

Année Universitaire 2019-2020

Modèle de Régression Linéaire Simple (MLRS)

Slim Zouaoui & Walid Barhoumi

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.

Exemple : 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.

Les objectifs principaux d'une analyse de la régression :

1. Comprendre comment et dans quelle mesure une variable X influence la variable dépendante Y .
2. Développer un modèle pour prévoir des valeurs de Y futures à partir de celles que pourraient prendre la variable X .

Modèle de régression linéaire simple : 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 sont :

- **Y** est la variable dépendante (dite expliquée ou endogène) à caractère aléatoire.
- **X** est la variable indépendante (dite explicative ou exogène) mesurée sans erreur ou fixée à des niveaux arbitraires.
- β_0 et β_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.
- ε 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**

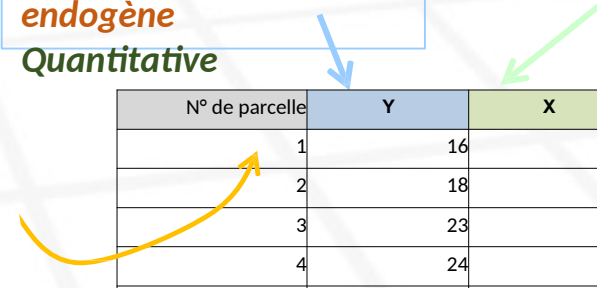
Exemple de régression simple : expliquer le rendement de maïs Y (en quintal) à partir de la quantité d'engrais utilisé (en kilo) sur des parcelles de terrain similaires.

Variable à prédire
Attribut classe
Variable
endogène
Quantitative

Variables prédictive
Descripteur Variable exogène
Quantitative ou binaire

Identifiant

Pas utilisé pour les calculs, mais peut
être utilisé pour les commentaires



N° de parcelle	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Modèle de régression simple : $Y = \beta_0 + \beta_1 X + \varepsilon$

- ❑ Nous disposons donc d'un échantillon de n couples de points (x_i, y_i) , et **on veut expliquer (prédire) les valeurs de Y en fonction des valeurs prises par X.**
- ❑ ε permet de résumer toute l'information qui n'est pas prise en compte dans la relation linéaire entre Y et X (problèmes de spécifications, approximation de la linéarité, variables absentes...)

I.2. Hypothèses du modèle :

Pour que le modèle soit bien défini, outre l'hypothèse de linéarité, il faut vérifier 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) , illustrant les différences entre les valeurs observées et celles estimées de y , devront satisfaire les hypothèses suivantes :

- H_1 : Les erreurs ont toutes une moyenne nulle.

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

- H_2 : L'homoscédasticité des erreurs.

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

- 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.

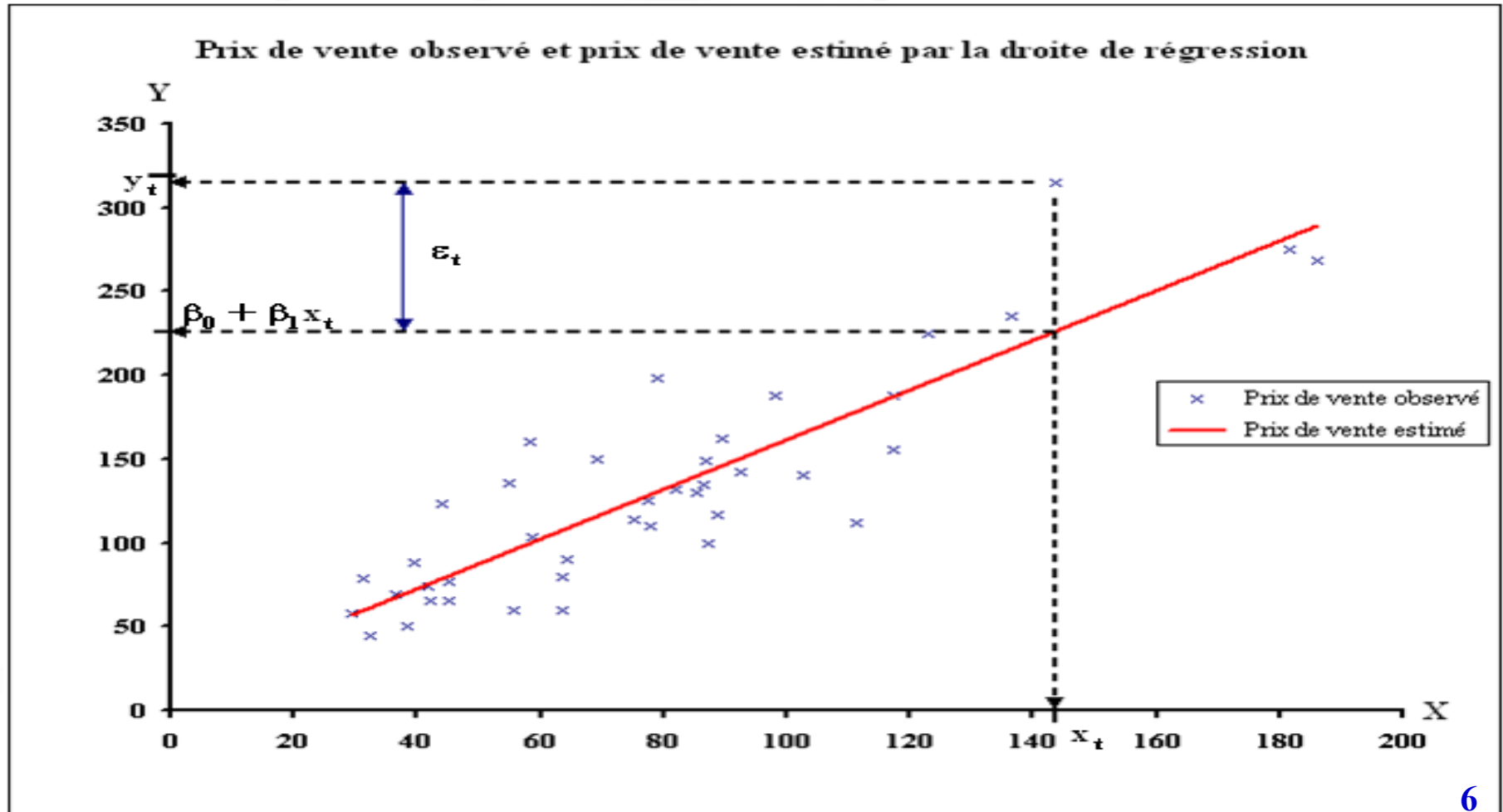
$$\text{Cov}(\varepsilon_t, \varepsilon_{t'}) = 0 \quad \forall t \neq t'$$

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

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

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 :



II – Les paramètres du modèle :

Les paramètres inconnus du modèle sont de deux sortes : Il y a les coefficients β_0 et β_1 d'une part et la variance des erreurs σ^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 β_0 et β_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 :

$$\varepsilon_t = y_t - (\beta_0 + \beta_1 x_t)$$

La méthode des moindres carrés ordinaires consiste à minimiser, par rapport aux 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 β_0 et β_1 . Les conditions de minimisation sont les suivantes :

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

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

On a alors :

$$\begin{cases} \frac{\partial SCR}{\partial \beta_0} = \frac{\partial \sum_{t=1}^n (y_t - \beta_0 - \beta_1 x_t)^2}{\partial \beta_0} = 0 \\ \frac{\partial SCR}{\partial \beta_1} = \frac{\partial \sum_{t=1}^n (y_t - \beta_0 - \beta_1 x_t)^2}{\partial \beta_1} = 0 \end{cases} \quad \begin{cases} \sum_{t=1}^n -2(y_t - \beta_0 - \beta_1 x_t) = 0 \\ \sum_{t=1}^n -2x_t(y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \sum_{t=1}^n (y_t - \beta_0 - \beta_1 x_t) = 0 \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \sum_{t=1}^n y_t - \sum_{t=1}^n \beta_0 - \sum_{t=1}^n \beta_1 x_t = 0 \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \sum_{t=1}^n y_t - n\beta_0 - \beta_1 \sum_{t=1}^n x_t = 0 \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \frac{\sum_{t=1}^n y_t}{n} - \frac{n\beta_0}{n} - \beta_1 \frac{\sum_{t=1}^n x_t}{n} = 0 \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \bar{y} - \beta_0 - \beta_1 \bar{x} = 0 \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \sum_{t=1}^n x_t (y_t - \beta_0 - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \sum_{t=1}^n x_t (y_t - (\bar{y} - \beta_1 \bar{x}) - \beta_1 x_t) = 0 \end{cases}$$

$$\begin{cases} \beta_0 = \bar{y} - \beta_1 \bar{x} \\ \sum_{t=1}^n x_t (y_t - \bar{y}) - \beta_1 \sum_{t=1}^n x_t (x_t - \bar{x}) = 0 \end{cases}$$

La condition de second ordre est vérifiée. On obtient alors les estimateurs des moindres carrés ordinaires suivants :

$$\begin{cases} \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 = \frac{\sum_{t=1}^n \mathbf{x}_t (\mathbf{y}_t - \bar{y})}{\sum_{t=1}^n \mathbf{x}_t (\mathbf{x}_t - \bar{\mathbf{x}})} \end{cases} \quad \text{Ou bien}$$

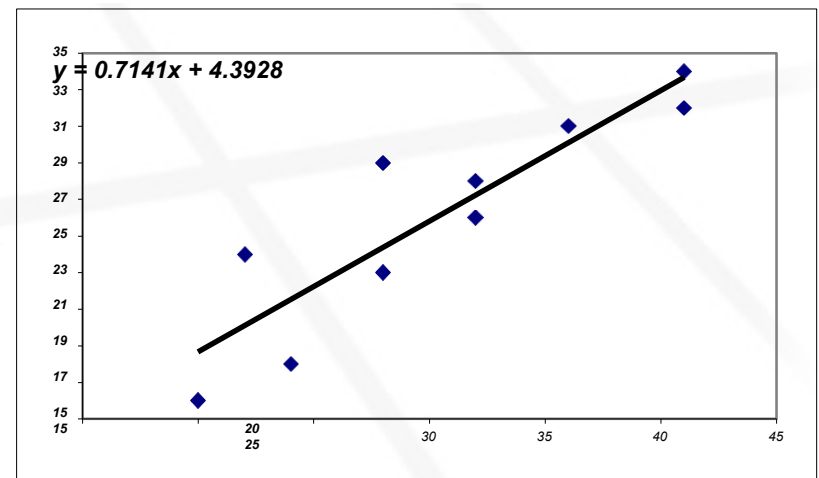
$$\begin{cases} \hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \\ \hat{\beta}_1 = \frac{\text{cov}(X, Y)}{\text{var}(X)} \end{cases}$$

Exemple des rendements agricoles

	Y	X	Y - \bar{Y}	X - \bar{X}	(Y - \bar{Y}) ²	(X - \bar{X}) ²
1	16	20	-10.1	-10.4	105.04	108.160
2	18	24	-8.1	-6.4	51.84	40.960
3	23	28	-3.1	-2.4	7.44	5.760
4	24	22	-2.1	-8.4	17.64	70.560
5	28	32	1.9	1.6	3.04	2.560
6	29	28	2.9	-2.4	6.96	5.760
7	26	32	-0.1	1.6	-0.16	2.560
8	31	36	4.9	5.6	27.44	31.360
9	32	41	5.9	10.6	62.54	112.360
10	34	41	7.9	10.6	83.74	112.360

Moyenne 26.1 30.4 Somme 351.6 492.4

$$\begin{cases} \hat{a} = \frac{351.6}{492.4} = 0.714 & \hat{\beta}_1 \\ \hat{b} = 26.1 - 0.714 \times 30.4 = 4.39 & \hat{\beta}_0 \end{cases}$$



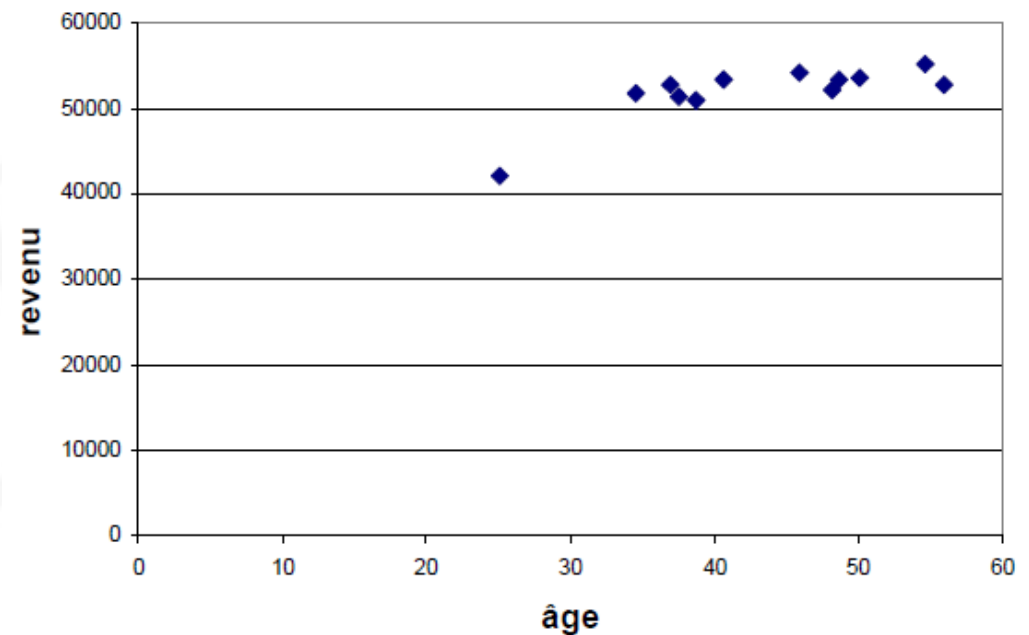
Exemple :

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 :

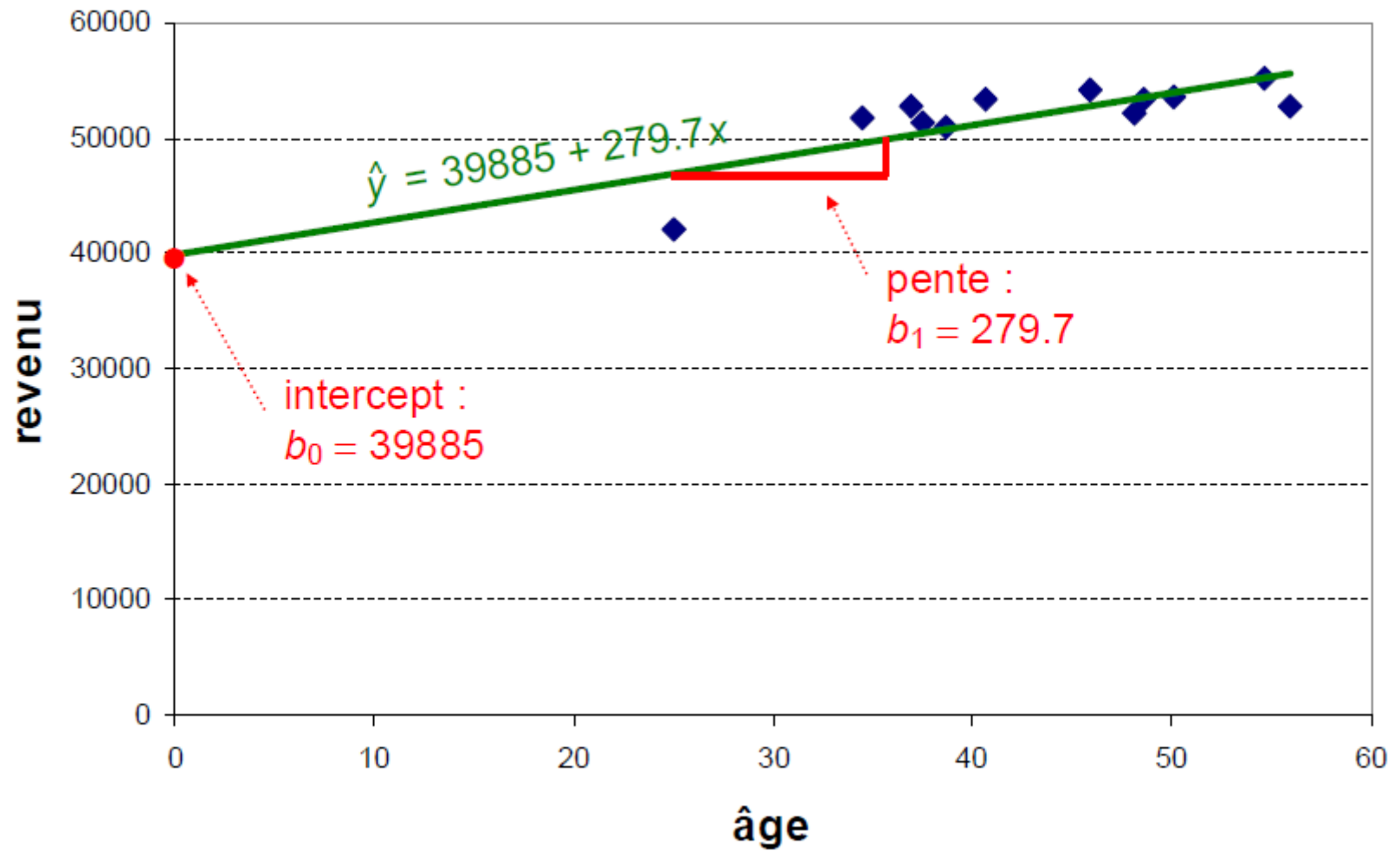


Exemple (suite) :

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

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

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



$\beta_0 = b_0$ et $\beta_1 = b_1$

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

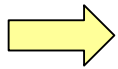
$$\varepsilon_t \sim N(0; \sigma^2)$$

$\sigma^2 = SCR/(n-2)$ Prévoir des valeurs de Y futures à partir de celles de X

III – Validation du modèle : coefficient de détermination R^2

On évalue la qualité de l'estimation du modèle de régression par : $R^2 = \frac{SCE}{SCT}$ Avec

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



Décomposition
de la variance

$$\sum_i (y_i - \bar{y})^2 = \sum_i (y_i^i - \hat{y}^i)^2 + \sum_i (\hat{y}^i - \bar{y})^2$$
$$SCT = SCE + SCR$$

SCT : somme des carrés totaux

SCE : somme des carrés expliqués par le modèle

SCR : somme des carrés résiduels, non expliqués par le modèle

III – Validation du modèle : coefficient de détermination R^2

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

IV – Validation du modèle : coefficient de détermination R²

$$\hat{y}_i = a\hat{x}_i + b$$

$$\hat{b} = 0.714x + 4.39$$

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

$$\varepsilon^2$$

	Y	X	(Y - \bar{Y})	(X - \bar{X})	(Y - \bar{Y})(X - \bar{X})	(X - \bar{X}) ²	(Y - \bar{Y}) ²	\hat{Y}	Résidus	Résidus ²
1	16	20	-10.1	-10.4	105.04	108.160	102.010	18.674	-2.674	7.149
2	18	24	-8.1	-6.4	51.84	40.960	65.610	21.530	-3.530	12.461
3	23	28	-3.1	-2.4	7.44	5.760	9.610	24.386	-1.386	1.922
4	24	22	-2.1	-8.4	17.64	70.560	4.410	20.102	3.898	15.195
5	28	32	1.9	1.6	3.04	2.560	3.610	27.242	0.758	0.574
6	29	28	2.9	-2.4	-6.96	5.760	8.410	24.386	4.614	21.286
7	26	32	-0.1	1.6	-0.16	2.560	0.010	27.242	-1.242	1.544
8	31	36	4.9	5.6	27.44	31.360	24.010	30.099	0.901	0.812
9	32	41	5.9	10.6	62.54	112.360	34.810	33.669	-1.669	2.785
10	34	41	7.9	10.6	83.74	112.360	62.410	33.669	0.331	0.110
Moyenne	26.1	30.4		Somme	351.6	492.4	314.9		Somme	63.838749
							SCT			SCR

ESTIMATION

a	0.714053615
b	4.392770106

$$SCE = SCT - SCR = 251.061251$$

$$R^2 = 0.79727295$$

$$R = 0.89290142$$