

## EXEMPLE QUI ILLUSTRE LA RÉGRESSION MULTIPLE (COURS 8)

FRÉDÉRIC BERTRAND

*Je vais traiter cet exemple sans me servir du logiciel R ou presque et faire tous les calculs « à la main » pour vous montrer au moins une fois dans ce cours comment nous appliquons les formules mathématiques qui sont données dans ce cours.*

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

**Tableau - Production, travail et capital**

Entreprise	Travail (heures)	Capital (machines/heures)	Production (100 tonnes)
$i$	$x_{i1}$	$x_{i2}$	$y_i$
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

Nous faisons donc l'hypothèse d'un modèle de régression linéaire 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

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

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

$$\widehat{\beta} = (\mathbf{tXX})^{-1}\mathbf{tXy}.$$

Pour cela, nous calculons :

$$\begin{aligned} (\mathbf{tXX}) &= \begin{bmatrix} 9 & 13\,790 & 4\,460 \\ 13\,790 & 21\,672\,100 & 7\,066\,200 \\ 4\,460 & 7\,066\,200 & 2\,323\,600 \end{bmatrix} \\ (\mathbf{tXX})^{-1} &= \begin{bmatrix} 6,304\,777 & -0,007\,800 & 0,011\,620 \\ -0,007\,800 & 0,000\,015 & -0,000\,031 \\ 0,011\,620 & -0,000\,031 & 0,000\,072 \end{bmatrix} \end{aligned}$$

et :

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

Nous obtenons ainsi :

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = (\mathbf{tXX})^{-1}\mathbf{tXy} = \begin{bmatrix} -437,714 \\ 0,336 \\ 0,410 \end{bmatrix}.$$

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

$$\hat{y}(x_1, x_2) = -437,714 + 0,336 x_1 + 0,410 x_2$$

Nous pouvons également calculer :

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

Nous pouvons alors calculer :

$$\begin{aligned} s^2(\widehat{\beta}) &= cm_{res}(\mathbf{tXX})^{-1} = 532 \begin{bmatrix} 6,304\,777 & -0,007\,800 & 0,011\,620 \\ -0,007\,800 & 0,000\,015 & -0,000\,031 \\ 0,011\,620 & -0,000\,031 & 0,000\,072 \end{bmatrix} \\ &= \begin{bmatrix} 3\,355,56 & -4,152 & 6,184 \\ -4,152 & 0,008 & -0,016 \\ 6,184 & -0,016 & 0,038 \end{bmatrix} \end{aligned}$$

Les écart-types  $s(\widehat{\beta}_j)$  des estimateurs  $\widehat{\beta}_j$  sont alors donnés par les racines carrées des éléments diagonaux de cette matrice. Nous avons ainsi :

$$\begin{aligned} s(\widehat{\beta}_0) &= 57,930\,97 \\ s(\widehat{\beta}_1) &= 0,089\,66 \\ s(\widehat{\beta}_2) &= 0,196\,14 \end{aligned}$$

**Nous allons 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.

Nous obtenons à l'aide du logiciel R :

```
> shapiro.test(residuals(modele1))
Shapiro-Wilk normality test
data: residuals(modele1)
W = 0.9157, p-value = 0.3578
```

**Nous ne pouvons donc pas rejeter l'hypothèse nulle  $\mathcal{H}_0$  de normalité au seuil de significativité  $\alpha = 5\%$ . Le risque associé à cette décision est un risque de seconde espèce,  $\beta$  qu'il faudrait calculer en toute théorie.**

Afin de tester l'hypothèse nulle :

$$\mathcal{H}_0 : \beta_j = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \beta_j \neq 0,$$

il s'agit de calculer les statistiques suivantes, pour respectivement  $j = 0$ ,  $j = 1$  et  $j = 2$  :

$$\begin{aligned} t_{obs} &= \frac{-437,714}{57,93097} = -7,556 \\ t_{obs} &= \frac{0,336}{0,08966} = 3,753 \\ t_{obs} &= \frac{0,410}{0,19614} = 2,090. \end{aligned}$$

Comme la valeur critique est donnée par  $t_{6;0,975} = 2,447$ , pour  $j = 0$  et  $j = 1$ , nous décidons de rejeter l'hypothèse nulle  $\mathcal{H}_0$  au seuil de significativité  $\alpha = 5\%$  et d'accepter l'hypothèse alternative  $\mathcal{H}_1$ . Le risque associé à cette décision est un risque de première espèce qui est égal à  $\alpha = 5\%$ .

Par contre, pour  $j = 2$ , nous décidons de ne pas rejeter et donc d'accepter l'hypothèse nulle  $\mathcal{H}_0$  au seuil de significativité  $\alpha = 5\%$ . Le risque associé à cette décision est un risque de seconde espèce qui est égal à  $\beta$  qu'il faudrait calculer en toute théorie.

**Conclusion : cela veut dire que la variable  $X_2$  n'est pas une variable significative dans le modèle.**

Nous calculons **les intervalles de confiance au niveau 95% pour les 3 paramètres  $\beta_0, \beta_1, \beta_2$ .**

$$\begin{aligned} -437,714 \pm 2,447 \times 57,93097 &= [-579,466; -295,961] \\ 0,336 \pm 2,447 \times 0,08966 &= [0,117; 0,556] \\ 0,410 \pm 2,447 \times 0,19614 &= [-0,070; 0,890] \end{aligned}$$

**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 :

$$\begin{aligned}
 SC_{reg} &= \widehat{\beta}^t \mathbf{Xy} - n\bar{y}^2 \\
 &= \begin{bmatrix} -437,71 & 0,336 & 0,41 \end{bmatrix} \times \begin{bmatrix} 2\,530 \\ 4\,154\,500 \\ 1\,378\,500 \end{bmatrix} - 428\,152,14 \\
 &= 144\,695 \\
 SC_{tot} &= \mathbf{y}^t \mathbf{y} - n\bar{y}^2 \\
 &= \begin{bmatrix} 60 & 120 & 190 & \dots & 440 \end{bmatrix} \times \begin{bmatrix} 60 \\ 120 \\ 190 \\ \cdot \\ \cdot \\ \cdot \\ 440 \end{bmatrix} - 428\,152,14 \\
 &= 147\,889
 \end{aligned}$$

Nous avons :

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

Nous obtenons le tableau d'ANOVA donné par le tableau ci-dessous. Nous voulons tester l'hypothèse nulle :

$$\mathcal{H}_0 : \beta_1 = \beta_2 = 0$$

contre l'hypothèse alternative :

$$\mathcal{H}_1 : \exists i = 1, 2 / \beta_i \neq 0.$$

Comme la statistique  $F_{obs} = 135,92$  est supérieure à la valeur critique  $F_{(0,95,2,6)} = 5,14$ , nous décidons de rejeter l'hypothèse nulle  $\mathcal{H}_0$  au seuil de significativité  $\alpha = 5\%$ .

Source de variation	Somme des carrés	ddl	Carrés moyens	$F_{obs}$
Régression	144 695	2	72 348	135,92
Résiduelle	3 194	6	532	
Totale	147 889	8		

Bien sûr, vous pouvez retrouver tous ces résultats avec le logiciel R. Voici les commandes à taper :

```

> production<-c(60,120,190,250,300,360,380,430,440)
> travail<-c(1100,1200,1430,1500,1520,1620,1800,1820,1800)
> capital<-c(300,400,420,400,510,590,600,630,610)
> modele<-lm(production~travail+capital)
> modele

```

Call:

```
lm(formula = production ~ travail + capital)
```

Coefficients:

```
(Intercept)      travail      capital
-437.7136      0.3365      0.4100
> summary(modele)
```

```
Call:
lm(formula = production ~ travail + capital)
```

```
Residuals:
    Min       1Q   Median       3Q      Max
-34.050 -10.129   4.526  17.080  21.850
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -437.71363    57.93097  -7.556 0.000279 ***
travail      0.33653     0.08966   3.753 0.009474 **
capital      0.41002     0.19614   2.090 0.081555 .
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 23.07 on 6 degrees of freedom
Multiple R-squared: 0.9784, Adjusted R-squared: 0.9712
F-statistic: 135.9 on 2 and 6 DF, p-value: 1.007e-05
```

Avec la commande `summary`, nous avons obtenu les 3 tests de Student. Pour obtenir les 3 intervalles de confiance pour chacun des paramètres, nous allons utiliser la commande `confint`.

```
> confint(modele)
              2.5 %      97.5 %
(Intercept) -579.46562195 -295.9616437
travail      0.11712905    0.5559315
capital      -0.06993004    0.8899614
```

Enfin, pour obtenir le tableau de l'ANOVA, nous allons utiliser la commande `anova`

```
> anova(modele)
Analysis of Variance Table

Response: production
      Df Sum Sq Mean Sq F value Pr(>F)
travail  1 142369  142369 267.4631 3.328e-06 ***
capital  1   2326    2326   4.3697  0.08155 .
Residuals 6   3194     532
---

```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)
```