EXEMPLE QUI ILLUSTRE LE COURS « RÉGRESSION LINÉAIRE MULTIPLE » (COURS 4 ET SUITE)

On va traiter cet exemple issu du livre « Analyse de régression appliquée » de Yadolah Dodge, Dunod, sans se servir de Minitab et faire tous les calculs « à la main » pour comprendre et voir au moins une fois comment appliquer les formules mathématiques qui sont introduites 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, les heures de travail X_1 et le capital utilisé X_2.

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 :

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon.$$

Comme on a à notre disposition qu'un échantillon de mesures, les variables aléatoires symbolisées par des grandes lettres deviennent des observations symbolisées par des petites lettres. Par conséquent, on a

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

Tableau - Travail, capital et production

Entreprise i	Travail (heures) x_{i1}	Capital (machines/heures) x_{i2}	Production (100 tonnes) 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 $\widehat{\beta}$ défini par l'égalité suivante :

$$\widehat{\beta} = ({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1\,{}^{\mathbf{t}}}\mathbf{X}\mathbf{y}.$$

Pour cela, on calcule :

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

$$({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1} = \begin{bmatrix} 6.304\,777 & -0.007\,800 & 0.011\,620 \\ -0.007\,800 & 0.000\,015 & -0.000\,031 \\ 0.011\,620 & -0.000\,031 & 0.000\,072 \end{bmatrix}$$

et :

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

On obtient ainsi :

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = ({}^{\mathbf{t}}\mathbf{X}\mathbf{X})^{-1\,{}^{\mathbf{t}}}\mathbf{X}\mathbf{y} = \begin{bmatrix} -437.710 \\ 0.336 \\ 0.410 \end{bmatrix}.$$

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

$$\widehat{y}(x_1, x_2) = -437.710 + 0.336\,x_1 + 0.410\,x_2.$$

On peut également calculer :

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

On peut alors calculer :

$$\begin{aligned} s^2\left(\widehat{\beta}\right) &= s^2({}^{\mathrm{t}}\mathbf{X}\mathbf{X})^{-1} = 532 \begin{bmatrix} 6.304\,777 & -0.007\,800 & 0.011\,620 \\ -0.007\,800 & 0.000\,015 & -0.000\,031 \\ 0.011\,620 & -0.000\,031 & 0.000\,072 \end{bmatrix} \\ &= \begin{bmatrix} 3\,355.56 & -4.152 & 6.184 \\ -4.152 & 0.008 & -0.016 \\ 6.184 & -0.016 & 0.038 \end{bmatrix}. \end{aligned}$$

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

$$\begin{aligned} s\left(\widehat{\beta}_0\right) &= 57.93 \\ s\left(\widehat{\beta}_1\right) &= 0.089\,66 \\ s\left(\widehat{\beta}_2\right) &= 0.196\,1. \end{aligned}$$

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 de signification $\alpha = 5\%$**.**

$$H_0 : \beta_0 = \beta_1 = \beta_2 = 0$$

contre l'hypothèse alternative :

$$H_1 : \exists\ j \in \{0, 1, 2\} \text{ tel que } \beta_j \neq 0,$$

il s'agit de calculer les statistiques suivantes :

$$\begin{aligned} t_0 &= \frac{-437.71}{57.93} = -7.56 \\ t_1 &= \frac{0.336}{0.089\,66} = 3.75 \\ t_2 &= \frac{0.41}{0.196\,1} = 2.09 \end{aligned}$$

Comme la valeur critique est donnée par $t_{0.025;6} = 2.45$, on rejette l'hypothèse nulle H_0 au seuil de signification $\alpha = 0.05$ pour $j = 0$ et $j = 1$, mais on accepte l'hypothèse nulle H_2 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** $\beta_0, \beta_1, \beta_2$.

$$\begin{aligned} -437.710 \pm 2.45 \times 57.93 &= [-579.64; -295.78] \\ 0.336 \pm 2.45 \times 0.089\,66 &= [0.116; 0.556] \\ 0.410 \pm 2.45 \times 0.196\,1 &= [-0.07; 0.89] \end{aligned}$$

Remarque : la valeur 0 est comprise dans l'intervalle de confiance pour β_2.

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

$$\begin{aligned} SC_{reg} &= \widehat{\beta} \times {}^{\mathbf{t}}\mathbf{X}\mathbf{y} - n\overline{y}^2 \\ &= \begin{bmatrix} -437.710 & 0.336 & 0.410 \end{bmatrix} \times \begin{bmatrix} 2\,530 \\ 4\,154\,500 \\ 1\,378\,500 \end{bmatrix} - 428\,152.14 \\ &= 144\,695 \end{aligned}$$

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

On a :

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

On obtient le tableau d'analyse de (la) variance donné par le tableau ci-dessous. On peut tester l'hypothèse nulle :

$$H_0 : \beta_1 = \beta_2 = 0$$

contre l'hypothèse alternative :

$$H_1 : \exists j \in \{1, 2\} \quad \beta_j \neq 0.$$

Comme la statistique $F_{obs} = 135.92$ est supérieure à la valeur critique $F_{(0.05;\,2;\,6)} = 5.14$, on rejette l'hypothèse nulle H_0 au seuil de significativité $\alpha = 0.05$.

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		