTIPLE

FRÉDÉRIC BERTRAND ET MYRIAM MAUMY

*Je vais traiter cet exemple sans me servir de R ou de Minitab et faire tous les calculs "à la main" pour vous montrer au moins une fois dans ce cours comment on applique les formules mathématiques qui sont données dans ce cours.*

Les données présentées dans le tableau ci-dessous concernent 9 entreprises de l'industrie chimique. On cherche à établir une relation entre la production $y_i$, les heures de travail $x_{i1}$ et le capital utilisé $x_{i2}$.

On fait donc l'hypothèse d'un modèle de régression multiple avec 2 variables explicatives, c'est-à-dire en notation vectorielle :

$$\vec{y} = \beta_0 \vec{1} + \beta_1 \vec{x}_1 + \beta_2 \vec{x}_2 + \vec{\varepsilon}$$

ou encore en notation matricielle :

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

où

$$\mathbf{y} = \begin{bmatrix} 60 \\ 120 \\ 190 \\ 250 \\ 300 \\ 360 \\ 380 \\ 430 \\ 440 \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & 1\,100 & 300 \\ 1 & 1\,200 & 400 \\ 1 & 1\,430 & 420 \\ 1 & 1\,500 & 400 \\ 1 & 1\,520 & 510 \\ 1 & 1\,620 & 590 \\ 1 & 1\,800 & 600 \\ 1 & 1\,820 & 630 \\ 1 & 1\,800 & 610 \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \\ \varepsilon_7 \\ \varepsilon_8 \\ \varepsilon_9 \end{bmatrix}.$$

**Tableau - Production, travail et capital**

| Entreprise<br>i | Travail (heures)<br>$x_{i1}$ | Capital (machines/heures)<br>$x_{i2}$ | Production (100 tonnes)<br>$y_i$ |
|---|---|---|---|
| 1 | 1 100 | 300 | 60 |
| 2 | 1 200 | 400 | 120 |
| 3 | 1 430 | 420 | 190 |
| 4 | 1 500 | 400 | 250 |
| 5 | 1 520 | 510 | 300 |
| 6 | 1 620 | 590 | 360 |
| 7 | 1 800 | 600 | 380 |
| 8 | 1 820 | 630 | 430 |
| 9 | 1 800 | 610 | 440 |

Il s'agit de calculer le vecteur des estimateurs $\hat{\beta}$ défini par l'égalité suivante :

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}.$$

Pour cela, on calcule :

$$(\mathbf{X}'\mathbf{X}) = \begin{bmatrix} 9 & 13\,790 & 4\,460 \\ 13\,790 & 21\,672\,100 & 7\,066\,200 \\ 4\,460 & 7\,066\,200 & 2\,323\,600 \end{bmatrix}$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 6,304\,777 & -0,007\,800 & 0.011\,620 \\ -0,007\,800 & 0.000\,015 & -0.000\,031 \\ 0.011\,620 & -0.000\,031 & 0.000\,072 \end{bmatrix}$$

et :

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 2\,530 \\ 4\,154\,500 \\ 1\,378\,500 \end{bmatrix}.$$

On obtient ainsi :

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} -437,71 \\ 0,336 \\ 0,41 \end{bmatrix}.$$

L'équation de l'hyperplan des moindres carrés est donc donnée par :

$$\hat{y}(x_1, x_2) = -437.71 + 0.336\, x_1 + 0.41\, x_2$$

On peut également calculer :

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n-p} = \frac{3194}{6} = 532.$$

On peut alors calculer :

$$\begin{aligned} s^2\left(\hat{\beta}\right) &= s^2(\mathbf{X}'\mathbf{X})^{-1} = 532 \begin{bmatrix} 6,304\,777 & -0,007\,800 & 0.011\,620 \\ -0,007\,800 & 0.000\,015 & -0.000\,031 \\ 0.011\,620 & -0.000\,031 & 0.000\,072 \end{bmatrix} \\ &= \begin{bmatrix} 3\,355.56 & -4,152 & 6,184 \\ -4,152 & 0,008 & -0,016 \\ 6,184 & -0,016 & 0,038 \end{bmatrix} \end{aligned}$$

Les écart-types $s\left(\hat{\beta}_j\right)$ des estimateurs $\hat{\beta}_j$ sont alors donnés par les racines carrées des éléments diagonaux de cette matrice. On a ainsi :

$$\begin{aligned} s\left(\hat{\beta}_0\right) &= 57,93 \\ s\left(\hat{\beta}_1\right) &= 0,089\,66 \\ s\left(\hat{\beta}_2\right) &= 0,196\,1. \end{aligned}$$

**On va maintenant réaliser des tests**.
Il faut donc s'intéresser à la normalité des résidus afin de savoir si les décisions que nous allons prendre sont légitimes ou non.

On obtient à l'aide de Minitab :

Test de normalité de Anderson-Darling
A-Carré : 0,324
Valeur de P : 0,449
**On ne peut donc pas rejeter l'hypothèse nulle de normalité au seuil** $\alpha = 5\%$**.**

Afin de tester l'hypothèse nulle :

$$H_0 : \beta_j = 0$$

contre l'hypothèse alternative :

$$H_1 : \beta_j \neq 0,$$

il s'agit de calculer les statistiques suivantes :

$$\begin{aligned} \mathrm{t}_c &= \frac{-437.71}{57.93} = -7.56 \\ \mathrm{t}_c &= \frac{0.336}{0.089\,66} = 3.75 \\ \mathrm{t}_c &= \frac{0.41}{0.196\,1} = 2.09 \end{aligned}$$

pour respectivement j=0, j= 1 et j= 2. Comme la valeur critique est donnée par $\mathrm{t}_{0.025;6} = 2.45$, on rejette l'hypothèse nulle au seuil de signification $\alpha = 0.05$ pour j= 0 et j= 1, mais on accepte l'hypothèse nulle pour j= 2.
**Conclusion : cela veut dire que la variable** $X_2$ **n'est pas significative dans le modèle.**

On calcule **les intervalles de confiance au niveau** 0.95 **pour les 3 variables** $\beta_1, \beta_2, \beta_3$.

$$\begin{aligned} -437.71 \pm 2.45 \times 57.93 &= [-579.64; -295.78] \\ 0.336 \pm 2.45 \times 0.089\,66 &= [0.116; 0.556] \\ 0.41 \pm 2.45 \times 0.196\,1 &= [-0.07; 0.89] \end{aligned}$$

**Remarque :** la valeur 0 est comprise dans l'intervalle de confiance pour $\beta_2$.

Calculons maintenant **le tableau d'ANOVA** pour notre exemple. Il s'agit de calculer les quantités suivantes :

$$\begin{aligned} \mathrm{SC}_{reg} &= \hat{\beta}\mathbf{X}'\mathbf{y} - n\overline{y}^2 \\ &= \begin{bmatrix} -437.71 & 0.336 & 0.41 \end{bmatrix} \times \begin{bmatrix} 2\,530 \\ 4\,154\,500 \\ 1\,378\,500 \end{bmatrix} - 428\,152.14 \\ &= 144\,695 \end{aligned}$$

$$\begin{aligned}
\mathrm{SC}_{tot} &= \mathbf{y}'\mathbf{y} - n\overline{y}^2 \\
&= \begin{bmatrix} 60 & 120 & 190 & \cdots & 440 \end{bmatrix} \times \begin{bmatrix} 60 \\ 120 \\ 190 \\ . \\ . \\ . \\ 440 \end{bmatrix} - 428\,152.14 \\
&= 147\,889
\end{aligned}$$

On a :

$$\mathrm{SC}_{res} = \mathrm{SC}_{tot} - \mathrm{SC}_{reg} = 147\,889 - 144\,695 = 3\,194.$$

On obtient le tableau d'ANOVA donné par le tableau ci-dessous. On peut tester l'hypothèse :

$$\mathrm{H}_0 : \beta_1 = \beta_2 = 0.$$

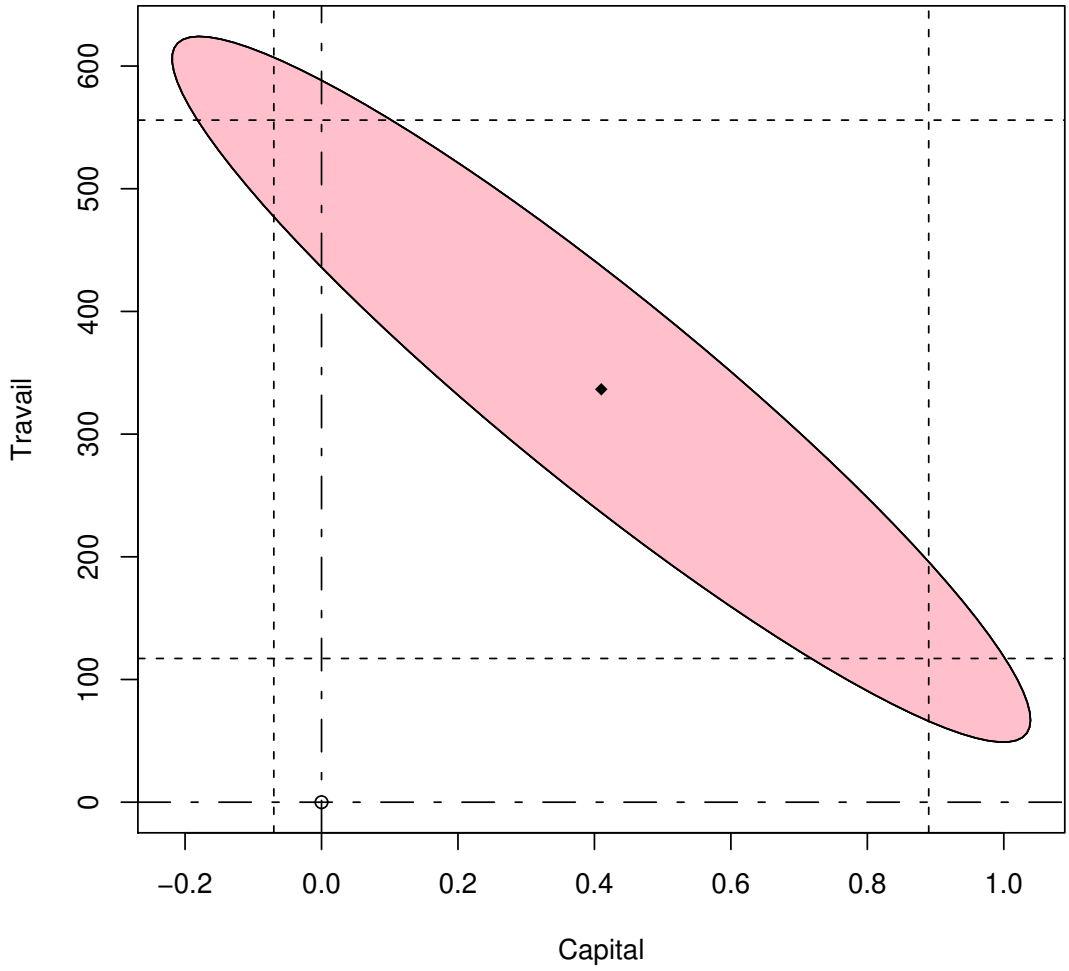
Comme la statistique $\mathrm{F}_c = 135.92$ est supérieure à la valeur critique $\mathrm{F}_{(0.05;\,2;\,6)} = 5.14$, on rejette cette hypothèse au seuil de significativité $\alpha = 0.05$.

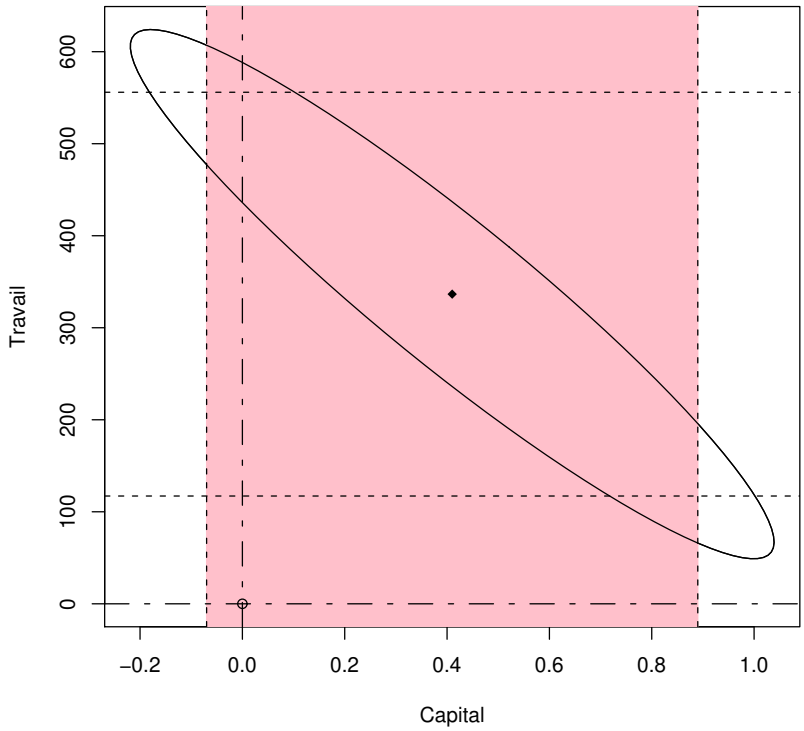
| Source de variation | Somme des carrés | ddl | Carrés moyens | $\mathrm{F}_c$ |
|---|---|---|---|---|
| Régression | 144 695 | 2 | 72 348 | 135.92 |
| Résiduelle | 3 194 | 6 | 532 | |
| Totale | 147 889 | 8 | | |

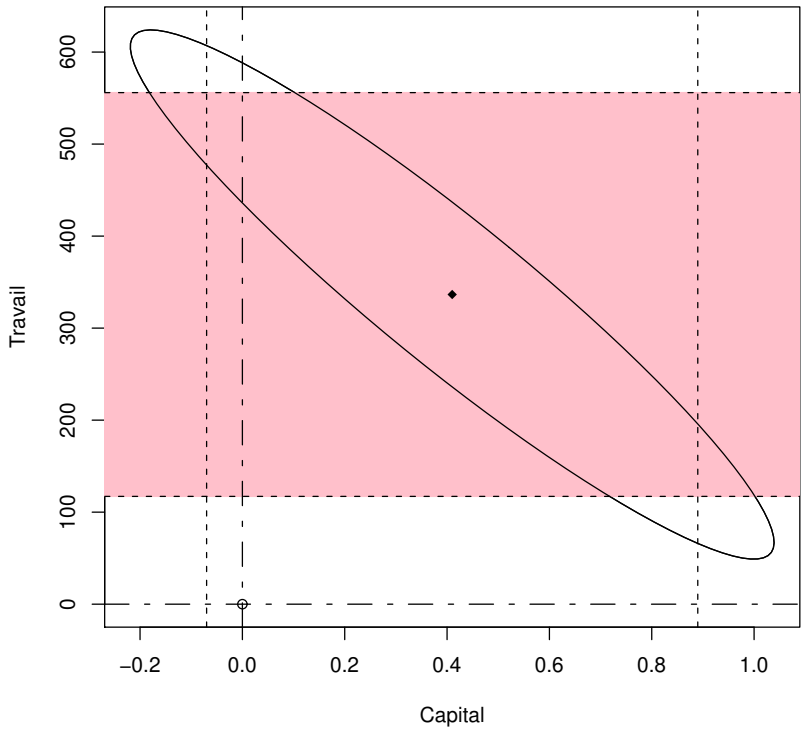
Les instructions suivantes permettent de construire des régions de confiance pour deux paramètres simultanément, c'est-à-dire une ellipse de confiance. Il est intéressant de comparer cette région de confiance simultanée avec les deux que nous obtenons en considérant chacun des paramètres séparément.

```
library(ellipse)
my.confidence.region <- function (g, a=2, b=3, which=0, col='pink') {
  e <- ellipse(g,c(a,b))
  x <- g$coef[a]
  y <- g$coef[b]
  cf <- summary(g)$coefficients
  ia <- cf[a,2]*qt(.975,g$df.residual)
  ib <- cf[b,2]*qt(.975,g$df.residual)
  xmin <- min(c(0,e[,1]))
  xmax <- max(c(0,e[,1]))
  ymin <- min(c(0,e[,2]))
  ymax <- max(c(0,e[,2]))
  plot(e,
       type="l",
       xlim=c(xmin,xmax),
       ylim=c(ymin,ymax),
      )
  if(which==1){ polygon(e,col=col) }
```

```
  else if(which==2){ rect(x-ia,par('usr')[3],x+ia,par('usr')[4],
   col=col,border=col) }
  else if(which==3){ rect(par('usr')[1],y-ib,par('usr')[2],y+ib,
   col=col,border=col) }
  lines(e)
  points(x,y,pch=18)
  abline(v=c(x+ia,x-ia),lty=2)
  abline(h=c(y+ib,y-ib),lty=2)
  points(0,0)
  abline(v=0,lty="F848")
  abline(h=0,lty="F848")
}
my.confidence.region(modele1, which=1)
my.confidence.region(modele1, which=2)
my.confidence.region(modele1, which=3)
```




Travail
0
100
200
300
400
500
600
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Capital


Travail
0
100
200
300
400
500
600
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
Capital