

Régression linéaire multiple

Olivier Levyne (2021)

1. Contexte

Une enquête a permis de collecter n observations de valeurs prises par des variables Y, X_1, X_2, \dots, X_p :

Observations	Y	X_1	X_2	\dots	X_p
1	y_1	$x_{1,1}$	$x_{1,2}$	\dots	$x_{1,p}$
2	y_2	$x_{2,1}$	$x_{2,2}$	\dots	$x_{2,p}$
\dots	\dots	\dots	\dots	\dots	\dots
n	y_n	$x_{n,1}$	$x_{n,2}$	\dots	$x_{n,p}$

Y est la variable à expliquer par les p variables explicatives X_1, X_2, \dots, X_p .

2. Finalité

La finalité est de déterminer les coefficients a_0, a_1, \dots, a_p de telle sorte que la régression suivante soit optimale :

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_p X_p + \varepsilon$$

où ε est le terme d'erreur aléatoire du modèle ; il définit une variable aléatoire non observée telle que $E(\varepsilon) = 0$ et $V(\varepsilon) = \sigma^2$

La méthode des moindres carrés consiste à déterminer les coefficients a_0, a_1, \dots, a_p de telle sorte que la somme des carrés des écarts entre les observations y_1, y_2, \dots, y_n de Y et les valeurs induites par la régression soit minimale :

Les écarts définissent le système suivant :

$$\begin{cases} \varepsilon_1 = y_1 - (a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \dots + a_p x_{1,p}) \\ \varepsilon_2 = y_2 - (a_0 + a_1 x_{2,1} + a_2 x_{2,2} + \dots + a_p x_{2,p}) \\ \dots \\ \varepsilon_n = y_n - (a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \dots + a_p x_{n,p}) \end{cases}$$

Il convient donc de déterminer les coefficients a_0, a_1, \dots, a_p de telle sorte la somme suivante soit minimale :

$$\sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n [y_n - (a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \dots + a_p x_{i,p})]^2$$

3. Formalisation

Le système ci-dessus peut aussi s'écrire :

$$\begin{cases} y_1 = a_0 + a_1 x_{1,1} + a_2 x_{1,2} + \dots + a_p x_{1,p} + \varepsilon_1 \\ y_2 = a_0 + a_1 x_{2,1} + a_2 x_{2,2} + \dots + a_p x_{2,p} + \varepsilon_2 \\ \dots \\ y_n = a_0 + a_1 x_{n,1} + a_2 x_{n,2} + \dots + a_p x_{n,p} + \varepsilon_n \end{cases}$$

Matriciellement :

$$\begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \dots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix}$$

Ce qui revient à :

$$Y = X.a + \varepsilon \Leftrightarrow \varepsilon = Y - X.a$$

Où X est la matrice à n lignes et $p+1$ colonnes ci-dessus et a la matrice à $p+1$ lignes et 1 colonnes. Ainsi, $X.a$, comme ε et Y , sont des matrices à n lignes et 1 colonne ce qui est bien cohérent avec le système initial.

La méthode des moindres carrés consiste alors à déterminer la matrice \hat{a} à n lignes et 1 colonne qui minimise :

$$\sum_{i=1}^n \varepsilon_i^2 = \varepsilon_1^2 + \varepsilon_2^2 + \dots + \varepsilon_n^2 = (\varepsilon_1 \quad \varepsilon_2 \quad \dots \quad \varepsilon_n) \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix} = {}^t\varepsilon.\varepsilon$$

On a alors :

$${}^t\varepsilon.\varepsilon = {}^t(Y - X\hat{a}).(Y - X\hat{a})$$

$${}^t\varepsilon.\varepsilon = ({}^tY - {}^t\hat{a} {}^tX.)(Y - X\hat{a})$$

$${}^t\varepsilon.\varepsilon = {}^tYY - {}^tYX\hat{a} - {}^t\hat{a} {}^tXY + {}^t\hat{a} {}^tXX\hat{a}$$

Or :

- ${}^t\hat{a}$ est une matrice à 1 ligne et $p+1$ colonnes ,
- tXY est le produit d'une matrice à $p+1$ lignes et n colonnes par une matrice à n lignes et 1 colonne. Donc tXY est une matrice à $p+1$ lignes et 1 colonne

Ainsi, ${}^t\hat{a} {}^tXY$ est une matrice à 1 et 1 colonne ; ${}^t\hat{a} {}^tXY$ est donc un scalaire.

Par conséquent, cette matrice est égale à sa transposée :

$${}^t\hat{a} {}^tXY = {}^tYX\hat{a}$$

Ainsi :-

$${}^t\varepsilon.\varepsilon = {}^tYY - 2 {}^t\hat{a} {}^tXY + {}^t\hat{a} {}^tXX\hat{a}$$

Pour déterminer la valeur de \hat{a} qui minimise la somme des carrés des écarts c'est-à-dire ${}^t\varepsilon.\varepsilon$, il convient de déterminer pour quelle valeur la dérivée de ${}^t\varepsilon.\varepsilon$ par rapport à \hat{a} s'annule. Ainsi :

$$\frac{\partial {}^t\varepsilon.\varepsilon}{\partial \hat{a}} = 0 \Leftrightarrow -2 {}^tXY + {}^tXX\hat{a} + {}^t\hat{a} {}^tXX = 0$$

Or :

- ${}^t\hat{a}$ est une matrice à 1 ligne et $p+1$ colonnes ,
- tX est une matrice à $p+1$ lignes et n colonnes
- X est une matrice à n lignes et $p+1$ colonnes
- \hat{a} est une matrice à $p+1$ lignes et 1 colonne

Donc ${}^t\hat{a} {}^tXX$ est un scalaire ; cette matrice est donc égale à sa transposée.

Ainsi :

$${}^t\hat{a} {}^tXX = {}^tXX\hat{a}$$

Ainsi :

$$\frac{\partial {}^t\varepsilon.\varepsilon}{\partial \hat{a}} = 0 \Leftrightarrow -2 {}^tXY + 2 {}^tXX\hat{a} = 0 \Leftrightarrow {}^tXY = {}^tXX\hat{a}$$

Dès lors

$${}^tXY = {}^tXX\hat{a}$$

Et finalement, sous réserve que tXX soit inversible :

$$\hat{a} = ({}^tXX)^{-1} {}^tXY$$

4. Exemple par calcul matriciel

Pour 10 banques cotées le tableau ci-dessous présente le rendement annuel de leur action, leur ROE et leur ratio de solvabilité. En outre, pour mener à bien la détermination d'une régression entre d'une part le rendement du titre, d'autre part le ROE et la solvabilité de la banque par le calcul matriciel et reconstituer la matrice X, une colonne composée uniquement de 1 a été ajoutée :

		X ₁	X ₂	Y
Banque	1	ROE	Solvabilité	Rendement
1	1	10%	10%	8%
2	1	15%	16%	12%
3	1	12%	12%	10%
4	1	20%	16%	18%
5	1	8%	10%	5%
6	1	4%	9%	-3%
7	1	16%	13%	14%
8	1	10%	12%	9%
9	1	11%	17%	15%
10	1	6%	8%	-1%
		Matrice X		Matrice Y

Les calculs intermédiaires et la matrice \hat{a} des coefficients de la régression sont les suivants :

Matrice tX.X			Inverse de tX.X		
10,00	1,12	1,23	1,90	3,78	-18,08
1,12	0,15	0,15	3,78	118,62	-138,76
1,23	0,15	0,16	-18,08	-138,76	273,32
Matrice tX.Y					
		\hat{a}			
0,87		-0,12	Constante		
0,12		0,80	. ROE		
0,12		0,94	. Solvabilité		

D'un point de vue syntaxique, les fonctions **produitmat**(matrice 1 ; matrice 2) et **transpose**(matrice) ont été utilisées.

Ainsi, $Y = -0,12 + 0,80X_1 + 0,94X_2$

soit encore : Rendement = -0,12 + 0,80 ROE + 0,94.Solvabilité

Ces résultats peuvent être vérifiés grâce aux utilitaires d'analyse d'Excel qui permettent également de juger de la significativité de chacun des coefficients obtenus.

5. Cohérence avec l'équation de la droite de régression

Dans l'hypothèse d'une seule variable explicative X , l'équation de la droite de régression est : $Y = aX + b$ ou encore, en utilisant les notations de la régression obtenue dans le paragraphe 3 : $Y = a_0 + a_1X$ où :

$$a_1 = \frac{\text{cov}(X, Y)}{V(X)}$$

$$a_0 = E(Y) - a_1 E(X) = E(Y) - \frac{\text{cov}(X, Y)}{V(X)} E(X) = \frac{E(Y)V(X) - \text{cov}(X, Y)E(X)}{V(X)}$$

$$\begin{aligned} a_0 &= \frac{E(Y)V(X) - [E(XY) - E(X)E(Y)]E(X)}{V(X)} \\ &= \frac{E(Y)V(X) - E(XY)E(X) + E(Y)[E(X)]^2}{V(X)} \end{aligned}$$

Or, d'après la formule de Kœnig-Huyghens : $V(X) = E(X^2) - [E(X)]^2$

Donc : $[E(X)]^2 = E(X^2) - V(X)$. Ainsi :

$$a_0 = \frac{E(Y)[V(X) + E(X^2) - V(X)] - E(XY)E(X)}{V(X)}$$

Finalement :

$$a_0 = \frac{E(Y)E(X^2) - E(XY)E(X)}{V(X)}$$

En effet, dans l'hypothèse d'une seule variable explicative :

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} \text{ et } X = \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ \dots & \dots \\ 1 & x_{n,1} \end{pmatrix} \Rightarrow {}^tX = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{n,1} \end{pmatrix}$$

Dès lors :

$${}^tXX = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{n,1} \end{pmatrix} \begin{pmatrix} 1 & x_{1,1} \\ 1 & x_{2,1} \\ \dots & \dots \\ 1 & x_{n,1} \end{pmatrix} = \begin{pmatrix} n & \sum_{i=1}^n x_{i,1} \\ \sum_{i=1}^n x_{i,1} & \sum_{i=1}^n x_{i,1}^2 \end{pmatrix}$$

Et :

$$({}^tXX)^{-1} = \frac{1}{n \sum_{i=1}^n x_{i,1}^2 - (\sum_{i=1}^n x_{i,1})^2} \begin{pmatrix} \sum_{i=1}^n x_{i,1}^2 & -\sum_{i=1}^n x_{i,1} \\ -\sum_{i=1}^n x_{i,1} & n \end{pmatrix}$$

Par ailleurs :

$${}^t_{XY} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ x_{1,1} & x_{2,1} & \dots & x_{n,1} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1} y_i \end{pmatrix}$$

Donc :

$$({}^t_{XX})^{-1} {}^t_{XY} = \frac{1}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]} \begin{pmatrix} \sum_{i=1}^n x_{i,1}^2 & - \sum_{i=1}^n x_{i,1} \\ - \sum_{i=1}^n x_{i,1} & n \end{pmatrix} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1} y_i \end{pmatrix}$$

$$({}^t_{XX})^{-1} {}^t_{XY} = \frac{1}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]} \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i,1} y_i \end{pmatrix}$$

$$({}^t_{XX})^{-1} {}^t_{XY} = \frac{1}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]} \begin{pmatrix} \sum_{i=1}^n x_{i,1}^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_{i,1} \sum_{i=1}^n x_{i,1} y_i \\ n \sum_{i=1}^n x_{i,1} y_i - \sum_{i=1}^n x_{i,1} \sum_{i=1}^n y_i \end{pmatrix}$$

En considérant la 2^{ème} ligne de la matrice :

$$a_1 = \frac{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1} y_i - \frac{1}{n} \sum_{i=1}^n x_{i,1} \frac{1}{n} \sum_{i=1}^n y_i \right]}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]} = \frac{cov(X, Y)}{V(X)}$$

En considérant la 1^{ère} ligne de la matrice :

$$a_0 = \frac{\sum_{i=1}^n x_{i,1}^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_{i,1} \sum_{i=1}^n x_{i,1} y_i}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]}$$

$$a_0 = \frac{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 \frac{1}{n} \sum_{i=1}^n y_i - \frac{1}{n} \sum_{i=1}^n x_{i,1} \frac{1}{n} \sum_{i=1}^n x_{i,1} y_i \right]}{n^2 \left[\frac{1}{n} \sum_{i=1}^n x_{i,1}^2 - \left(\frac{1}{n} \sum_{i=1}^n x_{i,1} \right)^2 \right]}$$

La simplification par n^2 permet de retrouver :

$$a_0 = \frac{E(Y)E(X^2) - cov(X, Y)E(X)}{V(X)}$$

6. Détermination des résidus

Les résidus ε correspondent aux écarts entre les valeurs prises par Y et celles induites par la régression ($X \hat{a}$) et qui correspondent à des prédictions souvent notées \hat{Y} . Ainsi :

$$\varepsilon = Y - X \hat{a} = Y - \hat{Y}$$

		X_1	X_2	Y	$X.\hat{a}$	$\varepsilon = Y - X.\hat{a}$		
Banque	1	ROE	Solvabilité	Rendement	Prédiction	Résidus		
1	1	10%	10%	8,0%	5,56%	2,44%		\hat{a}
2	1	15%	16%	12,0%	15,25%	-3,25%		-0,12
3	1	12%	12%	10,0%	9,06%	0,94%		0,80
4	1	20%	16%	18,0%	19,28%	-1,28%		0,94
5	1	8%	10%	5,0%	3,95%	1,05%		
6	1	4%	9%	-3,0%	-0,21%	-2,79%		
7	1	16%	13%	14,0%	13,22%	0,78%		
8	1	10%	12%	9,0%	7,45%	1,55%		
9	1	11%	17%	15,0%	12,98%	2,02%		
10	1	6%	8%	-1,0%	0,45%	-1,45%		
		Matrice X			Matrice Y			

7. Décomposition de la variance

La régression recherchée et l'expression de la variable à expliquer, en fonction des variables explicatives, qui minimise la somme SCR des carrés des résidus. Ainsi :

$$SCR = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

La prédiction issue de la régression est alors parfaite si $SCR = 0$.

Par ailleurs la somme SCT des carrés totaux, égale à $n.V(Y)$, est définie par :

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2$$

Il est possible d'établir que $SCT = SCE + SCR$ où SCE est la somme des carrés expliqués. SCE est la variabilité expliquée par le modèle c'est-à-dire la variation de Y expliquée par X . Sa formule est :

$$SCE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

En effet :

$$SCT = \sum_{i=1}^n [(y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})]^2$$

$$SCT = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 - 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$SCT = \sum_{i=1}^n (y_i - \hat{y}_i)^2 + \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 - 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

$$SCT = SCR + SCE - 2 \sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Montrons que :

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = 0$$

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_{i=1}^n (\hat{y}_i - \bar{y}) \varepsilon_i$$

Or, on sait que $\varepsilon = Y - \hat{Y}$; dès lors :

$$\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n (y_i - \hat{y}_i)$$

et :

$$\varepsilon_i = y_i - \hat{y}_i \Leftrightarrow y_i = \hat{y}_i + \varepsilon_i = \widehat{a}_0 + \widehat{a}_1 x_{i,1} + \widehat{a}_2 x_{i,2} + \cdots \widehat{a}_p x_{i,p} + \varepsilon_i$$

$$y_i = \widehat{a}_0 + \sum_{k=1}^p \widehat{a}_k x_{i,k} + \varepsilon_i$$

$$\sum_{i=1}^n y_i = n \cdot \widehat{a}_0 + \sum_{i=1}^n \sum_{k=1}^p \widehat{a}_k x_{i,k} + \sum_{i=1}^n \varepsilon_i$$

$$\sum_{i=1}^n \varepsilon_i = n\bar{y} - n \cdot \widehat{a}_0 - \sum_{i=1}^n \sum_{k=1}^p \widehat{a}_k x_{i,k}$$

$$\sum_{i=1}^n \varepsilon_i = n\bar{y} - n \cdot \widehat{a}_0 - \sum_{k=1}^p \widehat{a}_k \sum_{i=1}^n x_{i,k}$$

$$\sum_{i=1}^n \varepsilon_i = n\bar{y} - n \cdot \widehat{a}_0 - n \sum_{k=1}^p \widehat{a}_k \bar{x}_k$$

Or :

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \cdots a_p X_p + \varepsilon = a_0 + \sum_{k=1}^n a_k X_k + \varepsilon$$

Donc :

$$\widehat{a}_0 = Y - \sum_{k=1}^n \widehat{a}_k X_k$$

En particulier :

$$\widehat{a}_0 = y_i - \sum_{k=1}^n \widehat{a}_k x_{i,k}$$

$$n. \widehat{a}_0 = n\bar{y} - \sum_{k=1}^n \widehat{a}_k \bar{x}_k$$

Ainsi :

$$\sum_{i=1}^n \varepsilon_i = n\bar{y} - \left[n\bar{y} - \sum_{k=1}^n \widehat{a}_k \bar{x}_k \right] - n \sum_{k=1}^p \widehat{a}_k \bar{x}_k$$

Finalement :

$$\sum_{i=1}^n \varepsilon_i = 0$$

Dès lors :

$$\sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n (y_i - \widehat{y}_i) = 0 \Rightarrow \bar{y} = \bar{\widehat{y}}$$

$$\sum_{i=1}^n (\widehat{y}_i - \bar{y}) \varepsilon_i = \sum_{i=1}^n \widehat{y}_i \varepsilon_i - \bar{y} \sum_{i=1}^n \varepsilon_i = \sum_{i=1}^n \widehat{y}_i \varepsilon_i - 0$$

Or

$$\sum_{i=1}^n \widehat{y}_i \varepsilon_i = \widehat{y}_1 \varepsilon_1 + \widehat{y}_2 \varepsilon_2 + \dots + \widehat{y}_n \varepsilon_n = (\widehat{y}_1 \quad \widehat{y}_2 \quad \dots \quad \widehat{y}_n) \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{pmatrix} = {}^t \widehat{Y} \cdot \varepsilon = {}^t (X \widehat{a}) \cdot \varepsilon$$

$$\sum_{i=1}^n \widehat{y}_i \varepsilon_i = {}^t \widehat{a} \cdot {}^t X \varepsilon$$

Par ailleurs :

$$\widehat{a} = ({}^t X X)^{-1} {}^t X Y \Rightarrow {}^t X X \widehat{a} = {}^t X Y \Rightarrow {}^t X X \widehat{a} - {}^t X Y = 0 \Rightarrow {}^t X (X \widehat{a} - Y) = 0$$

En d'autres termes :

$${}^tX(\hat{Y} - Y) = 0 \Rightarrow {}^tX\varepsilon = 0 \Rightarrow \sum_{i=1}^n \hat{y}_i \varepsilon_i = 0$$

Conclusion :

$$\sum_{i=1}^n (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

Et :

$$SCT = SCR + SCE$$

D'un point de vue pratique :

$$SCT = \sum_{i=1}^n (y_i - \bar{y})^2 = n.V(Y)$$

$$SCR = \sum_{i=1}^n \varepsilon_i^2$$

$$SCE = SCT - SCR$$

		X ₁	X ₂	Y	X.â	ε = Y - X.â
Banque	1	ROE	Solvabilité	Rendement	Prédiction	Résidus
1	1	10%	10%	8,0%	5,56%	2,44%
2	1	15%	16%	12,0%	15,25%	-3,25%
3	1	12%	12%	10,0%	9,06%	0,94%
4	1	20%	16%	18,0%	19,28%	-1,28%
5	1	8%	10%	5,0%	3,95%	1,05%
6	1	4%	9%	-3,0%	-0,21%	-2,79%
7	1	16%	13%	14,0%	13,22%	0,78%
8	1	10%	12%	9,0%	7,45%	1,55%
9	1	11%	17%	15,0%	12,98%	2,02%
10	1	6%	8%	-1,0%	0,45%	-1,45%
		Matrice X		Matrice Y		
Variance	0,000%	0,208%	0,090%	0,412%	0,375%	0,037%
Somme des carrés				4,121%	0,375%	0,002%
				SCT	SCE	SCR

8. Coefficient de détermination

Le coefficient de détermination R^2 vise à quantifier la capacité du modèle à expliquer les variations de Y .

Dès lors :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT}$$

Si R^2 est proche de 1 alors le modèle est proche de la réalité.

En revanche, si R^2 est proche de 0 alors le modèle explique très mal la réalité. Il faut alors trouver un meilleur modèle

Analyse de la variance				
	Somme carrés	d° de liberté	Carrés moy	F_{emp}
SCE	3,75%	2	1,87%	35,36
SCR	0,37%	7	0,05%	
SCT	4,12%	9		
$R^2 = SCE/SCT$	91%			

Le nombre de degrés de liberté est égal au nombre de termes impliqués dans les sommes (c'est-à-dire le nombre d'observations) moins le nombre de paramètres estimés dans ces sommes.

Ainsi :

- Le calcul de la SCT passe par l'estimation de \bar{y} . Donc $DDL(SCT) = n - 1$ où n est le nombre d'observations. Dans l'exemple ci-dessus, $DDL(SCT) = 10 - 1 = 9$;
- 3 coefficients estimés sont requis ici pour obtenir la projection et former la SCR . Donc $DDL(SCR) = 10 - 3 = 7 = n - p$
- $DDL(SCE) = DDL(SCT) - DDL(SCR) = 9 - 7 = 2 = n - 1 - (n - p) = p - 1$

Les carrés moyens correspondent au rapport entre la somme des carrés et le nombre de degrés de liberté. Ils permettent de calculer la statistique F_{emp} décrite dans le paragraphe suivant.

9. Test de Fisher Snedecor de nullité de l'ensemble des coefficients

On dispose de n observations d'une variable Y à expliquer et de p variables explicatives X_1, X_2, \dots, X_p

Il s'agit de tester l'hypothèse H au seuil (ou risque d'erreur) α : $a_0 = a_1 = \dots = a_p = 0$

Sur la base des observations empiriques disponibles, il convient de calculer la statistique F_{emp} définie par :

$$F_{emp} = \frac{\frac{SCE}{p-1}}{\frac{SCR}{n-p}} = \frac{CME}{CMR}$$

Si l'hypothèse H de nullité de tous les coefficients est vraie alors :

$$F_{emp} \hookrightarrow \mathcal{F}(p-1, n-p)$$

La table de la distribution de la loi de Fisher Snedecor fournit la valeur t telle que :

$$P(F_{emp} > t) = \alpha$$

Le critère de décision est le suivant :

- Si $F_{emp} \geq F_{th}$ alors l'hypothèse H de nullité de tous les coefficients est rejetée au seuil α
- Si $F_{emp} < F_{th}$ alors l'hypothèse H est acceptée au seuil α .

Dans l'exemple précédent :

$$F_{emp} = 35,36$$

Et :

$$F_{emp} \hookrightarrow \mathcal{F}(10-1, 10-3) = \mathcal{F}(9, 7)$$

La table de la distribution de la loi de Fisher Snedecor (cf. : annexe 1) fournit la valeur t qui a seulement 5% de chances d'être dépassée :

$$P(F_{emp} > t) = 5\%$$

Pour $\mathcal{F}(9, 7)$, la table permet de lire : $t = 3,68$

$$P(F_{emp} > 3,68) = 5\%$$

En d'autres termes, si l'hypothèse H de nullité de tous les coefficients est vraie alors il y a 5% de chances pour que F dépasse 3,68. Or, $F_{emp} = 35,36$

$F_{emp} \geq F_{th}$, alors l'hypothèse H est rejetée au seuil de 5%.

Il peut donc être affirmé, avec un risque d'erreur de 5%, qu'il y a donc au moins un coefficient non nul

10. Test de Student de nullité de chaque coefficient

La statistique t est calculée sur la base des observations disponibles :

$$t = \frac{\hat{a}_j}{s(\hat{a}_j)}$$

où :

- $s^2(\hat{a}_j)$ est l'élément diagonal d'indice j de $CMR \cdot ({}^tXX)^{-1}$
- CMR est la somme des carrés résiduels moyens

Dans l'exemple précédent :

Inverse de ${}^tXX = ({}^tXX)^{-1}$				
1,90	3,78	-18,08		
3,78	118,62	-138,76		
-18,08	-138,76	273,32		
CMR	0,05%			
Matrice des var et cov des coefficients			\hat{a}	t
0,10%	0,20%	-0,96%	-0,12	-3,76
0,20%	6,29%	-7,36%	0,80	3,21
-0,96%	-7,36%	14,49%	0,94	2,48

Par ailleurs, si l'hypothèse de nullité du coefficient a_j est vraie, alors :

$$T_{emp} \hookrightarrow \mathcal{S}(n - p)$$

La table de la distribution de Student (cf. : annexe 2) fournit t tel que :

$$P(T_{emp} < t) = 1 - \alpha$$

Or :

$$P(-t < T_{emp} < t) = 1 - \alpha \Leftrightarrow P(T_{emp} < t) - P(T_{emp} < -t) = 1 - \alpha$$

Soit encore :

$$2 \cdot P(T_{emp} < t) - 1 = 1 - \alpha \Leftrightarrow P(T_{emp} < t) = 1 - \frac{\alpha}{2}$$

En résumé :

$$P(-t < T_{emp} < t) = 1 - \alpha \Leftrightarrow P(T_{emp} < t) = 1 - \frac{\alpha}{2}$$

$$\text{Si } \alpha = 5\% \text{ alors } 1 - \frac{\alpha}{2} = 0,975$$

Le critère de décision est alors le suivant :

- Si la valeur de T_{emp} observée sur l'échantillon de données est comprise dans l'intervalle $[-t, t]$ alors l'hypothèse de nullité du coefficient \hat{a}_j est acceptable au seuil α
- Sinon, l'hypothèse de nullité du coefficient \hat{a}_j est rejetée

En retenant les hypothèses de l'exemple précédent, la table permet de lire :

$$T_{emp} \hookrightarrow \mathcal{S}(10 - 3) = \mathcal{S}(7) \Rightarrow P(T_{emp} < 2,36) = 0,975$$

En d'autres termes :

$$P(-2,36 < T_{emp} < 2,36) = 0,95$$

Les valeurs de t issues de l'échantillon (-3,76, 3,21 et 2,48) et correspondant aux 3 coefficients sont en dehors de l'intervalle $[-2,36, 2,36]$. L'hypothèse de nullité de chacun des 3 coefficients doit donc être rejetée au seuil de 5%

Annexe 1

Fisher Snedecor distribution																		
Assuming T -> F(m,n), provision of t so that P(T > t) = 0,05																		
Example: when T -> F(5,4), P(T > 6,26) = 0,05																		
n	m																	
	1	2	3	4	5	6	7	8	9	10	11	12	13	20	30	38	44	52
1	161	200	216	225	230	234	237	239	241	242	243	244	245	248	250	251	251	252
2	18,5	19,0	19,2	19,2	19,3	19,3	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,4	19,5	19,5	19,5	19,5
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,76	8,74	8,73	8,66	8,62	8,60	8,59	8,58
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,94	5,91	5,89	5,80	5,75	5,72	5,71	5,70
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,70	4,68	4,66	4,56	4,50	4,47	4,46	4,44
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00	3,98	3,87	3,81	3,78	3,76	3,75
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,60	3,57	3,55	3,44	3,38	3,35	3,33	3,32
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,31	3,28	3,26	3,15	3,08	3,05	3,03	3,02
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,10	3,07	3,05	2,94	2,86	2,83	2,82	2,80
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,94	2,91	2,89	2,77	2,70	2,67	2,65	2,63
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,82	2,79	2,76	2,65	2,57	2,54	2,52	2,50
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,72	2,69	2,66	2,54	2,47	2,43	2,41	2,40
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,63	2,60	2,58	2,46	2,38	2,35	2,33	2,31
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,57	2,53	2,51	2,39	2,31	2,27	2,25	2,24
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,51	2,48	2,45	2,33	2,25	2,21	2,19	2,17
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,46	2,42	2,40	2,28	2,19	2,16	2,14	2,12
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,41	2,38	2,35	2,23	2,15	2,11	2,09	2,07
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,37	2,34	2,31	2,19	2,11	2,07	2,05	2,03
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,34	2,31	2,28	2,16	2,07	2,03	2,01	1,99
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,31	2,28	2,25	2,12	2,04	2,00	1,98	1,96
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,28	2,25	2,22	2,10	2,01	1,97	1,95	1,93
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,26	2,23	2,20	2,07	1,98	1,95	1,93	1,90
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,24	2,20	2,18	2,05	1,96	1,92	1,90	1,88
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,22	2,18	2,15	2,03	1,94	1,90	1,88	1,86
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,20	2,16	2,14	2,01	1,92	1,88	1,86	1,84
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22	2,18	2,15	2,12	1,99	1,90	1,86	1,84	1,82
27	4,21	3,35	2,96	2,73	2,57	2,46	2,37	2,31	2,25	2,20	2,17	2,13	2,10	1,97	1,88	1,84	1,82	1,80
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19	2,15	2,12	2,09	1,96	1,87	1,83	1,81	1,79
29	4,18	3,33	2,93	2,70	2,55	2,43	2,35	2,28	2,22	2,18	2,14	2,10	2,08	1,94	1,85	1,81	1,79	1,77
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,13	2,09	2,06	1,93	1,84	1,80	1,78	1,76
36	4,11	3,26	2,87	2,63	2,48	2,36	2,28	2,21	2,15	2,11	2,07	2,03	2,00	1,87	1,78	1,73	1,71	1,69
38	4,10	3,24	2,85	2,62	2,46	2,35	2,26	2,19	2,14	2,09	2,05	2,02	1,99	1,85	1,76	1,72	1,69	1,67
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,04	2,00	1,97	1,84	1,74	1,70	1,68	1,65
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,95	1,92	1,89	1,75	1,65	1,60	1,58	1,55
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,87	1,83	1,80	1,66	1,55	1,50	1,48	1,45

Annexe 2

Assuming T defined a Student random variable with n degrees of freedom, this table provides t so that: $P(T < t) = 1 - \alpha$													
Example, for $1 - \alpha = 0,975$, assuming T defines a Student random variable with 7 degrees of freedom: $P(T < 2,36) = 0,975$													
Then: $P(-2,36 < T < 2,36) = 0,95$													
as $P(-t < T < t) = 1 - \alpha \Leftrightarrow P(T < t) - P(T < -t) = 1 - \alpha \Leftrightarrow P(T < t) - [1 - P(T < t)] = 1 - \alpha \Leftrightarrow 2P(T < t) - 1 = 1 - \alpha \Leftrightarrow P(T < t) = 1 - \alpha/2$													
α	$1 - \alpha$												
	0,5	0,6	0,7	0,75	0,8	0,85	0,9	0,95	0,975	0,99	0,995	0,999	0,9995
1	0,00	0,32	0,73	1,00	1,38	1,96	3,08	6,31	12,71	31,82	63,66	318,31	636,62
2	0,00	0,29	0,62	0,82	1,06	1,39	1,89	2,92	4,30	6,96	9,92	22,33	31,60
3	0,00	0,28	0,58	0,76	0,98	1,25	1,64	2,35	3,18	4,54	5,84	10,21	12,92
4	0,00	0,27	0,57	0,74	0,94	1,19	1,53	2,13	2,78	3,75	4,60	7,17	8,61
5	0,00	0,27	0,56	0,73	0,92	1,16	1,48	2,02	2,57	3,36	4,03	5,89	6,87
6	0,00	0,26	0,55	0,72	0,91	1,13	1,44	1,94	2,45	3,14	3,71	5,21	5,96
7	0,00	0,26	0,55	0,71	0,90	1,12	1,41	1,89	2,36	3,00	3,50	4,79	5,41
8	0,00	0,26	0,55	0,71	0,89	1,11	1,40	1,86	2,31	2,90	3,36	4,50	5,04
9	0,00	0,26	0,54	0,70	0,88	1,10	1,38	1,83	2,26	2,82	3,25	4,30	4,78
10	0,00	0,26	0,54	0,70	0,88	1,09	1,37	1,81	2,23	2,76	3,17	4,14	4,59
11	0,00	0,26	0,54	0,70	0,88	1,09	1,36	1,80	2,20	2,72	3,11	4,02	4,44
12	0,00	0,26	0,54	0,70	0,87	1,08	1,36	1,78	2,18	2,68	3,05	3,93	4,32
13	0,00	0,26	0,54	0,69	0,87	1,08	1,35	1,77	2,16	2,65	3,01	3,85	4,22
14	0,00	0,26	0,54	0,69	0,87	1,08	1,35	1,76	2,14	2,62	2,98	3,79	4,14
15	0,00	0,26	0,54	0,69	0,87	1,07	1,34	1,75	2,13	2,60	2,95	3,73	4,07
16	0,00	0,26	0,54	0,69	0,86	1,07	1,34	1,75	2,12	2,58	2,92	3,69	4,01
17	0,00	0,26	0,53	0,69	0,86	1,07	1,33	1,74	2,11	2,57	2,90	3,65	3,97
18	0,00	0,26	0,53	0,69	0,86	1,07	1,33	1,73	2,10	2,55	2,88	3,61	3,92
19	0,00	0,26	0,53	0,69	0,86	1,07	1,33	1,73	2,09	2,54	2,86	3,58	3,88
20	0,00	0,26	0,53	0,69	0,86	1,06	1,33	1,72	2,09	2,53	2,85	3,55	3,85
21	0,00	0,26	0,53	0,69	0,86	1,06	1,32	1,72	2,08	2,52	2,83	3,53	3,82
22	0,00	0,26	0,53	0,69	0,86	1,06	1,32	1,72	2,07	2,51	2,82	3,50	3,79
23	0,00	0,26	0,53	0,69	0,86	1,06	1,32	1,71	2,07	2,50	2,81	3,48	3,77
24	0,00	0,26	0,53	0,68	0,86	1,06	1,32	1,71	2,06	2,49	2,80	3,47	3,75
25	0,00	0,26	0,53	0,68	0,86	1,06	1,32	1,71	2,06	2,49	2,79	3,45	3,73
26	0,00	0,26	0,53	0,68	0,86	1,06	1,31	1,71	2,06	2,48	2,78	3,43	3,71
27	0,00	0,26	0,53	0,68	0,86	1,06	1,31	1,70	2,05	2,47	2,77	3,42	3,69
28	0,00	0,26	0,53	0,68	0,85	1,06	1,31	1,70	2,05	2,47	2,76	3,41	3,67
29	0,00	0,26	0,53	0,68	0,85	1,06	1,31	1,70	2,05	2,46	2,76	3,40	3,66
30	0,00	0,26	0,53	0,68	0,85	1,05	1,31	1,70	2,04	2,46	2,75	3,39	3,65