ENI CARTHAGE
المدرســـة الوطنيــة للمهندسيــن بقرطــاج
Ecole Nationale d'Ingénieurs de Carthage

Année Universitaire 2020-2021

Analyse de données

Applications Modèle Linéaire

Modèle linéaire

En étudiant le comportement simultané de deux variables **X** et **Y**, on pourrait trouver une certaine variation simultanée dans les valeurs que peuvent prendre ces deux variables et ce dans une certaine proportion et même dans deux sens opposés.

Citons par exemple le cas où **X** est une variable qui décrit le facteur travail dans une entreprise et **Y** est une variable relative à la production de l'entreprise. On constate que plus la valeur de **X** s'élève, celle de **Y** s'élève aussi. Ceci nous ramène à prédire qu'il pourrait y avoir une relation entre **X** et **Y**. On parle alors de régression.

Parmi les objectifs principaux d'une analyse de la régression, on peut citer les deux points suivants :

1. comprendre comment et dans quelles mesures une variable **X** influence la variable dépendante **Y**.
2. développer un modèle pour prévoir des valeurs de **Y** futures à partir de celles que pourrait prendre la variable **X**.

Dans ce qui suit nous allons nous intéresser au modèle de régression linéaire simple. C'est-à-dire le cas où la variable **Y** est en relation linéaire avec une variable **X**. Autrement, **Y** peut s'écrire sous la forme d'une constante donnée à laquelle on ajoute un coefficient multiplié par **X**.

I - Présentation et hypothèses du modèle :

I.1. Présentation du modèle :

On cherche à établir s'il y a un lien linéaire entre deux variables **X** et **Y**. Le modèle est :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Dans ce modèle, appelé modèle de régression linéaire simple, les composantes ont la signification suivante :

- **Y** est la variable dépendante (expliquée ou endogène) à caractère aléatoire.
- **X** est la variable indépendante (explicative ou exogène) mesurée sans erreur ou fixée à des niveaux arbitraires.

• $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$ sont les coefficients de régression théoriques du modèle que l'on devra estimer à l'aide d'un échantillon. Ce sont les paramètres du modèle.

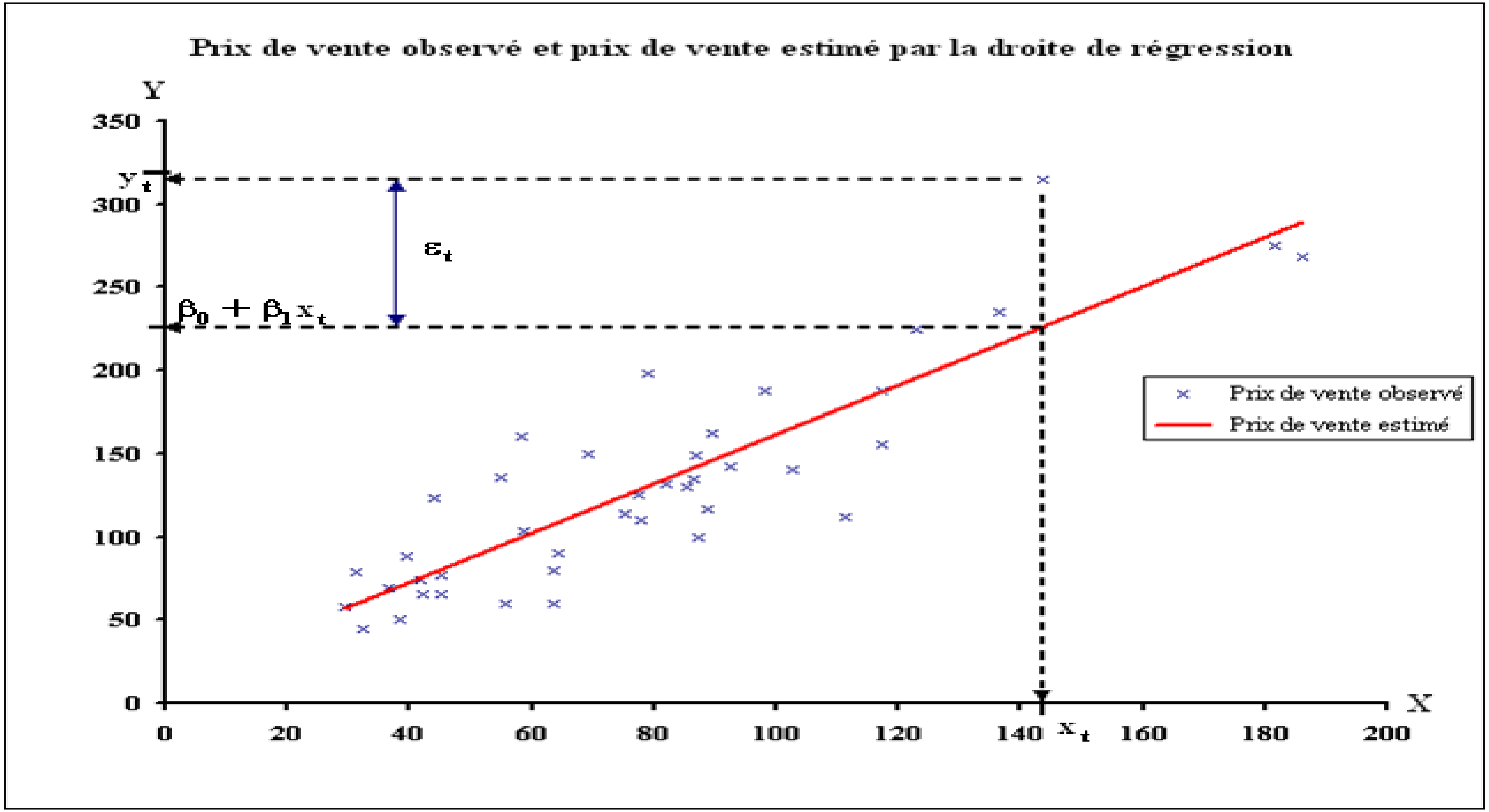
• $\boldsymbol{\varepsilon}$ représente l'erreur théorique aléatoire associée à la variable dépendante **Y** : c'est une variable aléatoire qui prend en compte l'existence éventuelle d'autres influences que celle de **X** sur **Y**.

Exemples :

- *X peut être le temps et Y une grandeur mesurée à différentes dates*

- *Y peut être la différence de potentiel mesurée aux bornes d'une résistance pour différentes valeurs de l'intensité X du courant*

Exemple : Soit un exemple dans lequel nous voudrions étudier l'existence d'une relation linéaire entre le prix de vente d'une maison et son estimation municipale et en analysant le nuage de points obtenu, on peut alors ajouter une droite de tendance qui illustre cette relation linéaire. On obtient alors le graphique suivant :



I.2. Hypothèses du modèle :

Pour que le modèle soit bien défini, outre l'hypothèse de linéarité, il faut ajouter un certain nombre d'autres hypothèses :

Pour les n couples $(x_t ; y_t)$ de valeurs observées dans la population, nous avons la relation suivante :

Les erreurs théoriques (ε_t) devront satisfaire les hypothèses suivantes :

- ***H_1*** : Les erreurs ont toutes une moyenne nulle.

$$E(\varepsilon_t) = 0 \quad \forall t \in \{1;2;...;n\}$$

- ***H_2*** : L'homoscédasticité des erreurs.

$$V(\varepsilon_t) = \sigma^2 \quad \forall t \in \{1;2;...;n\}$$

- ***H_3*** : Les erreurs sont indépendantes entre elles (les erreurs de deux observations différentes ne sont pas corrélées) et forment une suite de variables aléatoires indépendantes.

$$Cov(\varepsilon_t, \varepsilon_{t'}) = 0 \quad \forall t \neq t'$$

- ***H_4*** : Les erreurs sont indépendantes et identiquement distribuées (i.i.d.) selon la loi normale d'espérance nulle et de variance .

$$\varepsilon_t \sim N(0;\sigma^2) \quad \forall t \in \{1;2;...;n\}$$

II – Les paramètres du modèle :

Les paramètres inconnus du modèle sont de deux sortes : Il y a les coefficients $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$ d'une part et la variance des erreurs $\boldsymbol{\sigma^2}$ d'autre part.

Dans ce qui suit, on va estimer respectivement ces paramètres par la méthode des moindres carrés ordinaires (MCO) qui s'avère appropriée pour l'obtention des estimateurs $\hat{\beta}_0$ et $\hat{\beta}_1$ respectifs des paramètres $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$.

II.1. Estimation des paramètres β_0 et β_1 :

Le principe consiste à calculer le terme d'erreur qui est l'écart entre y_t observé et y_t estimé. On aura alors :

$$\boldsymbol{\varepsilon_t = y_t - (\beta_0 + \beta_1 X)}$$

La méthode des moindres carrés ordinaires consiste à minimiser, par rapport au paramètres inconnus du modèle, la somme des carrés des écarts (ou des résidus) appelée **SCR** et qui est égale à :

$$SCR = \sum_{t=1}^{n} \varepsilon_t^2$$

Nous allons alors minimiser l'expression **SCR** par rapport à $\boldsymbol{\beta_0}$ et $\boldsymbol{\beta_1}$. Les conditions de minimisation sont les suivantes :

$$Min\ \ SCR = \underset{\beta_0 ; \beta_1}{Min} \sum_{t=1}^{n} \varepsilon_t^2$$

Conditions de premier ordre : $\begin{cases} \dfrac{\partial SCR}{\partial \beta_0} = 0 \\ \dfrac{\partial SCR}{\partial \beta_1} = 0 \end{cases}$ *Condition de deuxième ordre :* $\dfrac{\partial^2 SCR}{\partial \beta_0 \partial \beta_1} \geq 0$

On a alors :

$$\boxed{\begin{cases} \hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x} \\ \hat{\beta}_1 = \dfrac{\operatorname{cov}(X,Y)}{\operatorname{var}(X)} \end{cases}}$$

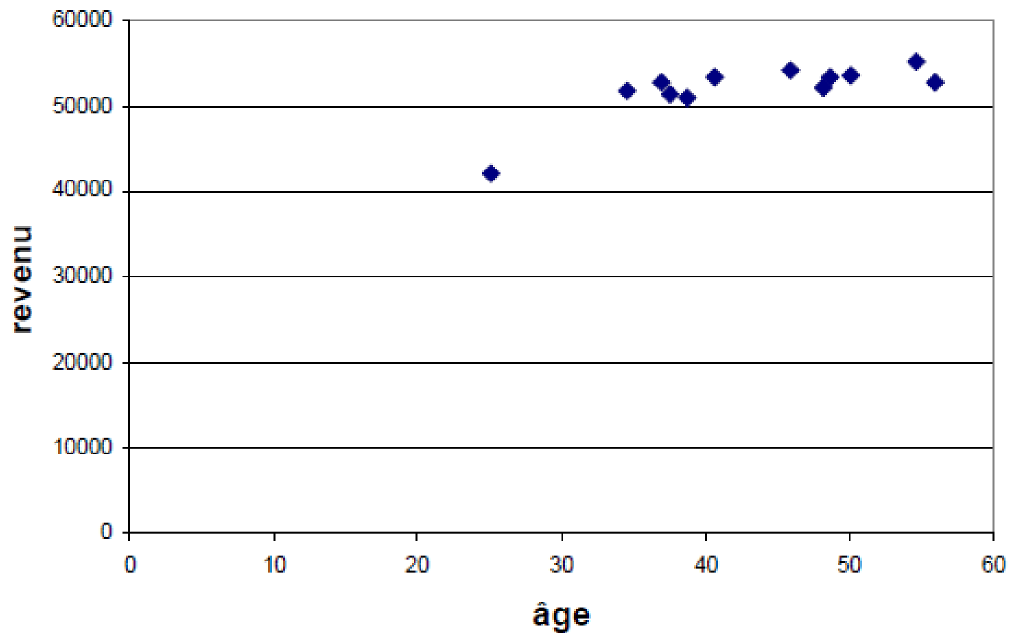
Exemple :

Le syndic s'intéresse au rapport entre l'âge et le revenu des résidents d'une ville. Il sélectionne un échantillon aléatoire simple de taille n = 12.   .

données de l'échantillon :

ind.	revenu	âge
1	52125.0	48.1
2	50955.9	38.7
3	53382.9	48.6
4	51286.9	37.5
5	55243.6	54.7
6	53384.7	40.7
7	53488.2	50.1
8	54134.1	45.9
9	52706.4	55.9
10	42144.3	25.1
11	52665.2	36.9
12	51656.7	34.5
Moyenne	*51931.2*	*43.1*
Ecart type	*3314.9*	*9.1*

nuage de points :

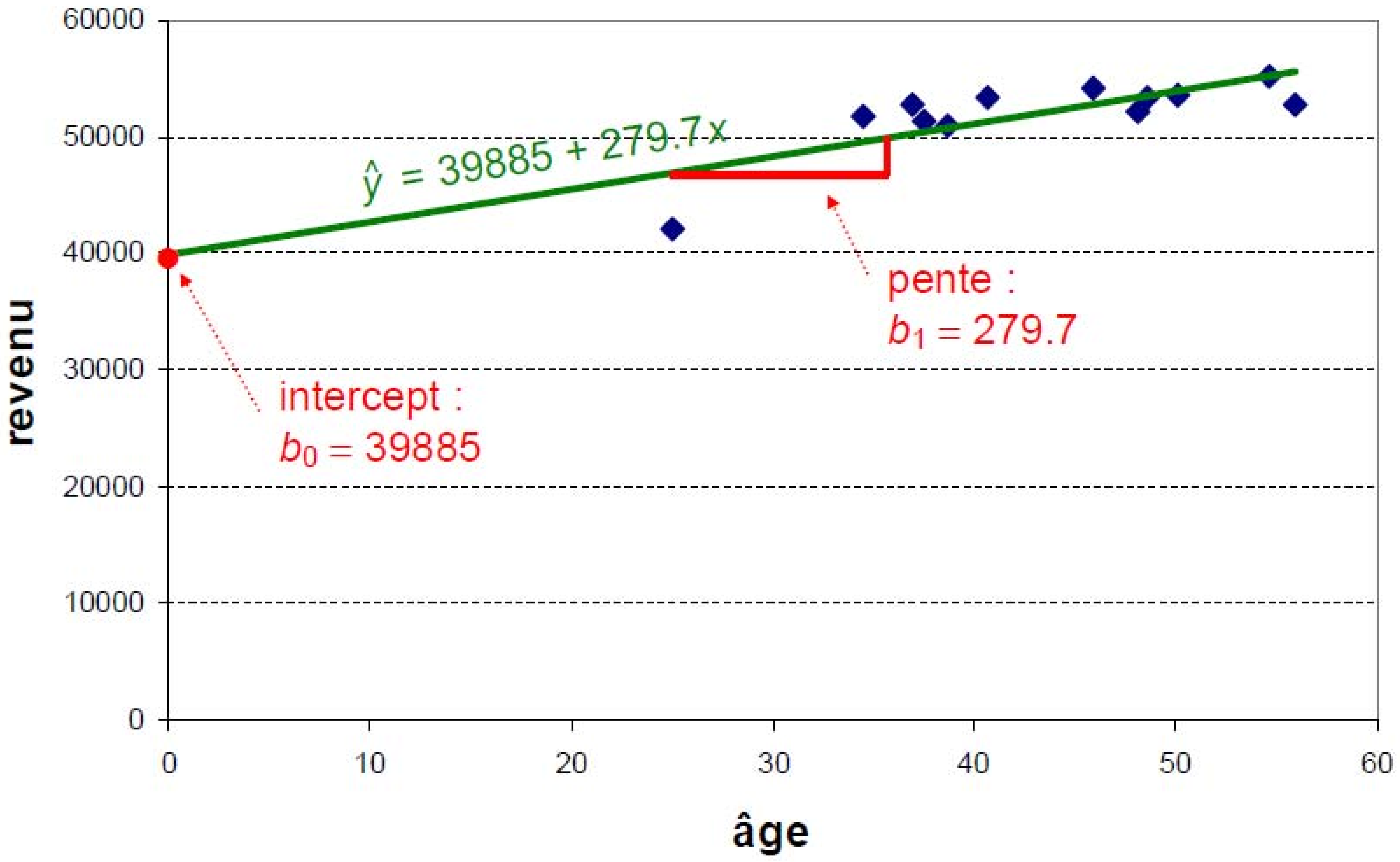


Exemple (suite) :

i (ind.)	y_i (revenu)	x_i (âge)	$y_i - \bar{y}$	$x_i - \bar{x}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	52125.0	48.1	193.9	5.0	978.6	25.5
2	50955.9	38.7	-975.3	-4.4	4245.4	18.9
3	53382.9	48.6	1451.7	5.6	8061.1	30.8
4	51286.9	37.5	-644.3	-5.5	3570.3	30.7
5	55243.6	54.7	3312.5	11.6	38434.3	134.6
6	53384.7	40.7	1453.5	-2.4	-3481.4	5.7
7	53488.2	50.1	1557.1	7.1	10982.0	49.7
8	54134.1	45.9	2202.9	2.9	6281.9	8.1
9	52706.4	55.9	775.2	12.9	9975.6	165.6
10	42144.3	25.1	-9786.9	-18.0	176033.4	323.5
11	52665.2	36.9	734.1	-6.1	-4503.3	37.6
12	51656.7	34.5	-274.5	-8.6	2350.7	73.3
Moyenne	***51931.2***	***43.1***	***0***	***0***	***21077.4***	***75.4***
Somme	***623174.0***	***516.8***	***0***	***0***	***252928.4***	***904.3***

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{252928.4}{904.3} = \underline{\underline{279.7}}$$

$$\hat{\beta}_0 = 51931.2 - 279.7 * 43.1 \doteq \underline{\underline{39885}}$$



60000
50000
40000
30000
20000
10000
0
revenu
$\hat{y} = 39885 + 279.7x$
pente :
$b_1 = 279.7$
intercept :
$b_0 = 39885$
0
10
20
30
40
50
60
âge

Calcul SCR :

i (ind.)	y_i (revenu)	x_i (âge)	$\hat{y}_i = 39885 + 279.7 * x_i$	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$
1	52125.0	48.1	53343.0	-1218.0	1483550.6
2	50955.9	38.7	50713.7	242.2	58665.3
3	53382.9	48.6	53484.3	-101.4	10274.8
4	51286.9	37.5	50381.1	905.8	820405.6
5	55243.6	54.7	55176.5	67.1	4507.4
6	53384.7	40.7	51261.3	2123.5	4509068.6
7	53488.2	50.1	53903.9	-415.6	172735.6
8	54134.1	45.9	52728.7	1405.4	1975015.2
9	52706.4	55.9	55530.2	-2823.8	7973726.7
10	42144.3	25.1	46900.3	-4756.1	22620189.0
11	52665.2	36.9	50215.3	2450.0	6002285.9
12	51656.7	34.5	49535.7	2121.0	4498484.4
Moyenne	***51931.2***	***43.1***	***51931.2***	***0***	***4177409.1***
Somme	***623174.0***	***516.8***	***623174.0***	***0***	***50128909.0***

II.2. Les lois de probabilité des estimateurs β_0 , β_1 :

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{t=1}^{n} x_t \varepsilon_t}{\sum_{t=1}^{n} x_t (x_t - \bar{x})}$$

Donc $\boldsymbol{\beta_1}$, peut s'écrire sous la forme d'une combinaison linéaire des termes $\boldsymbol{\varepsilon_t}$ qui suivent chacun une loi normale $\mathbf{N(0;\sigma^2)}$

La combinaison linéaire de lois normales est une loi normale. Par conséquent, $\boldsymbol{\hat{\beta}_1}$ suit une loi normale d'espérance $\mathbf{E(\hat{\beta}_1)}$ et de variance $\mathbf{V(\hat{\beta}_1)}$.

Donc

$$\hat{\beta}_1 \sim N\left(\beta_1 ; V(\hat{\beta}_1)\right) \quad \Rightarrow \quad \frac{\hat{\beta}_1 - \beta_1}{\sqrt{V(\hat{\beta}_1)}} \sim N(0;1)$$

Par Analogie, on obtient pour la loi de $\hat{\beta}_0$

$$\hat{\beta}_0 \sim N\left(\beta_0 ;\ V(\hat{\beta}_0)\right) \quad \Rightarrow \quad \frac{\hat{\beta}_0 - \beta_0}{\sqrt{V(\hat{\beta}_0)}} \sim N(0;1)$$

III – Tests de significativité des paramètres :

Après avoir estimé les paramètres du modèle par leurs estimateurs respectifs, on voudrait savoir si le modèle qu'on a établi peut admettre la constante ou non. D'autre part, on voudrait connaître si la variable **X** explique la variable **Y** ou non.

Autrement, dans notre modèle, nous avons postulé qu'il y a une relation de cause à effet entre les variables X et Y. D'après l'échantillon composé de **n** observations qu'on a, nous allons confirmer le modèle ou bien l'infirmer. Pour cela, nous allons procéder à des tests de significativité des paramètres du modèle.

III.1 – Tests de significativité du paramètre β_0 :

Nous allons chercher la loi de probabilité à utiliser pour effectuer le test de significativité. On connaît que

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{V(\hat{\beta}_0)}} \sim N(0\,;1)$$

Comme une loi de **Student** à **m** degrés de liberté est égal au rapport d'une loi **normale centrée réduite** par la racine carrée d'une loi de **Khi deux** à **m** degrés de liberté divisée par ce nombre de degrés de liberté, alors, on obtient que :

$$\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{V}(\hat{\beta}_0)}} \sim T_{(n-2)}$$

En conclusion, nous allons utiliser la loi de **Student** à **(n-2)** degrés de liberté pour effectuer le test de significativité.

On pose une hypothèse nulle $\mathbf{H_0}$ selon laquelle la valeur du paramètre est nulle (le modèle n'admet pas de constante) contre une hypothèse alternative $\mathbf{H_1}$ selon laquelle le paramètre est significatif (différent de 0). Les hypothèses sont présentées ainsi :

$$\begin{cases} \mathbf{H_0 : \beta_0 = 0} \\ \mathbf{H_1 : \beta_0 \neq 0} \end{cases}$$

Nous allons comparer le terme $\mathbf{t_c}$ (la statistique **t** calculée) à la valeur $\mathbf{t_{\alpha/2}}$ (la valeur critique qui est extraite à partir de la table de la loi de **Student** à **(n-2)** degrés de liberté).

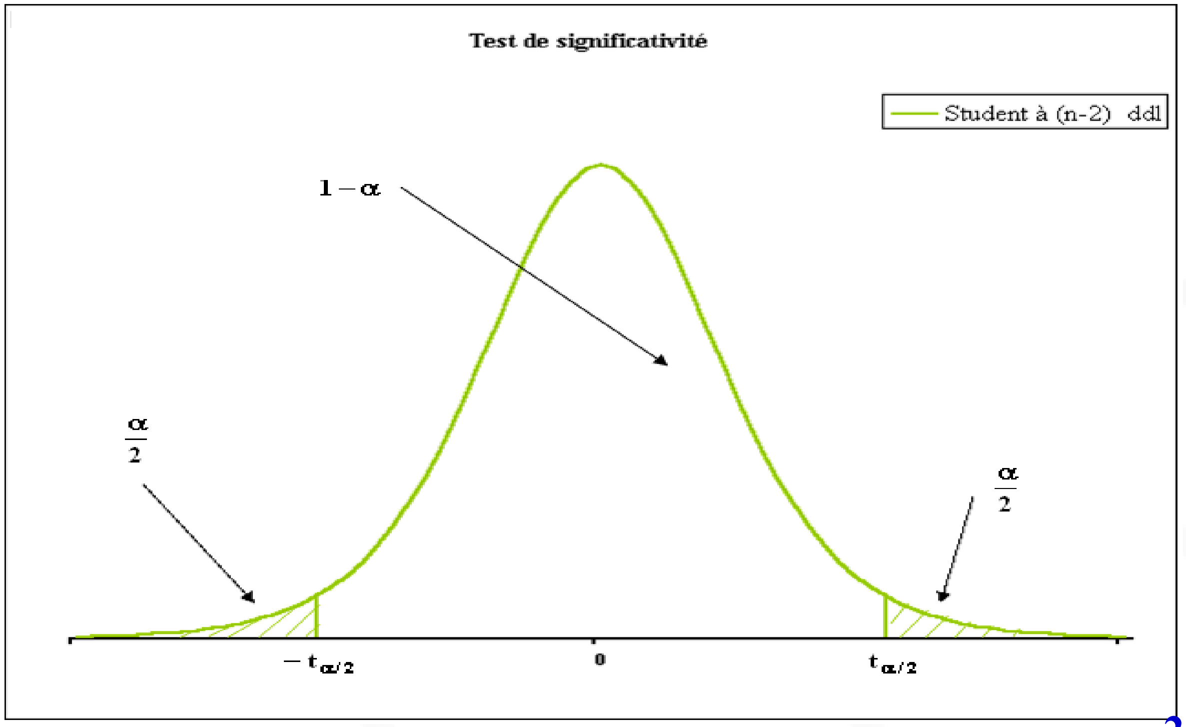
$$t_c = \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{V}(\hat{\beta}_0)}}$$

Sous l'hypothèse $\mathbf{H_0}$, on a $\beta_0 = 0$ donc $t_c = \dfrac{\hat{\beta}_0}{\sqrt{\hat{V}(\hat{\beta}_0)}}$

Si $|t_c| \langle t_{\alpha/2}$ c'est-à-dire $-t_{\alpha/2} \langle t_c \langle t_{\alpha/2}$, nous allons accepter l'hypothèse **H_0** qui stipule que la constante $\beta_0 = 0$ avec un risque de se tromper égal à α (pour cela α doit être faible et généralement égale à 0.05).

Si $|t_c| \rangle t_{\alpha/2}$ c'est-à-dire $t_c \rangle t_{\alpha/2}$ ou $t_c \langle -t_{\alpha/2}$, nous allons rejeter l'hypothèse **H_0** et accepter l'hypothèse alternative **H_1** selon laquelle la constante β_0 est significative.

On peut représenter le test par le schéma suivant où la zone hachurée représente la zone de rejet alors que la zone vide représente la zone d'acceptation du test



Intervalle de confiance pour le paramètre β_0

Si on accepte l'hypothèse $\mathbf{H_0}$ alors on aura $p\left(|t_c| \geq t_{\alpha/2}\right) = \alpha$

$$p\left(|t_c| \langle t_{\alpha/2}\right) = 1 - \alpha \rightarrow p\left(\left|\frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{V}(\hat{\beta}_0)}}\right| \leq t_{\alpha/2}\right) = 1 - \alpha$$

$$-t_{\alpha/2} \leq \frac{\hat{\beta}_0 - \beta_0}{\sqrt{\hat{V}(\hat{\beta}_0)}} \leq t_{\alpha/2}$$ Avec un niveau de confiance égale à **1- α**

$$\hat{\beta}_0 - t_{\alpha/2}\sqrt{\hat{V}(\hat{\beta}_0)} \langle \beta_0 \langle \hat{\beta}_0 + t_{\alpha/2}\sqrt{\hat{V}(\hat{\beta}_0)}$$

L'intervalle de confiance de β_0 « IC(β_0) » à un niveau de confiance égale à **1- α**

$$\boldsymbol{\beta_0} \in \left[\hat{\boldsymbol{\beta}}_0 - \mathbf{t}_{\alpha/2}\sqrt{\hat{V}(\hat{\beta}_0)}\ ;\ \hat{\boldsymbol{\beta}}_0 + \mathbf{t}_{\alpha/2}\sqrt{\hat{V}(\hat{\beta}_0)}\right] = IC(\boldsymbol{\beta_0})$$

III.2 – Tests de significativité du paramètre β_1 :

Nous allons tester si la variable X influence la variable Y ou non. Autrement, est-ce que le paramètre est significatif ou bien il peut être assimilé à 0.

Par analogie, nous allons utiliser la statistique suivante :

$$t_c = \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\hat{V}(\hat{\beta}_1)}} \sim T_{(n-2)}$$

On pose une hypothèse nulle $\mathbf{H_0}$ selon laquelle la valeur du paramètre est nulle (la variable explicative n'est pas significative) contre une hypothèse alternative $\mathbf{H_1}$ selon laquelle le paramètre est significatif (différent de 0). Les hypothèses sont présentées ainsi :

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

La règle de décision :

Nous allons comparer le terme $\mathbf{t_c}$ (la statistique **t** calculée) sous H_0 à la valeur $\mathbf{t_{\alpha/2}}$ (la valeur critique qui est extraite à partir de la table de la loi de **Student** à **(n-2)** degrés de liberté).

Si $|t_c| \langle t_{\alpha/2}$ c'est-à-dire $-t_{\alpha/2} \langle t_c \langle t_{\alpha/2}$, nous allons accepter l'hypothèse $\mathbf{H_0}$ qui stipule que la constante $\beta_1 = 0$ avec un risque de se tromper égal à α (pour cela α doit être faible et généralement égale à 0.05).

Si $|t_c| \rangle t_{\alpha/2}$ c'est-à-dire $t_c \rangle t_{\alpha/2}$ ou $t_c \langle -t_{\alpha/2}$, nous allons rejeter l'hypothèse $\mathbf{H_0}$ et accepter l'hypothèse alternative $\mathbf{H_1}$ selon laquelle la constante β_1 est significative.

Intervalle de confiance pour le paramètre :

Par analogie, on obtient l'intervalle de confiance de β_1 pour à un niveau **1-α.**

$$\boldsymbol{\beta_1} \in \left[\hat{\boldsymbol{\beta}}_1 - \mathbf{t}_{\alpha/2} \sqrt{\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_1)} \ ; \hat{\boldsymbol{\beta}}_1 + \mathbf{t}_{\alpha/2} \sqrt{\hat{\mathbf{V}}(\hat{\boldsymbol{\beta}}_1)} \right] = \mathrm{IC}(\boldsymbol{\beta_1})$$

Une autre manière, pour procéder au test serait la suivante

Si $\mathbf{0} \in \mathbf{IC}(\boldsymbol{\beta}_1)$, alors on accepte l'hypothèse selon laquelle $\boldsymbol{\beta}_1 = \mathbf{0}$
Si $\mathbf{0} \notin \mathbf{IC}(\boldsymbol{\beta}_1)$, alors on rejette l'hypothèse selon laquelle $\boldsymbol{\beta}_1 = \mathbf{0}$

IV – Validation du modèle : coefficient de détermination R²

On note l'expression suivante : $\mathbf{R^2 = \frac{SCE}{SCT}}$ Avec

$$SCT = \sum_{t=1}^{n} (y_t - \bar{y})^2 = m_{YY} \qquad SCE = \sum_{t=1}^{n} (\hat{y}_t - \bar{y})^2 = \hat{\beta}_1^2 m_{XX} \qquad SCR = \sum_{t=1}^{n} \hat{\varepsilon}_t^{\,2} = m_{YY} - \hat{\beta}_1^2 m_{XX}$$

Décomposition de la variance

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

$$SCT = SCR + SCE$$

SCT *: somme des carrés totaux*
SCE *: somme des carrés expliqués par le modèle*
SCR *: somme des carrés résiduels, non expliqués par le modèle*

IV – Validation du modèle : coefficient de détermination R²

Coefficient de détermination.
Exprime la part de variabilité de Y expliquée par le modèle.
$R^2 \rightarrow 1$, le modèle est excellent
$R^2 \rightarrow 0$, le modèle ne sert à rien

$$R^2 = \frac{SCE}{SCT}$$

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

SCE représente la variation expliquée.
SCR représente la variation inexpliquée due aux variables omises dans le modèle.

Si $R^2=0{,}9$; on dit que 90% de la variation de **X** est expliquée par la variation de Y **.**

Si $R^2=0{,}1$; la variation de **X** contribue à hauteur de 10% dans l'explication de la variation de **Y.** Par conséquent, la variable explicative ne suffit pas à elle seule à expliquer la variable expliquée **.** On doit dans ce cas introduire d'autres variables dans le modèle sans pour autant rejeter automatiquement la variable X. Ce qu'on appelle modèle linéaire multiple

V – Modèles dérivés et interprétation des coefficients

Modèle linéaire

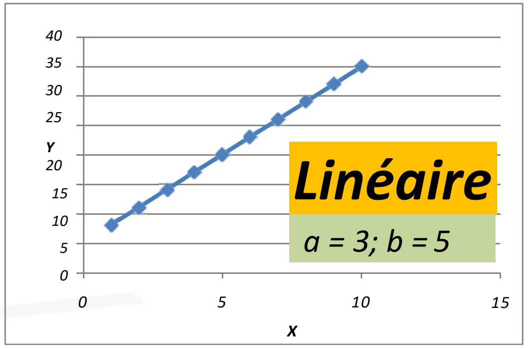
Lecture de la pente

$$Y = aX + b$$

Ex. ventes = -12 * prix + 1000

→ Lecture en niveau : si prix = 10 euros alors ventes = 980 unités

→ Lecture en termes d'évolution : si prix augmente de 1 euro , les ventes vont diminuer de 12 unités.



$$\Rightarrow \quad a = \frac{dy}{dx}$$

La variation de Y est proportionnelle à la variation de X

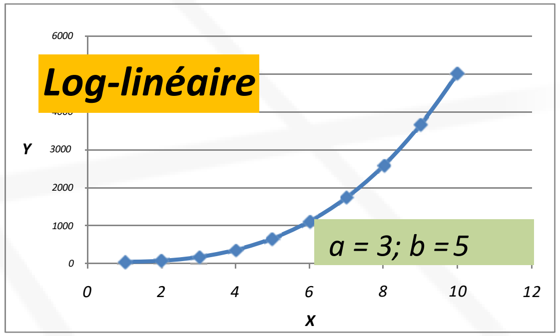
Avantages

→ Simplicité

→ Utilisé dans une première approche

→ Estimation directe des paramètres par la méthode des MCO

Modèle log-linéaire : $Y = bX^a$



$$\Rightarrow \quad a = \frac{dy/y}{dx/x}$$

Le taux de variation de Y est proportionnelle au taux de variation de X

Avantages

→ Modèle à élasticité constante : favori des économistes

→ Ex. emploi = f(production), demande = f(prix)

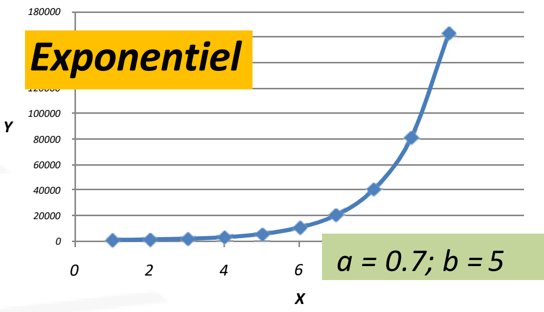
→ Linéarisation : ln(y) = a ln(x) + ln(b)

V – Modèles dérivés et interprétation des coefficients

Modèle exponentiel (géométrique)

$$Y = e^{aX+b}$$

Le taux de variation de Y est proportionnelle à la variation de X



$$\Rightarrow \quad a = \frac{dy/y}{dx}$$

Avantages

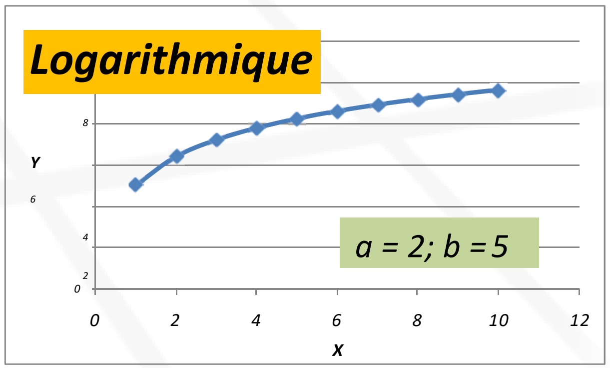
- → ***Surtout utilisé quand x = temps, ainsi dx= 1***
- → ***Dans ce cas, la croissance (décroissance) de Y est constante dans le temps***
- → ***Ce type d'évolution (croissance exponentielle) ne dure pas longtemps***
- → ***Linéarisation : ln(y) = a x + ln(b)***

Modèle logarithmique

$$Y = a\ln(X) + b$$

$$\Rightarrow \quad a = \frac{dy}{dx/x}$$

La variation de Y est proportionnelle au taux de variation de X



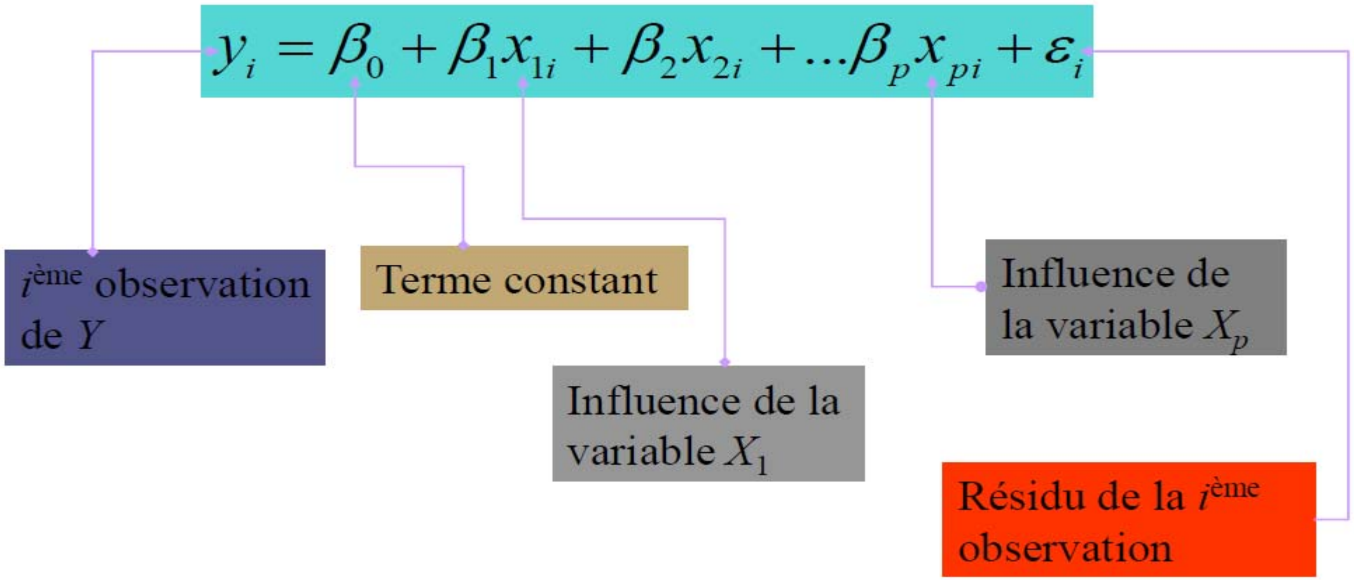
Avantages

- → ***Archétype de la croissance (décroissance) qui s'épuise***
- → ***Ex. salaire = f(ancienneté) ; vente = f(publicité)***

Modèle linéaire multiple

Estimer la relation entre une variable dépendante(Y) quantitative et plusieurs variables indépendantes (X1,X2, …)

•Equation de régression multiple : Cette équation précise la façon dont la variable dépendante est reliée aux variables explicatives :



Ecriture matricielle du modèle

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \cdots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$y = X\beta + \varepsilon$$

La méthode des moindres carrés donne pour résultat :

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

$$E(\hat{\beta}) = \beta \qquad V(\hat{\beta}) = \sigma^2 \left(X^T X\right)^{-1}$$

Le principale du test et l'intervalle de confiance sont les même comme dans le cas du modèle linéaire simple pour chaque paramètre $\boldsymbol{\beta_i}$ quelconque.

II – Les paramètres du modèle :

Estimateur des moindres carrés ordinaires

Pour trouver les paramètres «β_i» qui minimise S :

$$S = \varepsilon'\varepsilon$$
$$= \sum_i \varepsilon_i^{\,2} = \sum_i [y_i - (\beta_0 + \beta_{i,1}x_1 + \ldots + \beta_{i,p}x_p)]^2$$

On doit résoudre

$$\frac{\partial S}{\partial \beta} = 0$$

Il y a (p+1) équations dites « équations normales » à résoudre

$$\Rightarrow \quad S = \varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$
$$= Y'Y - 2\beta' X'Y + \beta' X' X\beta$$

$$\Rightarrow \quad \frac{\partial S}{\partial \beta} = -2(X'Y) + 2(X'X)\beta = 0$$

$$\Rightarrow \quad \boxed{\hat{\beta} = (X'X)^{-1}X'Y}$$

II – Les paramètres du modèle :

Commentaires

$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$(X'X) = \begin{pmatrix} n & \sum_i x_{i,1} & \cdots & \sum_i x_{i,p} \\ \sum_i x_{i,1} & \sum_i x_{i,1}^2 & & \sum_i x_{i,1} \times x_{i,p} \\ & & & \\ & & & \sum_i x_{i,p}^2 \end{pmatrix}$$

(p+1,p+1)

Matrice des sommes des produits croisés entre les variables exogènes – Symétrique (son inverse aussi est symétrique)

Si les variables sont centrées

•1/n (X'X) = matrice de variance covariance

Si les variables sont centrées et réduites

•1/n (X'X) = matrice de corrélation

$$(X'Y) = \begin{pmatrix} \sum_i y_i \\ \sum_i y_i x_{i,1} \\ \vdots \\ \sum_i y_i x_{i,p} \end{pmatrix}$$

(p+1, 1)

Vecteur des sommes des produits croisés entre l'endogène et les variables exogènes

Si les variables sont centrées

- 1/n (X'Y) = vecteur des covariances entre Y et X

Si les variables sont centrées et réduites

- 1/n (X'Y) = vecteur des corrélations entre Y et X

II – Les paramètres du modèle :

Exemple :

Cigarette	TAR (mg)	NICOTINE (m	WEIGHT (g)	CO (mg)
Alpine	14.1	0.86	0.9853	13.6
Benson&Hedges	16	1.06	1.0938	16.6
CamelLights	8	0.67	0.928	10.2
Carlton	4.1	0.4	0.9462	5.4
Chesterfield	15	1.04	0.8885	15
GoldenLights	8.8	0.76	1.0267	9
Kent	12.4	0.95	0.9225	12.3
Kool	16.6	1.12	0.9372	16.3
L&M	14.9	1.02	0.8858	15.4
LarkLights	13.7	1.01	0.9643	13
Marlboro	15.1	0.9	0.9316	14.4
Merit	7.8	0.57	0.9705	10
MultiFilter	11.4	0.78	1.124	10.2
NewportLights	9	0.74	0.8517	9.5
Now	1	0.13	0.7851	1.5
OldGold	17	1.26	0.9186	18.5
PallMallLight	12.8	1.08	1.0395	12.6
Raleigh	15.8	0.96	0.9573	17.5
SalemUltra	4.5	0.42	0.9106	4.9
Tareyton	14.5	1.01	1.007	15.9
TrueLight	7.3	0.61	0.9806	8.5
ViceroyRichLight	8.6	0.69	0.9693	10.6
VirginiaSlims	15.2	1.02	0.9496	13.9
WinstonLights	12	0.82	1.1184	14.9

Identifiant
(Pas utilisé pour les calculs, mais peut être utilisé pour les commentaires : points atypiques, etc.)

Variables prédictives
Descripteurs ***Variables exogènes***

Quantitative

Variable à prédire
Attribut classe
Variable endogène

Quantitative

II – Les paramètres du modèle :

Exemple des cigarettes :

constante	TAR (mg)	ICOTINE (mg	WEIGHT (g)	CO (mg)
1	14.1	0.86	0.9853	13.6
1	16	1.06	1.0938	16.6
1	8	0.67	0.928	10.2
1	4.1	0.4	0.9462	5.4
1	15	1.04	0.8885	15
1	8.8	0.76	1.0267	9
1	12.4	0.95	0.9225	12.3
1	16.6	1.12	0.9372	16.3
1	14.9	1.02	0.8858	15.4
1	13.7	1.01	0.9643	13
1	15.1	0.9	0.9316	14.4
1	7.8	0.57	0.9705	10
1	11.4	0.78	1.124	10.2
1	9	0.74	0.8517	9.5
1	1	0.13	0.7851	1.5
1	17	1.26	0.9186	18.5
1	12.8	1.08	1.0395	12.6
1	15.8	0.96	0.9573	17.5
1	4.5	0.42	0.9106	4.9
1	14.5	1.01	1.007	15.9
1	7.3	0.61	0.9806	8.5
1	8.6	0.69	0.9693	10.6
1	15.2	1.02	0.9496	13.9
1	12	0.82	1.1184	14.9

(X'X)

24	275.6	19.88	23.0921
275.6	3613.16	254.177	267.46174
19.88	254.177	18.0896	19.266811
23.0921	267.46174	19.266811	22.3637325

(X'X)^-1

6.56299	0.06290	-0.93908	-6.71991
0.06290	0.02841	-0.45200	-0.01528
-0.93908	-0.45200	7.86328	-0.39900
-6.71991	-0.01528	-0.39900	7.50993

X'Y

289.7
3742.85
264.076
281.14508

$\hat{\beta}$

-0.55170	constante
0.88758	tar
0.51847	nicotine
2.07934	weight

$$\hat{\beta} = (X'X)^{-1} X'Y$$

Distribution de $\hat{\beta}$

Par hypothèse, $\varepsilon \equiv N(0, \sigma_\varepsilon)$ $\Rightarrow$ $\begin{cases} \dfrac{\hat{\beta}_j - \beta_j}{\sigma_{\hat{a}_j}} \equiv N(0,1) \\ (n-p-1)\dfrac{\hat{\sigma}^2_\varepsilon}{\sigma^2_\varepsilon} \equiv \chi^2(n-p-1) \end{cases}$

Toujours par analogie avec la régression simple, on peut montrer que $\Rightarrow$ $(n-p-1)\dfrac{\hat{\sigma}^2_\varepsilon}{\sigma^2_\varepsilon} = (n-p-1)\dfrac{\hat{\sigma}^2_{\hat{a}_j}}{\sigma^2_{\hat{a}_j}}$

$\Rightarrow$ $\boxed{\dfrac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\hat{a}_j}} \equiv \tau(n-p-1)}$ ***Loi de Student à (n – p – 1) degrés de liberté.***

On peut la mettre en œuvre dans différents schémas.

- *Test de conformité à un standard c.-à-d. H0 : $a_j = c$ vs. H1: $a_j \neq c$*
  Bilatéral ou unilatéral
- *Test de significativité c.-à-d. H0 : $a_j = 0$ vs. H1: $a_j \neq 0$*
  Permet de déterminer si la variable X_j a un impact sur Y !!!
- *Intervalle de confiance au niveau (1 - α)*

Évaluation globale de la régression
Tableau d'analyse de variance et Coefficient de détermination

Équation d'analyse de variance – Décomposition de la variance

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

SCT
Variabilité totale

SCE
Variabilité expliquée par le modèle

SCR
Variabilité non-expliquée (Variabilité résiduelle)

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Modèle	SCE	p	SCE/p
Résiduel	SCR	n-p-1	SCR/(n-p-1)
Total	SCT	n-1	

Tableau d'analyse de variance

Un indicateur de qualité du modèle : le coefficient de détermination. Il exprime la proportion de variabilité de Y qui est retranscrite par le modèle

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

R^2#1, le modèle est parfait
R^2#0, le modèle est mauvais