

Prof. Mohamed Aly LOULY

Data Science et Machine Learning

1

Le modèle de la régression linéaire
avec les tests statistiques
correspondants

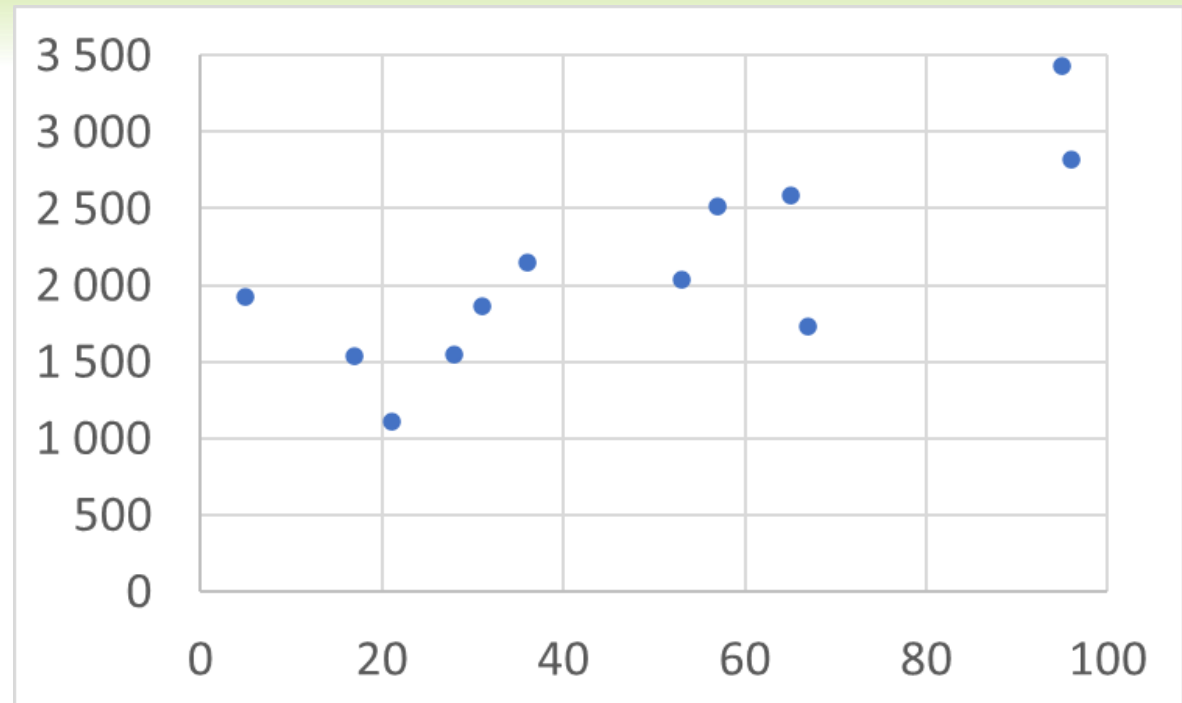
RÉGRESSION LINÉAIRE SIMPLE

Objectif : expliquer une variable quantitative dépendante Y par une variable quantitative indépendante X .

Exemple : expliquer le coût de production mensuel par les quantités fabriquées d'un produit.

DONNÉES DE LA RÉGRESSION LINÉAIRE SIMPLE

Y	X
1 733	67
2 144	36
2 824	96
1 115	21
1 922	5
2 519	57
3 428	95
1 862	31
1 551	28
2 035	53
2 589	65
1 540	17



Le coût **Y** vu comme une fonction de la quantité fabriquée **X**.

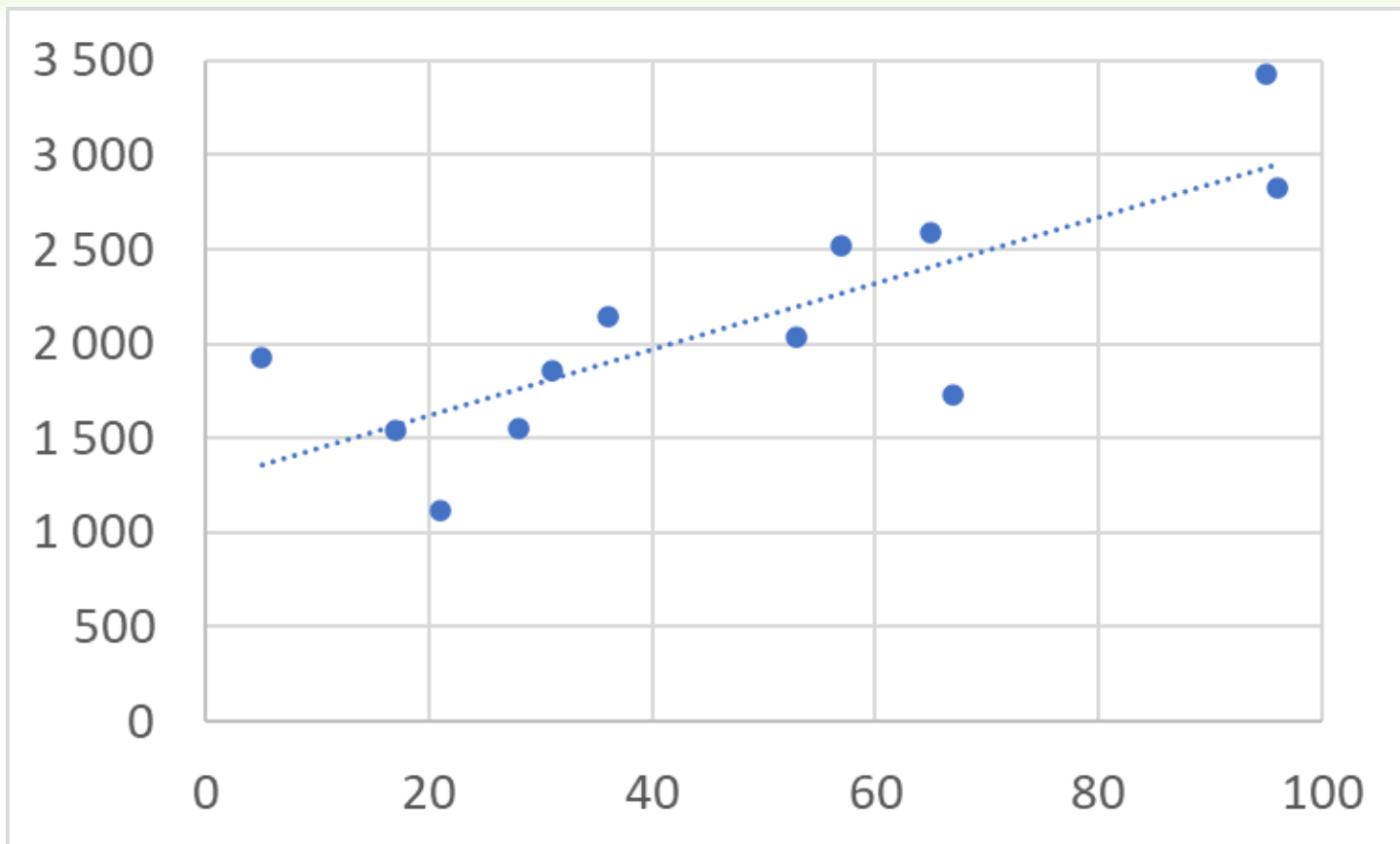
$$y_i = f(x_i)$$

LE MODÈLE THÉORIQUE DE LA RÉGRESSION LINÉAIRE SIMPLE

Nous cherchons les paramètres a et b qui permettent d'écrire Y en fonction de X selon le modèle suivant :

$$y_i = (ax_i + b) + e_i = \hat{y}_i + e_i, \quad i=1,\dots,n.$$

DROITE DE LA RÉGRESSION LINÉAIRE SIMPLE



PRINCIPE DE LA MÉTHODE DES MOINDRES CARRÉES

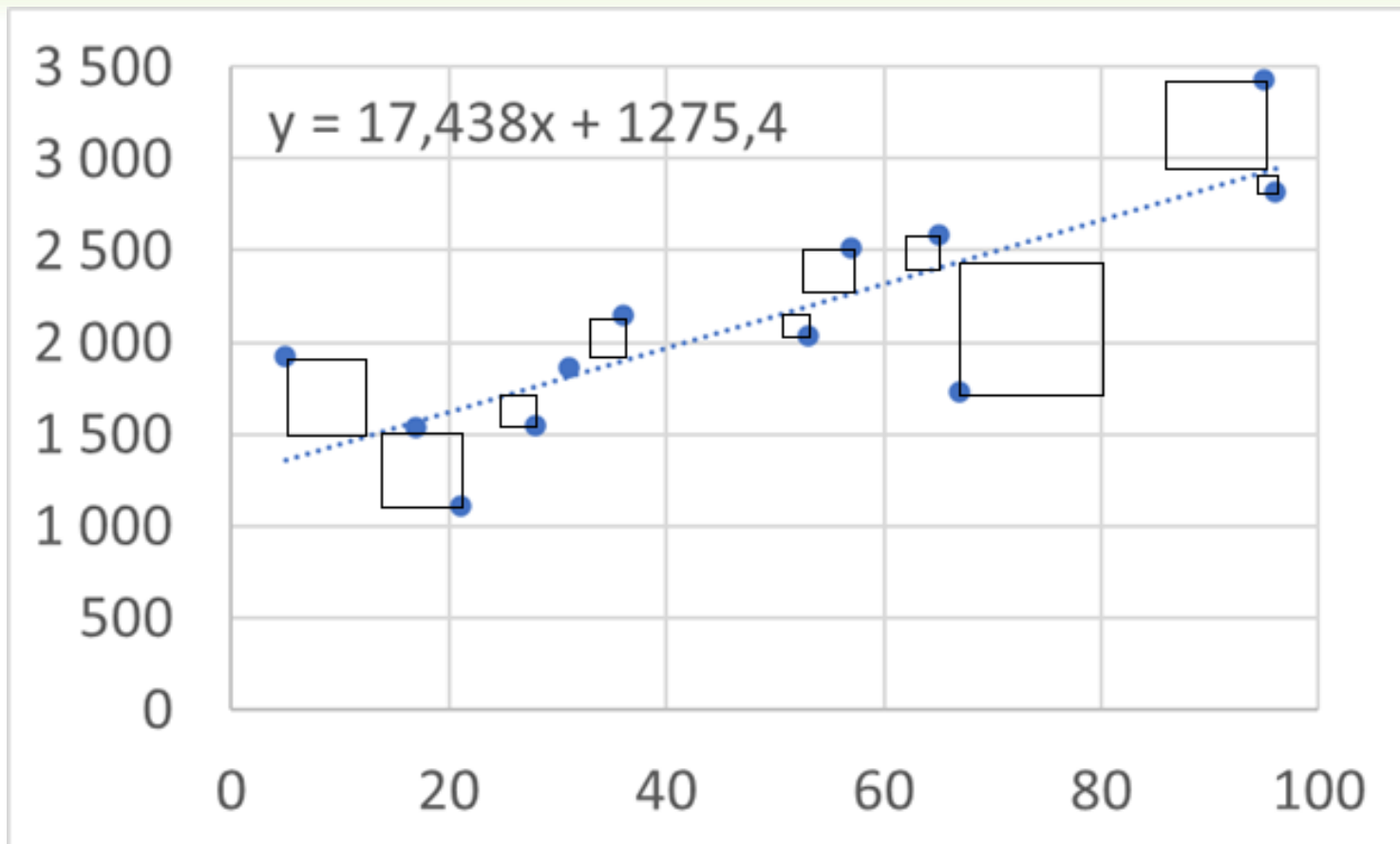
Les paramètres du modèles sont fixés de manière à minimiser la sommes des carrés résiduels :

$$\min_{a,b} \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - ax_i - b)^2$$

En supposant X de variance non nulle, on obtient :

$$a^* = \frac{Cov(X,Y)}{Var(X)} = 17,4 \quad ; \quad b^* = \bar{Y} - a^* \bar{X} = 1275,4$$

DROITE DES MOINDRES CARRÉES



L'ÉQUATION FONDAMENTALE D'ANALYSE DE LA VARIANCE

La Somme des Carrés Totaux (SCT) = La Somme des Carrés Expliqués (SCE) + La Somme des Carrés Résiduels (SCR)

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

$$4\,516\,526 = 2\,931\,158 + 1\,585\,368$$

La variabilité totale (SCT) est égale à la somme de la variabilité expliquée (SCE) et la variabilité résiduelle (SCR)

PREMIÈRES MESURES DE PERTINENCE DU MODÈLE

Coefficient de corrélation :

$$R = \cos(X, Y) = \text{cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{\overline{XY} - \bar{X} \bar{Y}}{\sigma_X \sigma_Y} = 0,81$$

Coefficient de détermination :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 0,65$$

TEST DE FISHER POUR LA PERTINENCE DU MODÈLE

Hypothèse H_0 : X ne peut pas expliquer Y : ($a=0$)

Hypothèse H_1 : X explique bien Y : ($a \neq 0$)

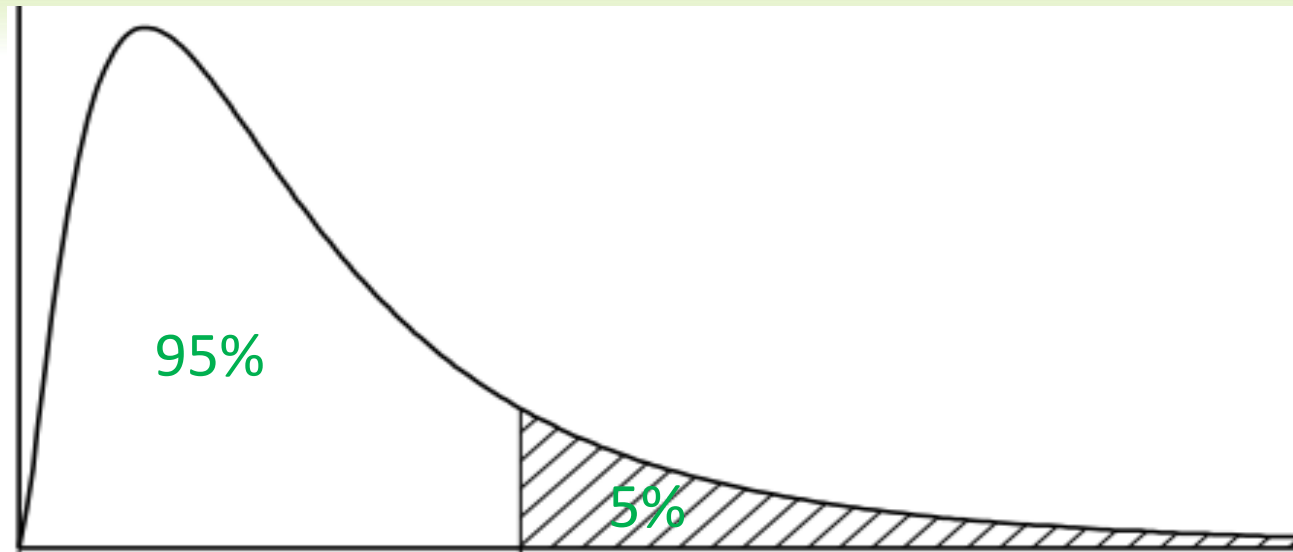
Sous l'hypothèse H_0 l'observation suivante est une réalisation de la loi de Fisher avec les degrés de liberté ($1, n-2$) :

$$F^* = \frac{VE}{VR} = \frac{SCE/1}{SCR/(n-2)} = \frac{R^2}{(1-R^2)/(n-2)}$$

Le modèle de la régression linéaire est jugé pertinent si l'observation F^* dépasse un seuil critique :

$$F^* > F_{1,n-2}^{0,05} \quad \longleftrightarrow \quad \text{Proba}(F_{1,n-2} > F^*) \leq 0,05$$

LA LOI ET LE TEST DE FICHER



Zone d'acceptation de H_0
La régression n'est pas fiable

Seuil
critique

Zone de rejet de H_0
La régression est fiable

TABLEAU D'ANALYSE DE LA VARIANCE ET TEST DE FISHER

Source	Somme des carrés	Degré de liberté	Carrés moyens	Observation	P-value	Seuil critique
Expliquée	$\sum (\hat{y}_i - \bar{y})^2$	1	$VE = \frac{SCE}{1}$	$F^* = \frac{CME}{CMR}$	$\Pr (F > F^*)$	$F_{1, n-2}^{0,05}$
Résiduelle	$\sum (y_i - \hat{y}_i)^2$	$n - 2$	$VR = \frac{SCR}{n-2}$			
Totale	$\sum (y_i - \bar{y})^2$	$n - 1$				

TABLEAU D'ANALYSE DE LA VARIANCE ET TEST DE FISHER

Source	Somme des carrés	Degré de liberté	Carrés moyens	Observation	P-value	Seuil critique
Expliquée	2 931 158	1	2 931 158	18,49	0,0016	4,96
Résiduelle	1 585 368	10	158 536,8			
Totale	4 516 526	11				

Le modèle est fiable

RÉGRESSION LINÉAIRE MULTIPLE

Objectif : expliquer une variable quantitatives Y par plusieurs variables quantitatives $X_1, X_2, \dots, X_p$.

Y	X1	X2
1 733	67	20
2 144	36	80
2 824	96	36
1 115	21	29
1 922	5	99
2 519	57	50
3 428	95	81
1 862	31	42
1 551	28	53
2 035	53	72
2 589	65	95
1 540	17	31

Exemple : expliquer le coût de production par les quantités fabriquées de 2 produits.

LE MODÈLE THÉORIQUE DE LA RÉGRESSION LINÉAIRE

Une variable quantitative Y est expliquée par plusieurs variables quantitatives $X_1, X_2, \dots, X_p$ selon le modèle suivant :

$$y_i = (a_0 + a_1 x_{i1} + \dots + a_p x_{ip}) + \varepsilon_i = \hat{y}_i + \varepsilon_i, \quad i=1, \dots, n.$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$Y = Xa + E$$

ESTIMATEUR DES MOINDRE CARRÉES DES COEFFICIENTS

Les paramètres du modèles sont fixés de manière à minimiser la sommes des carrés résiduels :

$$\min_{a_0, \dots, a_p} \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (y_i - (a_0 + a_1 x_{i1} + \dots + a_p x_{ip}))^2$$

En supposant $X^T X$ inversible, on obtient les paramètres suivants :

$$a^* = (X^T X)^{-1} X^T Y$$

Les paramètres permettent de prédire la variable expliquée :

$$\hat{Y} = X a^*$$

EXEMPLE DE LA PRÉVISION DU COÛT DE PRODUCTION

 X

1	67	20
1	36	80
1	96	36
1	21	29
1	5	99
1	57	50
1	95	81
1	31	42
1	28	53
1	53	72
1	65	95
1	17	31

 X^T

1	1	1	1	1	1	1	1	1	1	1	1
67	36	96	21	5	57	95	31	28	53	65	17
20	80	36	29	99	50	81	42	53	72	95	31

EXEMPLE DE LA PRÉVISION DU COÛT DE PRODUCTION

$$X^T X$$

12	571	688
571	36809	32629
688	32629	47542

$$(X^T X)^{-1}$$

0,731887	-0,005017	-0,007148
-0,005017	0,000104	0,000001
-0,007148	0,000001	0,000124

$$\hat{Y} = Xa^*$$

2039
2149
2723
1330
1811
2191
3196
1647
1714
2360
2822
1281

$$X^T Y$$

25262
1370137
1534697

$$a^* = (X^T X)^{-1} X^T Y$$

644,69
17,56
10,90

L'ÉQUATION FONDAMENTALE D'ANALYSE DE LA VARIANCE

La Somme des Carrés Totaux (SCT) = La Somme des Carrés Expliqués (SCE) + La Somme des Carrés Résiduels (SCR)

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

La variabilité totale (SCT) est égale à la somme de la variabilité expliquée (SCE) et la variabilité résiduelle (SCR)

PREMIÈRES MESURES DE PERTINENCE DU MODÈLE

Coefficient de corrélation :

$$R = \text{cor}(Y, \hat{Y}) = 0,93$$

Coefficient de détermination :

$$R^2 = \frac{SCE}{SCT} = 1 - \frac{SCR}{SCT} = 0,86$$

TEST DE FISHER POUR LA PERTINENCE DU MODÈLE

$H_0 : a_1 = a_2 = \dots = a_p = 0$

H_1 : il existe au moins un coefficient non nul

Sous l'hypothèse H_0 l'observation suivante est une réalisation de la loi de Fisher avec les degrés de liberté $(p, n-p-1)$:

$$F^* = \frac{VE}{VR} = \frac{SCE/p}{SCR/(n-p-1)} = \frac{R^2/p}{(1-R^2)/(n-p-1)}$$

Le modèle de la régression linéaire est jugé pertinent si l'observation F^* dépasse un seuil critique :

$$F^* > F_{p,n-p-1}^{0,05} \iff \text{Proba}(F_{p,n-p-1} > F^*) \leq 0,05$$

TABLEAU D'ANALYSE DE LA VARIANCE ET TEST DE FISHER

Source	Somme des carrés	Degré de liberté	Carrés moyens	Observation	P-value	Seuil critique
Expliquée	$\sum (\hat{y}_i - \bar{y})^2$	p	$VE = \frac{SCE}{p}$	$F^* = \frac{CME}{CMR}$	$\Pr (F > F^*)$	$F_{p, n-p-1}^{0,05}$
Résiduelle	$\sum (y_i - \hat{y}_i)^2$	n - p - 1	$VR = \frac{SCR}{n-p-1}$			
Totale	$\sum (y_i - \bar{y})^2$	n - 1				

TABLEAU D'ANALYSE DE LA VARIANCE ET TEST DE FISHER

Source	Somme des carrés	Degré de liberté	Carrés moyens	Observation	P-value	Seuil critique
Expliquée	3 892 784	2	1 946 392	28,08	0,0001	4,26
Résiduelle	623 742	9	69 305			
Totale	4 516 526	11				

Le modèle
est fiable

TEST DE STUDENT POUR LA PERTINENCE DE CHAQUE VARIABLE

H_0 : X_i ne participe à l'explication de Y : ($a_i = 0$)

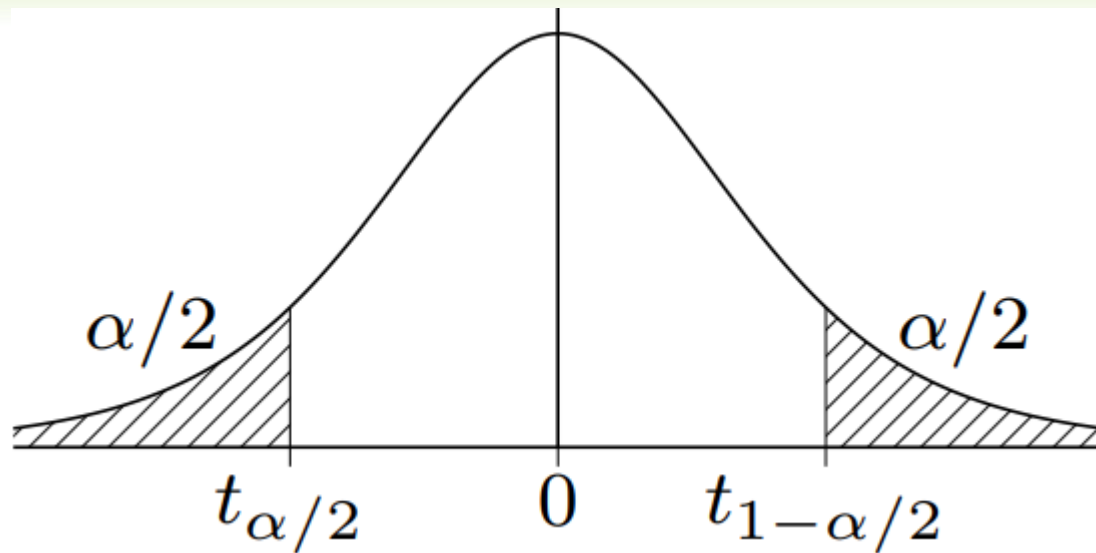
H_1 : X_i participe bien à l'explication de Y : ($a_i \neq 0$)

Sous l'hypothèse H_0 l'observation suivante est une réalisation de la loi de Student de degré de liberté $n-p-1$:

$$t_i^* = \frac{a_i^*}{\sigma_{a_i^*}}$$

Où la variance empirique de a_i^* est le i ème coefficient de la diagonale de la matrice des variances :

$$\text{Variances} = VR (X^T X)^{-1}$$



Zone d'acceptation de H_0 : Proba $(|T| \leq t_{n-p-1}^{0,975}) = 0,95$

Zone de rejet de H_0 : Proba $(|T| \geq t_{n-p-1}^{0,975}) = 0,05$

LES OBSERVATIONS DE STUDENT

$$\text{Variances} = VR (X^T X)^{-1}$$

50723,2	-347,7	-495,4
-347,7	7,2	0,1
-495,4	0,1	8,6

$$t_i^*$$

2,86		
	6,55	
		3,72

Toutes les observations dépassent le seuil critique $t_{10}^{0,975} = 2,2281$.

Chaque variable X_i contribue donc bien à l'explication de Y .