CHAPITRE I.
MODELE LINEAIRE DE REGRESSION SIMPLE (MRS)

Régression
Généralités

- **Faire de la régression**, c'est trouver la meilleure prédiction possible d'une grandeur (numérique) lorsqu'on connaît les valeurs d'autres grandeurs.

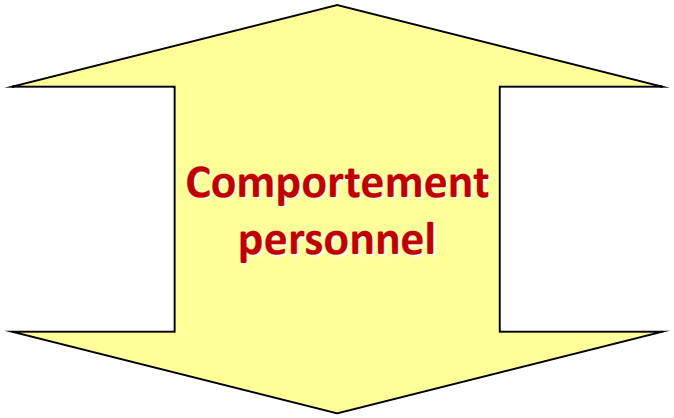
- La grandeur à prédire est dite « **variable à expliquer** », les autres grandeurs étant des « **variables explicatives** ».

- **Par exemple**, vous pourriez chercher à prédire **le budget** qu'un foyer va consacrer à l'achat d'une nouvelle automobile (variable à expliquer) en fonction des âges et revenus respectifs des parents, du nombre d'enfants, du nombre de kilomètres annuels de la famille, et de l'âge de leur voiture actuelle (variables explicatives).

Régression

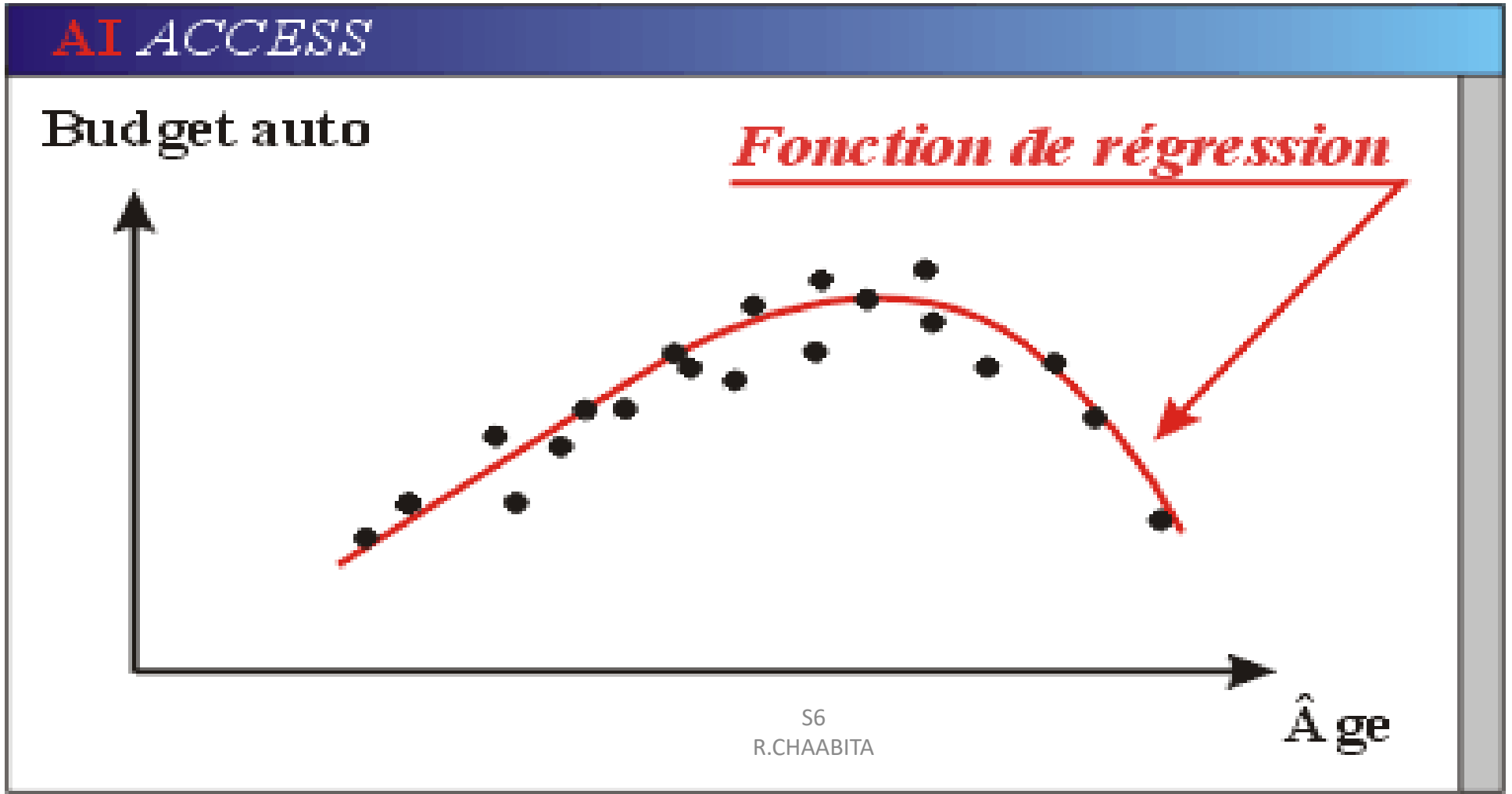
- **Toute la difficulté vient de ce que** :
  - **Le lien entre variables explicatives et variable à expliquer peut être assez complexe.**

  Ex: Deux foyers décrits par des profils identiques peuvent avoir consacré des sommes différentes à l'achat d'une nouvelle voiture.



- **C'est cette dernière raison qui empêche la prédiction d'être parfaite.**

- Dans le cas où il n'y a qu'une seule variable explicative et une seule variable à expliquer, la régression se visualise comme le problème qui consiste à faire passer "au mieux" une courbe dans un nuage de points



Modèle Linéaire

- Attention, ce terme a un double sens.

- Un modèle est dit "**linéaire dans les variables**" si les variables n'y interviennent que sous la forme de combinaisons linéaires.
- Un modèle est dit "**linéaire dans les paramètres**" si les paramètres n'y interviennent que par des combinaisons linéaires.

Régression Linéaire Simple

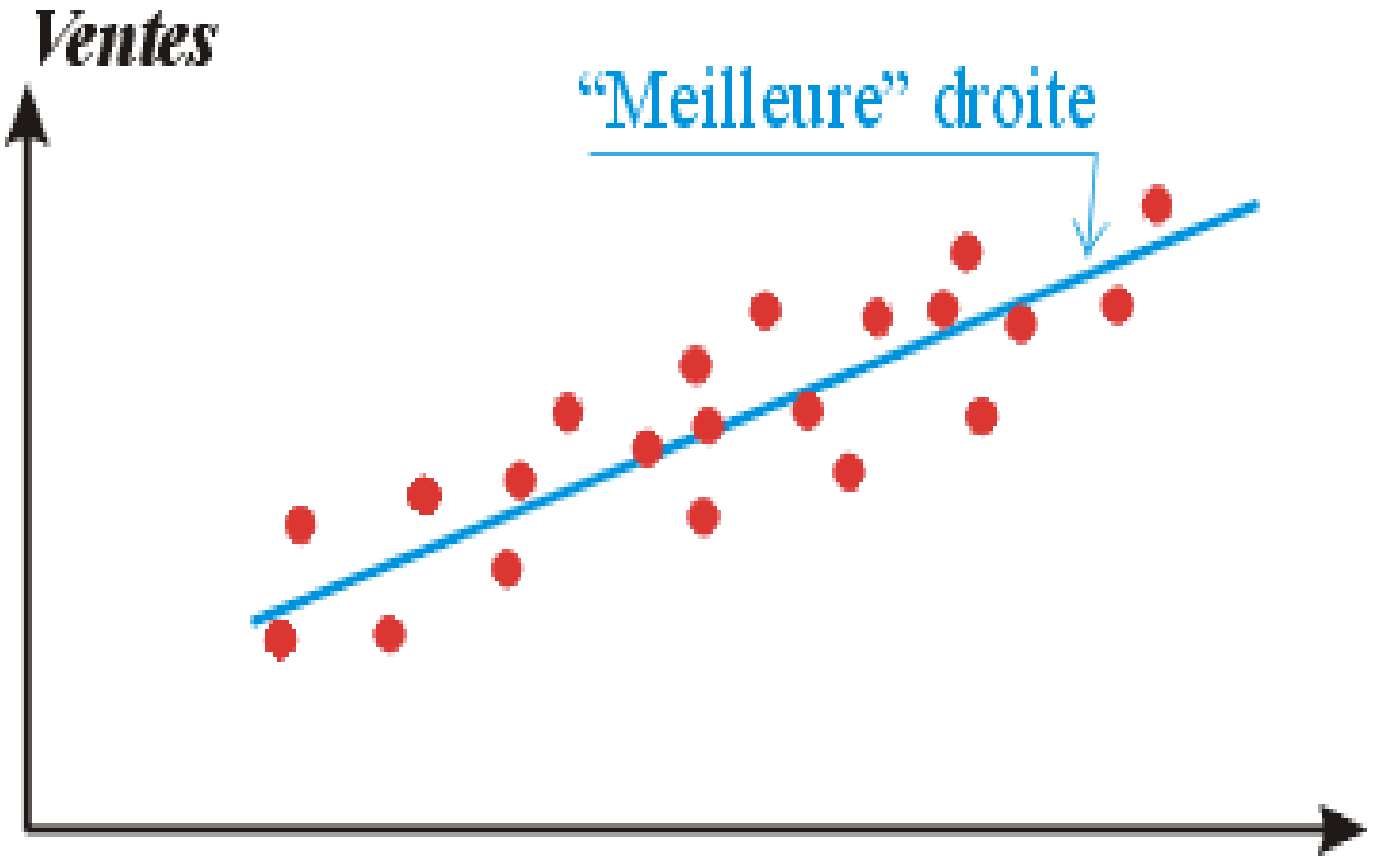
- La plus simple et la plus populaire des techniques de régression.

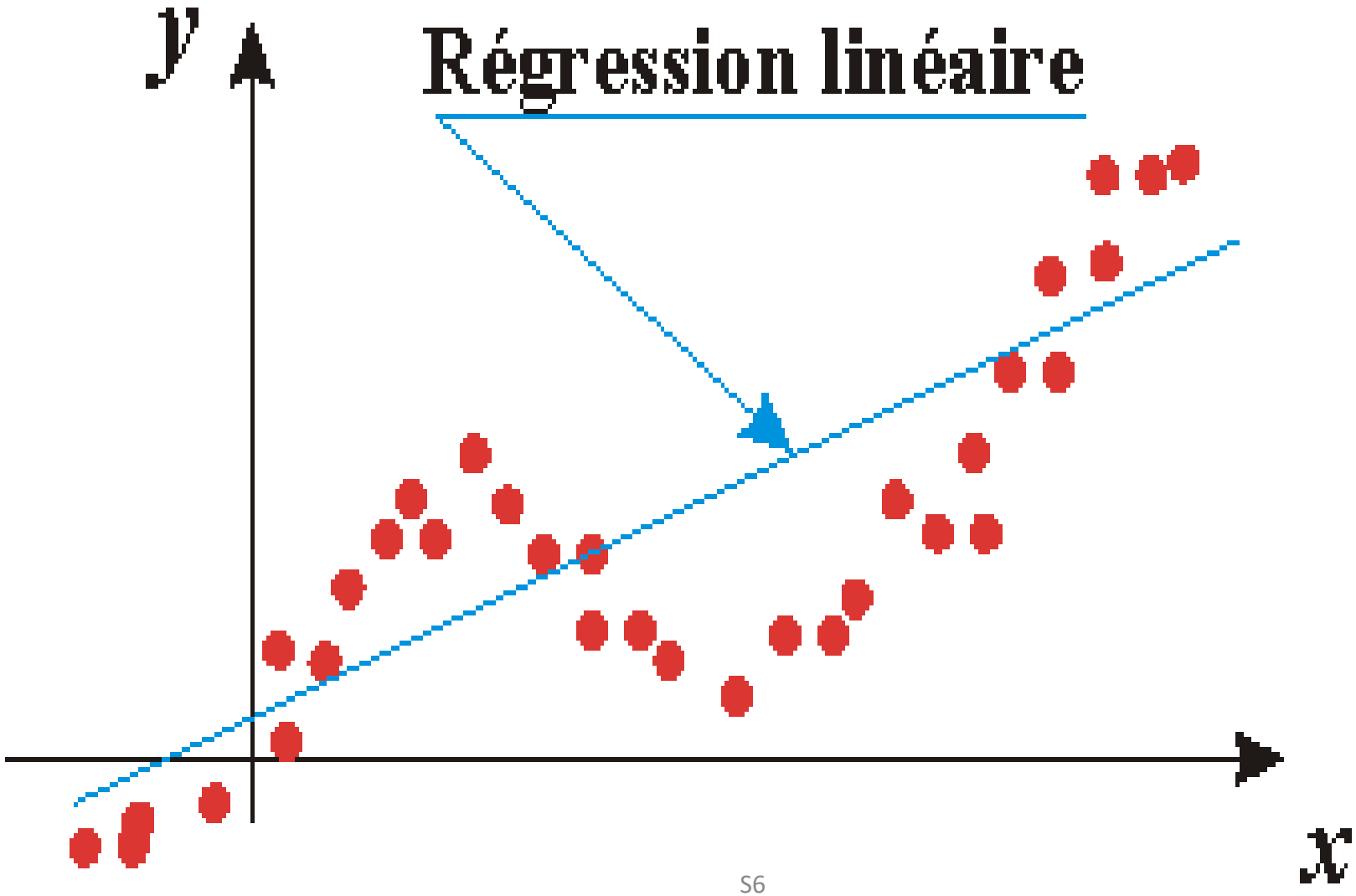
- La Régression Linéaire Simple (RLS) est un cas particulier de régression dans lequel:

  * Il n'y a qu'une seule **variable explicative** (**numérique**),

  * Le modèle est linéaire dans la variable et dans les paramètres.

- Par exemple, supposons que vous vouliez "expliquer" la variable "**Ventes**" par la variable "**Budget_Promotionel**", ces chiffres étant normalisé à travers les différents magasin d'un groupe de distribution. Il est vraisemblable qu'il est possible de faire passer une ligne droite à travers un "**nuage de points**", chaque point représentant un magasin. La RLS a pour objectif d'identifier la "**meilleure droite**" dans un sens précis.


Ventes
"Meilleure" droite
Budget promotionnel


Régression linéaire
y
x

Régression Linéaire Simple

- Dans la pratique, la RLS n'est pas considérée comme un "cas particulier", mais **la** technique de régression par excellence. Il y a plusieurs raisons à cela :

  1) **La détermination de la "meilleure droite", ou "Droite des Moindres Carrés"** (**La courbe de régression est la ligne qui rend minimum la somme des carrés des écarts**) est simple, et repose sur des principes assez intuitifs.

  2) **Les paramètres (ou coefficients) de la régression peuvent être facilement interprétés**.

  3) Sous réserve d'hypothèses assez restrictives, la RLS s'appuie sur une théorie mathématique complète qui lui permet de résoudre les délicats problèmes de la stabilité du **modèle et de son pouvoir de généralisation, sans avoir recours aux lourdes techniques de validation.**

Exemple Introductif

- Soit la fonction de consommation Keynésienne: $C = a_o + a_1 Y$

Où : C = Consommation,

Y = revenu,

a_1 = propension marginale à consommer

a_o = consommation autonome ou incompressible

Vocabulaire

- La variable consommation est appelée

«**variable à expliquer**» ou «**variable endogène**»

- La variable revenu est appelée

«**variable explicative**» ou «**variable exogène** »

- a_0 et a_1 sont **les paramètres du modèle** ou encore **les coefficients de régression**

Spécification

On distingue 2 types

- **Les modèles en série temporelle**, les variables représentent des phénomènes observés à intervalles de temps réguliers, par exemple la Consommation et le revenu annuel de 1985 à 2005 pour le Maroc. Le modèle s'écrit alors: $Ct = a_o + a_1 Y_t$ $t= 1985, \ldots, 2005$

  Où C_t = Consommation au temps t, Y_t = Revenu au temps t

- **Les modèles en coupe instantanée, les variables représentent des phénomènes observés au même instant mais concernant plusieurs individus par exemple, la consommation et le revenu observés sur un échantillon de 20 pays en 2005.** Le modèle s'écrit alors: $C_i = a_o + a_1 Y_i$ $i= 1, \ldots, 20$

  Où C_i = Consommation pour le pays, Y_i = Revenu pour le pays i.

CONSTAT

- Le modèle tel qu'il vient d'être spécifie n'est **qu'une caricature de la réalité et une présentation incomplète.**
- Ne retenir que le revenu pour expliquer la consommation est à l'évidence même **insuffisant**: il existe une multitude d'autres facteurs (variables) susceptibles d'expliquer la consommation (tels que: l'âge, situation matrimoniale, nombre d'enfants, niveau d'instruction, catégorie socioprofessionnelles,etc.
- **Question:**

comment remédie à ce constat?

Rôle du terme aléatoire

- Pour remédier à ce problème, nous ajoutons un terme (U_t). Qui synthétise l'ensemble de ces informations non explicitées dans le modèle: $\mathbf{C_t = a_o + a_1 Y_t + U_t}$ si le modèle est spécifié en série temporelle

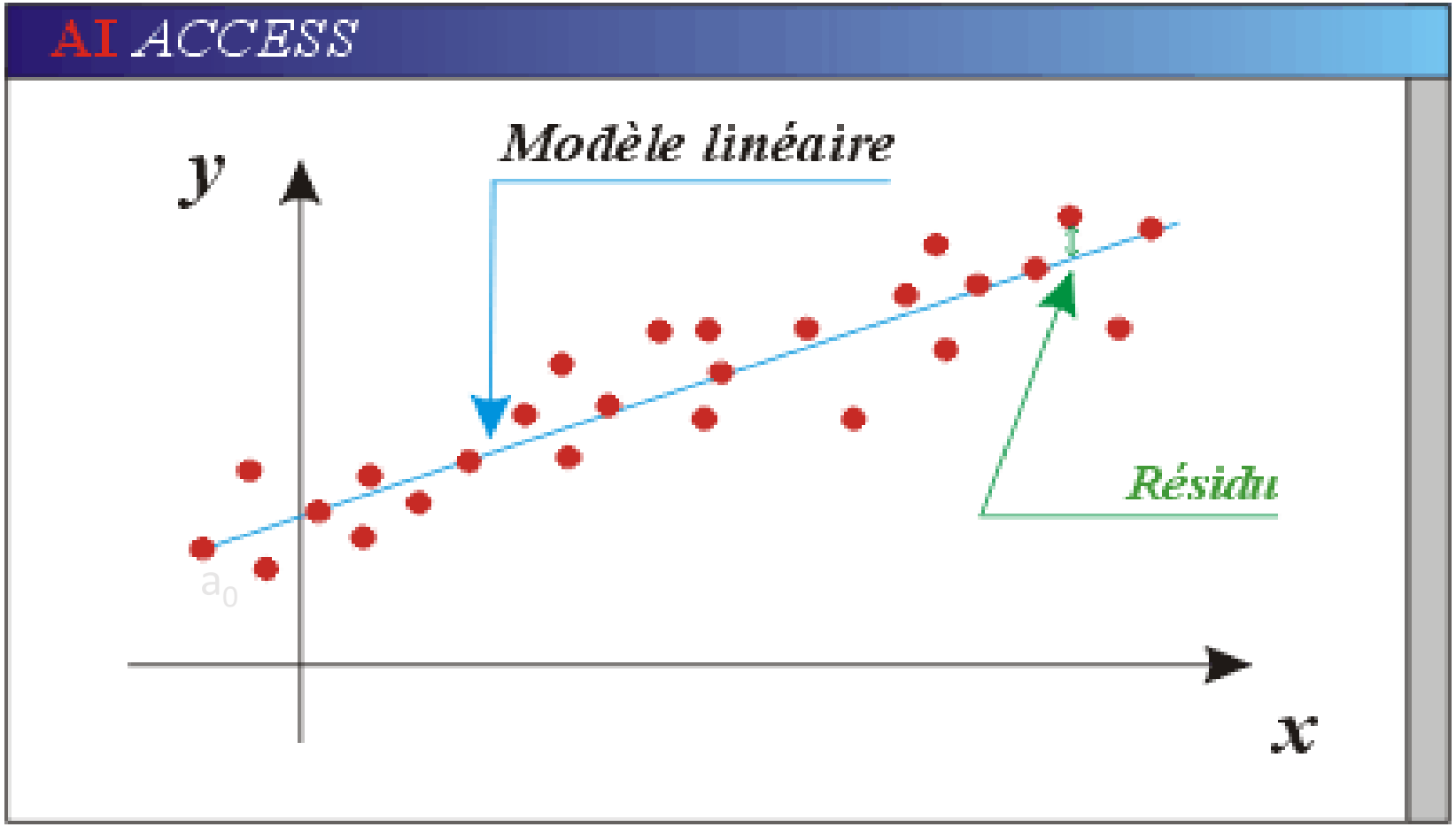
$\mathbf{C_i = a_o + a_1 Y_i + U_i}$ si le modèle est spécifié en coupe instantané

- **Où Ut représente l'erreur de spécification du modèle, c'est à dire l'ensemble des phénomènes explicatifs de la consommation non liés au revenu**
  - **Le terme Ut mesure quoi?**

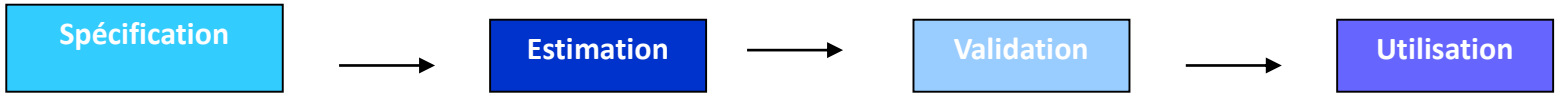
Le terme aléatoire

- **Il mesure la différence entre les valeurs réellement observées de C_t et les valeurs qui auraient été observées**. Le terme Ut regroupe donc trois erreurs:
- **Une erreur de spécification**, càd le fait que la seule variable exogène n'est pas suffisante pour rendre compte la totalité du phénomène non expliquer.
- **Une erreur de mesure**, les données ne représentent pas exactement le phénomène
- **Une erreur de fluctuation d'échantillonnage**, d'un échantillon à un autre les observations, et donc les estimations, sont légèrement différentes.

Comment on calcul l'aléa U_t graphiquement?



Les principales phases de la modélisation



On cherchera à répondre à deux questions principales au niveau de chaque phase :

En quoi elle consiste ?

Comment y procéder ?

I - PRESENTATION GENERALE :

$$Y_i = aX_i + b + U_i$$

Y : variable endogène.
X : variable exogène.

U : Variable aléatoire appelée résidu, elle est une mesure de l'ignorance.
On dispose de "n" observations sur Y et X (i=1,2,…,n). Nous avons donc "n" couples (Yi, Xi) qui sont des réalisations des variables Y et X.
a et b sont des paramètres réels et inconnus que l'on se propose d'estimer à l'aide des observations

HYPOTHESES « CLASSIQUES » POUR LE M.R.S

Hypothèse 1 : *Le modèle est correctement spécifié*

la variable explicative retenue soit la « meilleure » sans omission d'autres variables, la vraie relation soit une relation linéaire dans ou par rapport aux paramètres à estimer et enfin la variable aléatoire intervienne de manière additive

Hypothèse 2 : Les Yi et Xi sont des grandeurs numériques observées sans erreur.

-

Y est une variable aléatoire par l'introduction de U

$$E(U_i) = 0 \text{ quelque soit } X_i (i = 1 \ldots n).$$

Ce qui importe est que l'espérance mathématique de Ui soit nulle ou la même pour tout i. Cette **hypothèse est une hypothèse de permanence structurelle.**

Hypothèse 3 : *L'homoscédasticité*

Ui est distribuée selon une loi de probabilité indépendante de "i" et de Xi

$$V(U_i) = E(U_i^2) = \delta_u^2$$ **Quantité finie**

Hypothèse 3 reprend l'hypothèse 2 mais elle est plus forte .

Si H3 n'est pas réalisée, on parle **d'hétéroscédasticité.**

Hypothèse 4 : *Hypothèse d'indépendance des erreurs (ou résidus)*

On suppose que le Ui et les Uj erreurs relatives à 2 observations différentes sont indépendantes entre elles c'est-à-dire

$$Cov(U_i, U_j) = 0 \quad \forall i \neq j.$$

Hypothèse 5 : Hypothèse de normalité

On suppose que les Ui sont distribuées selon une loi normale.

Hypothèse 6 : *Hypothèse concerne la variable exogène*

Lorsque **n tend vers l'infini**, la suite des X_i est telle que

$$\frac{\sum X_i}{n} = \overline{X} \text{ tend vers } X_0.$$

$$\frac{\sum (Xi - \overline{X})^2}{n} \text{ tend vers } S^2 \text{ avec } X_0 \text{ et } S^2 \text{ des quantités finies.}$$

S^2 l'estimation de la variance de l'erreur

$$s^2 = \frac{\sum \hat{u}_i^2}{n-2}$$

Cette hypothèse est utile pour l'étude des propriétés des estimateurs de a et b.

$\hat{a}$ et $\hat{b}$ **sont des estimateurs convergents.**

<u>Hypothèse 7</u> : *On ne dispose d'aucune information (restriction) sur les paramètres a et b à estimer.*

Ils peuvent prendre n'importe quelle valeur réelle positive, négative ou nulle.

RÉGRESSION

- **Définir les concepts suivants ainsi que leurs fonctions**

a) Modèles de la régression simple

b) Modèle linéaire de la régression

c) Diagramme de dispersion

d) Terme d'erreur

révision

- **Formuler la relation générale entre la consommation, Y et le revenu, X,**

a/ sous un forme linéaire exacte

b/ sous une forme aléatoire

c/ pourquoi peut-on s'attendre à ce que la plupart des valeurs observées de Y ne donnent pas des points situés exactement en ligne droite

révision

- En quel sens la méthode de MCO permet-elle d'estimer la « meilleure » droite d'ajustement pour un échantillon d'observations XY?
- Pourquoi choisir les écarts verticaux?
- Pourquoi ne pas prendre simplement des écarts sans les portes au carré?
- Pourquoi ne pas prendre la somme des valeurs absolues des écarts

Définir les concepts suivants ainsi que leurs fonctions

a) Modèles de la régression simple

b) Modèle linéaire de la régression

c) Diagramme de dispersion

d) Terme d'erreur

a/ Le modèle de régression simple est utilisé pour tester des hypothèses portant sur la relation entre une variable dépendante, Y, et une variable indépendante, X ; il sert également à la prévision.

b/ Le modèle linéaire de la régression suppose qu'il existe une relation linéaire approchée entre X et Y : autrement dit, l'ensemble des couples de valeurs Xi et Yi appartenant à l'échantillon aléatoire observé donne des points (Xi,Yi) répartis sur une droite ou au voisinage immédiat de celle-ci.

c/ Un diagramme de dispersion est un graphe qui associe à chaque couple d'observations dépendantes et indépendantes un point dans un plan euclidien orthonormé XY.

d/ Le terme d'erreur (encore appelé terme stochastique ou perturbation aléatoire) mesure l'écart (d'ordinaire en projection verticale) entre chaque valeur observée Y et la valeur « vraie » mais inobservable, donnée par la courbe de régression.

1/ En quel sens la méthode de MCO permet-elle d'estimer la « meilleure » droite d'ajustement pour un échantillon d'observations XY?
2/ Pourquoi choisir les écarts verticaux?
3/ Pourquoi ne pas prendre simplement des écarts sans les portes au carré?
4/ Pourquoi ne pas prendre la somme des valeurs absolues des écarts

1/ Une droite ajuste les données (les observations de l'échantillon XY) au sens des moindres carrés lorsque, sur le graphe de dispersion, **la somme des carrés des distances verticales entre les points observés et la droite est minimale**.

2/ On utilise **les écarts verticaux** parce qu'on s'efforce d'expliquer ou de prédire les changements de Y, lequel est mesuré sur l'axe vertical.

3/ Si l'on somme simplement les écarts, deux écarts de même valeur absolue mais de signes opposées s'éliminent, de sorte que la somme totale est nulle : **la méthode serait inapplicable.**

4/ On pourrait éviter la difficulté précédente en prenant la somme des valeurs absolues des écarts. On préfère toutefois utiliser **la somme des écarts quadratique** de manière à défavoriser relativement les grands écarts par rapport aux petits.

DETERMINATION de â et de ^b PAR LES MOINDRES CARRES ORDINAIRES : (M.C.O)

$$\overline{Y} - \hat{a}\overline{X} - \hat{b} = 0 \qquad \Rightarrow \qquad \hat{b} = \overline{Y} - \hat{a}\overline{X}$$

$$\hat{a} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

Exemple N°1

- Pendant 10 ans, de 1996 à 2005, une ferme a expérimenté le rendement du maïs (**Y en kg/ha**) associé à l'emploi de quantité croissantes d'un fertilisant (**X en g**). Le tableaux suivant rassemble ces données .

1/Reporter ces données sur le diagramme de dispersion

2/ Donner l'équation estimée de la droite de régression

3/ Tracer la droite de régression et calculer l'écart (ei) entre Yi et ^Yi

Année	n	Y_i	X_i
1996	1	40	6
1997	2	44	10
1998	3	46	12
1999	4	48	14
2000	5	52	16
2001	6	58	18
2002	7	60	22
2003	8	68	24
2004	9	74	26
2005	10	80	32

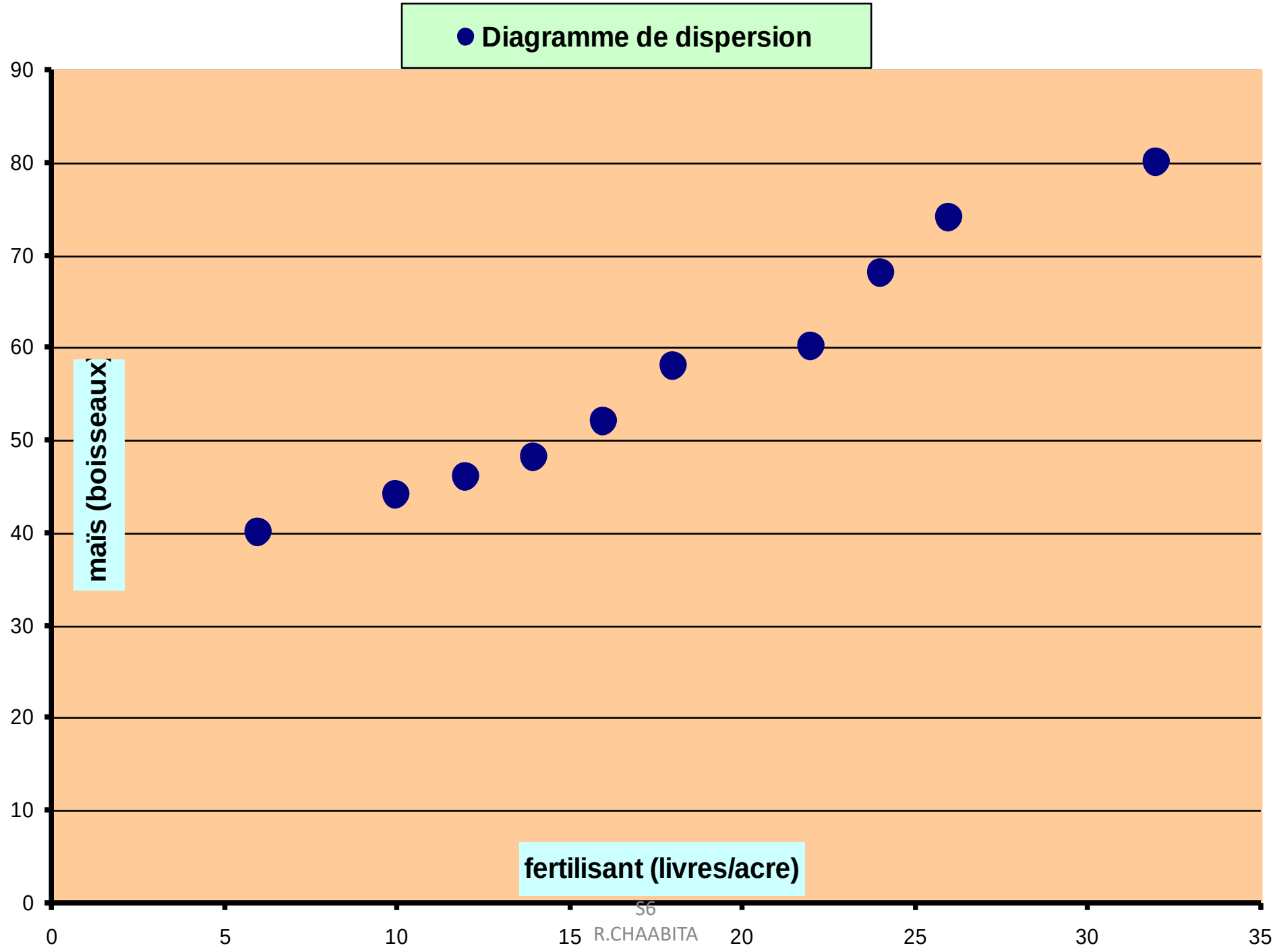


Diagramme de dispersion
maïs (boisseaux)
fertilisant (livres/acre)
0
10
20
30
40
50
60
70
80
90
0
5
10
15
20
25
30
35

On sait que

$$\hat{a} = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2}$$

et

$$\hat{b} = \overline{Y} - \hat{a}\overline{X}$$

- On écrit que

$$x_i = Xi - \overline{X} \quad et \quad y_i = Yi - \overline{Y}$$

- alors

$$\hat{a} = \frac{\sum x_i y_i}{\sum x_i^2}$$

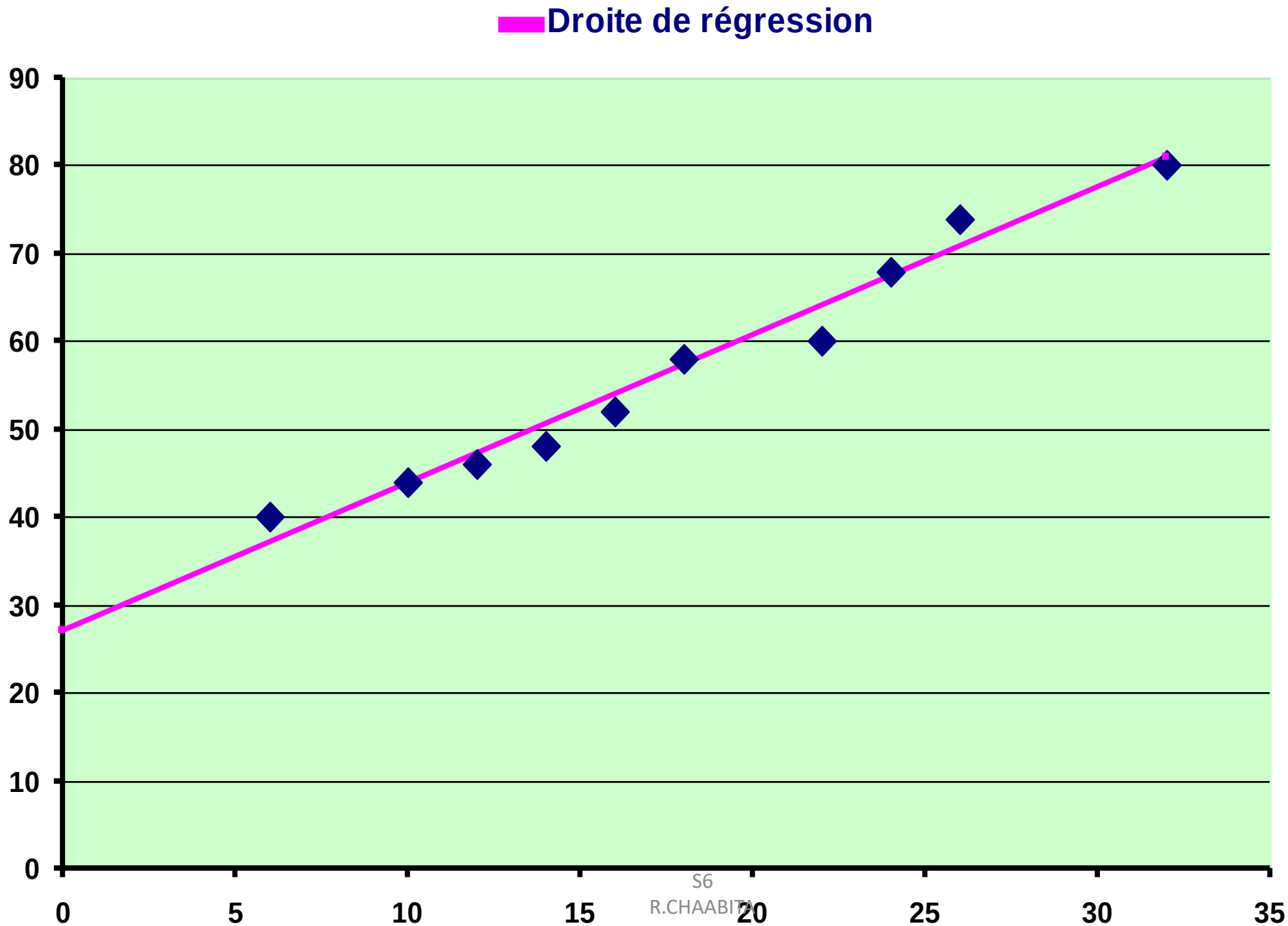
n	Yi	Xi	yi	xi	xiyi	x^2_i
1	40	6	-17	-12	204	144
2	44	10	-13	-8	104	64
3	46	12	-11	-6	66	36
4	48	14	-9	-4	36	16
5	52	16	-5	-2	10	4
6	58	18	1	0	0	0
7	60	22	3	4	12	16
8	68	24	11	6	66	36
9	74	26	17	8	136	64
10	80	32	23	14	322	196
somme	570	180	0	0	956	576
moyenne	57	18				

Donc â = 956/576 = 1,66
(pente estimée de la droite de régression)

et b^ = 57 – (1,66x18)=
57 – 29,88 = 27,12 (ordonnée à l'origine).
Equation estimée de la droite de régression:
^Yi = 27,12 +1,66 Xi

Par conséquent, si Xi = 0 alors Y^= 27,12= b^.
Et lorsque Xi= 18 = moyenne de x alors
Y^= 27,12 + (1,66x18)=57 = moyenne de Y.

Il en résulte que la droite de régression passe par le point $(\overline{X}, \overline{Y})$


Droite de régression
90
80
70
60
50
40
30
20
10
0
0
5
10
15
20
25
30
35

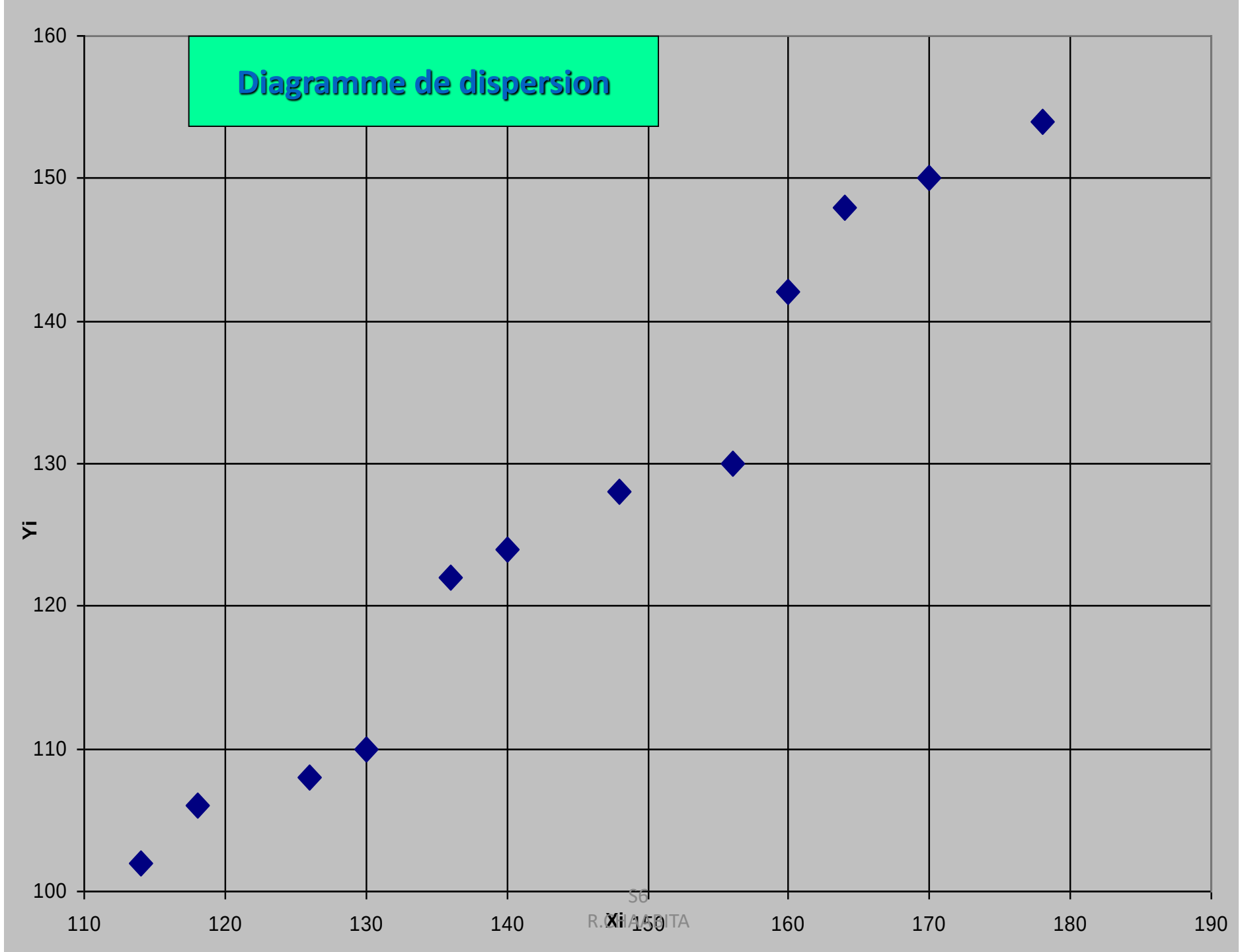
n	Y_i	$\hat{Y}_i$	e_i
1	40	37,08	2,92
2	44	43,72	0,28
3	46	47,04	-1,04
4	48	50,36	-2,36
5	52	53,36	-1,68
6	58	57	1
7	60	63,64	-3,64
8	68	66,96	1,04
9	74	70,28	3,72
10	80	80,24	-0,24
somme	570		0
moyenne	57		

Exemple N°2

1/Construire un diagramme de dispersion pour
les données en milliards de Dhs:
dépenses de consommation, Y,
et revenu disponible, X , pendant douze années
de 1994 à 2005.

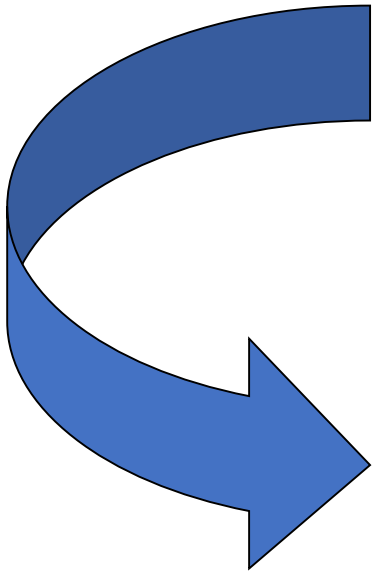
2/ Etablir équation de régression .
3/ Tracer la droite de régression correspondante,
en indiquant l'écart spécifiant chaque couple

année	n	Yi	Xi
1994	1	102	114
1995	2	106	118
1996	3	108	126
1997	4	110	130
1998	5	122	136
1999	6	124	140
2000	7	128	148
2001	8	130	156
2002	9	142	160
2003	10	148	164
2004	11	150	170
2005	12	154	178



ni	Yi	Xi	yi	xi	xi^2	xiyi
1	102	114	-25	-31	961	775
2	106	118	-21	-27	729	567
3	108	126	-19	-19	361	361
4	110	130	-17	-15	225	255
5	122	136	-5	-9	81	45
6	124	140	-3	-5	25	15
7	128	148	1	3	9	3
8	130	156	3	11	121	33
9	142	160	15	15	225	225
10	148	164	21	19	361	399
11	150	170	23	25	625	575
12	154	178	27	33	1089	891
Somme	1524	1740	0	0	**4812**	**4144**
Moyenne	127	145				

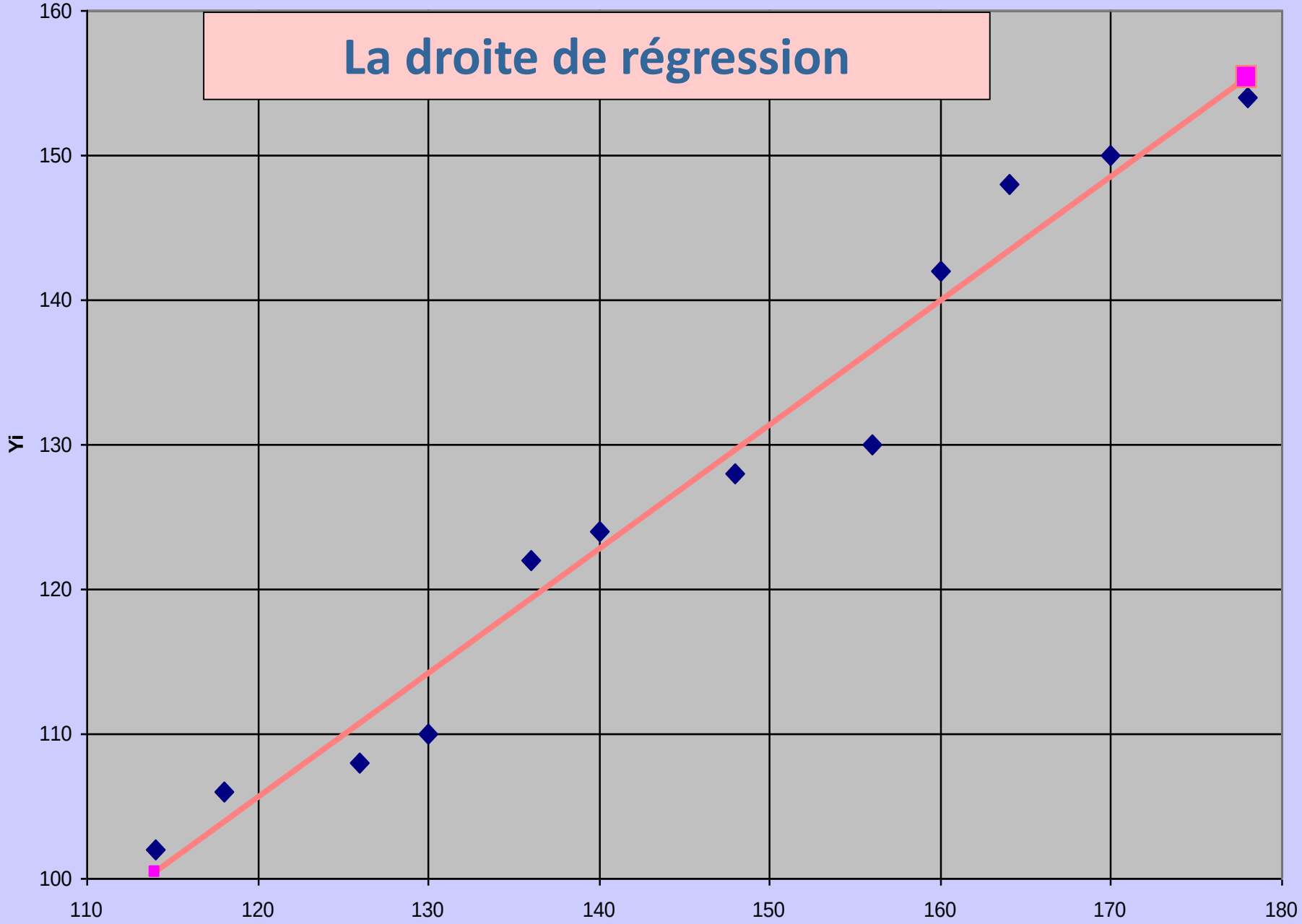
â= 4144/4812= 0,86



^b= 127-(0,86*145)= 2,30

$$\hat{Y} = 0{,}86\hat{X} + 2{,}30$$

La droite de régression



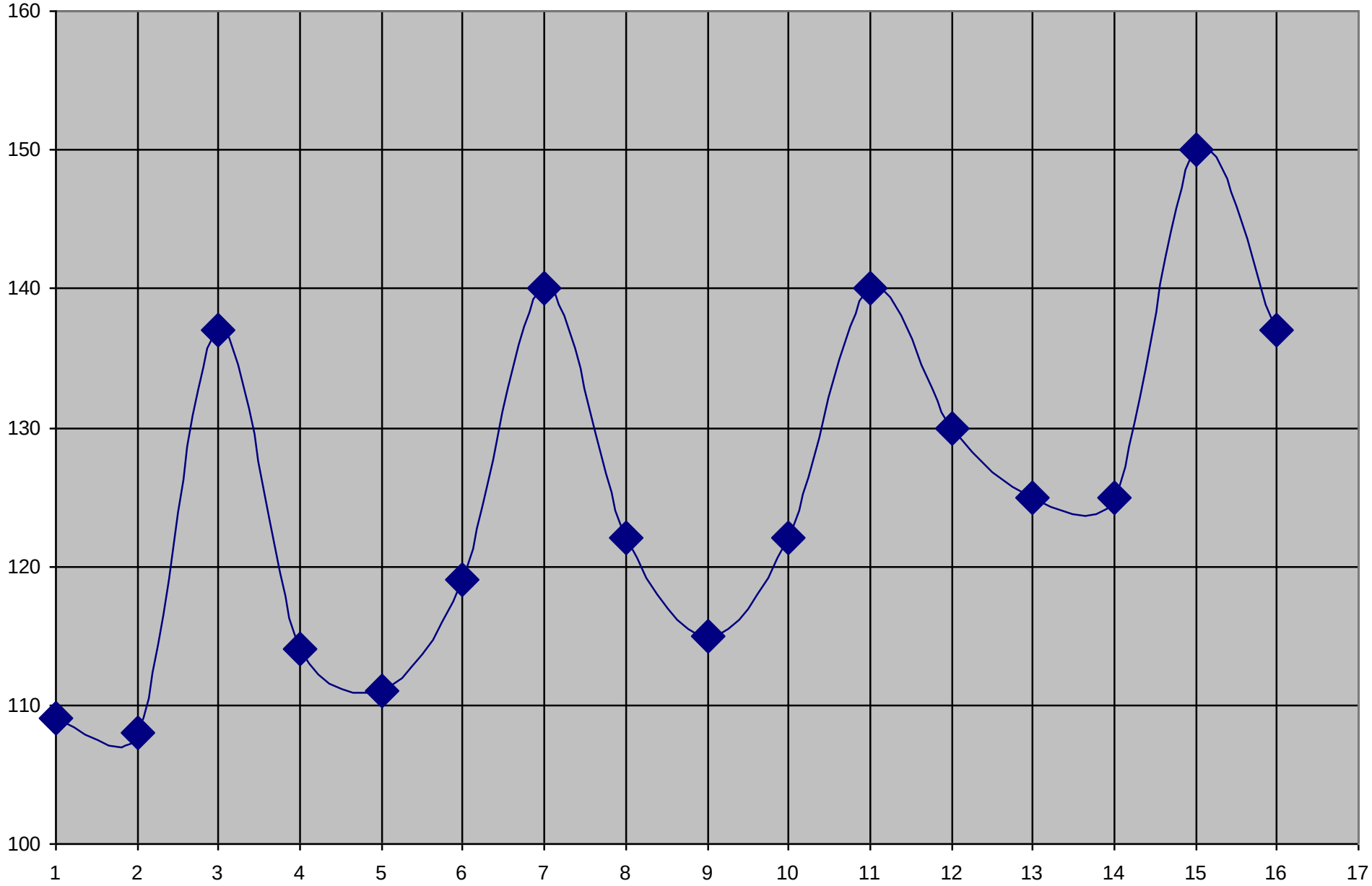
Exemple N°3

Considérons la sérié des indices de la livraison trimestrielle d'essence au Maroc pour années consécutives.

Années	1er trim	2eme trim	3eme trim	4eme trim
1997	109	108	137	114
1998	111	119	140	122
1999	115	122	140	130
2000	125	125	150	137

Travaille à faire

- **Tracer le nuage de points, que fait-il apparaître?**
- **Donner la droite de régression**
- **Quel sera l'indice de la livraison en 2ème trimestre 2011.**


160
150
140
130
120
110
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Commentaire

- **Le graphique fait apparaître une tendance générale à l'augmentation. Il convient de mentionner qu'un mouvement saisonnier se produit chaque année. Un maximum absolu est atteint le troisième trimestre de chaque année (arrivée des travailleurs marocain à l'étranger, touristes.**
- **NB:Pour une bonne interprétation il est nécessaire de dessaisonnaliser cette série.**

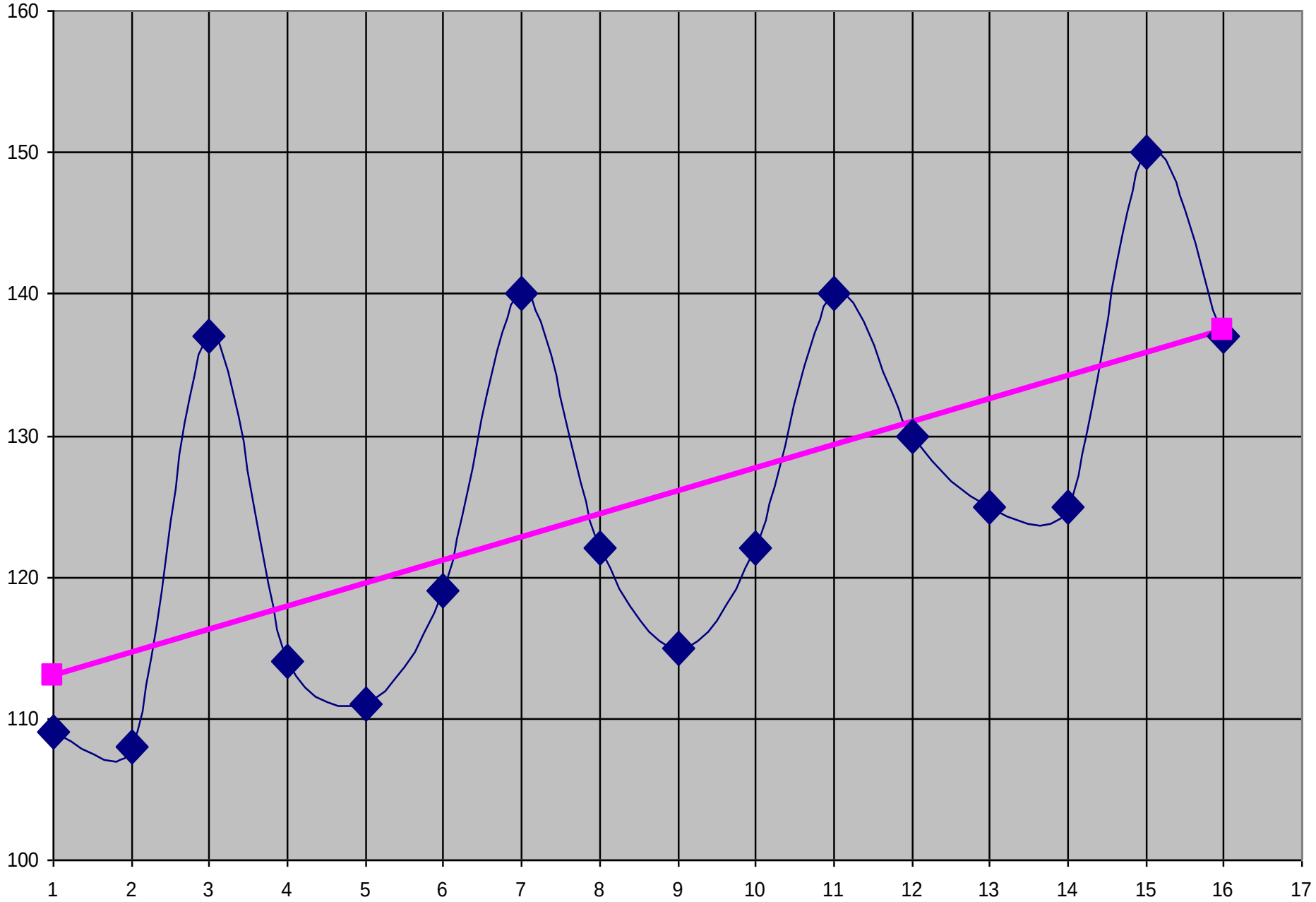
t_i	Y_i	$t_i - \bar{t}$	$y_i - \bar{y}$	$(t_i - \bar{t})^2$	$(t_i - \bar{t})(y_i - \bar{y})$
1	109	-7,5	-16,25	56,25	121,875
2	108	-6,5	-17,25	42,25	112,125
3	137	-5,5	11,75	30,25	-64,625
4	114	-4,5	-11,25	20,25	50,625
5	111	-3,5	-14,25	12,25	49,875
6	119	-2,5	-6,25	6,25	15,625
7	140	-1,5	14,75	2,25	-22,125
8	122	-0,5	-3,25	0,25	1,625
9	115	0,5	-10,25	0,25	-5,125
10	122	1,5	-3,25	2,25	-4,875
11	140	2,5	14,75	6,25	36,875
12	130	3,5	4,75	12,25	16,625
13	125	4,5	-0,25	20,25	-1,125
14	125	5,5	-0,25	30,25	-1,375
15	150	6,5	24,75	42,25	160,875
16	137	7,5	11,75	56,25	88,125
somme	136	2004			340	555
Moy	8,5	125,25				

Détermination des coefficients â et ^b

- **â= 555/340=1,63**
- **^b= 125,25-(1,63 x 8,5)=111,4**

- Donc

$$\hat{Y} = 1{,}63\hat{X} + 111{,}4$$


160
150
140
130
120
110
100
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Calcule de l'indice de la livraison en 2ème trimestre 2011

- On sait que $\hat{Y} = 1{,}63\hat{X} + 111{,}4$

- t= 58
- Alors

$$\hat{Y} = 1{,}63 * 58 + 111{,}4 = 205{,}94$$

Exemple N°4

On reprend les observations consignées dans le tableau « exemple 2 », à propos de la relation entre la consommation globale et le revenu disponible. Déterminer S^2, $S^2\hat{b}$ et $S^2\hat{a}$.

Où: S2 est l'estimateur de la variance de l'erreur
$S^2\hat{a}$ est l'estimateur de la variance de $\hat{a}$
$S^2\hat{b}$ est l'estimateur de la variance de $\hat{b}$

Il faut poser deux questions sur les estimateurs (voir hypothèse 6)

- **Les estimateurs sont-ils sans biais?**
- **Les estimateurs sont-ils convergents?**

Comme on a trouvé que E (â) = a

<u>â est donc un estimateur sans biais de a</u>

De même on remarque que

$$E(\hat{b}) = b$$

Alors $\hat{b}$ est un estimateur sans biais de b

D’autre part on a

$$V(\hat{a}) = \frac{\delta_u^2}{(\sum (X_i - \overline{X})^2}$$

Lorsque n → ∞ alors le dénominateur tend vers ∞et V(â)→ 0
Donc â est convergent

et

$$V(\hat{b}) = \delta_u^2 \left[\frac{1}{n} + \left(\frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right] \right.$$

De même Lorsque n → ∞ V(^b)→ 0 Donc ^b est convergent

alors la distribution des $\hat{u}_i$ (ou $\hat{e}_i$) converge en probabilité vers celle du u_i (ou e_i)

et la valeur

$$S^2 = \frac{\sum e_i^2}{n-2}$$

estimateur de la variance des résidus.

Où

n : le nombre d'observation.

n-2 : le nombre de degré de liberté.

2 : le nombre des paramètres (a et b).

Ce qui signifie que

$$S^2_{\hat{a}} = \frac{S^2}{\sum (X_i - \overline{X})^2}$$

De même

$$S^2_{\hat{b}} = S^2 \left[\frac{1}{n} + \frac{\overline{X}^2}{\sum (X_i - \overline{X})^2} \right]$$

Correction exemple 5

n	Yi	Xi	^Yi	ei, ou ui	ei²	$(X_i - \overline{X})^2$
1	102	114	100,34	1,66	2,76	961
2	106	118	103,78	2,22	4,93	729
3	108	126	110,66	-2,66	7,08	361
4	110	130	114,1	-4,10	16,81	225
5	122	136	119,26	2,74	7,51	81
6	124	140	122,7	1,30	1,69	25
7	128	148	129,58	-1,58	2,50	9
8	130	156	136,46	-6,46	41,73	121
9	142	160	139,9	2,10	4,41	225
10	148	164	143,34	4,66	21,72	361
11	150	170	148,5	1,50	2,25	625
12	154	178	155,38	-1,38	1,90	1089
Somme	**1524**	**1740**		**0,00**	**115,28**	**4812,00**
Moyenne	**127**	**145**				

Selon le tableau

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{115{,}28}{12-2} = 11{,}53$$

$$S^2_{\hat{a}} = \frac{S^2}{\sum (X_i - \bar{X})^2} = \frac{11{,}53}{4812} = 0{,}0024$$

$$S^2_{\hat{b}} = S^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2}\right] = 11{,}53\left(\frac{1}{12} + \frac{(145)^2}{4812}\right) = 51{,}34$$

Selon le tableau

$$S^2 = \frac{\sum e_i^2}{n-2} = \frac{115{,}28}{12-2} = 11.53$$

$$S^2_{\hat{a}} = \frac{S^2}{\sum (X_i - \bar{X})^2} = \frac{11{,}53}{4812} = 0.0024$$

$$S^2_{\hat{b}} = S^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right] = 11{,}53 \left(\frac{1}{12} + \frac{(145)^2}{4812} \right) = 51{,}34$$

$$S_{\hat{a}} = \sqrt{0{,}0024} = 0{,}05$$

$$S\hat{b} = \sqrt{51{,}34} = 7{,}17$$

Pour savoir si la variable exogène choisie est pertinente, on passe au TEST DE STUDENT

- **Test de studente de â =tâ= â/Sâ**
- **avec ddl = n-k = n-2(avec k nombre de paramètre)**

- soit on compare le tâ par rapport à t tabuler
Si tâ> à t tabulé → on rejette Ho (Ho : a =0)
La variable exogène est pertinente .

- Soit on compare tâ à 2, Si tâ> 2→ on rejette Ho (Ho : a =0): La variable exogène est pertinente.

- **Si non la variable exogène n'est pas pertinente.**

Application

- **tâ= â/Sâ = 0,86/0,05 = 17,5.**
  avec ddl = n-k (avec k nombre de paramètre)
  = 12-2=10
  t tabulé à 5% = 2,228
  alors tâ> à t tabulé →17,5>2,228
  on rejette Ho (Ho : a =0)
  La variable exogène est pertinente

COEFFICIENT DE DETERMINATION : (Validation du modèle)

- Le principe de validation d'un modèle consiste à s'interroger sur l'égalité de $(\hat{Y} \equiv Y)$
- et On dit que le modèle reproduit bien la réalité si $(\hat{Y} \equiv Y)$, donc la différence s'explique par d'autres variables (X).
- On ne peut pas dire qu'une variable (X_i) explique un modèle, mais la variation de (X_i) explique la variation de Y.
- Ce qui implique la nécessité de calculer:

$$R^2 = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2} = \frac{\text{variation expliquée}}{\text{variation totale}}$$

R^2 est appelé coefficient de détermination (carré du coefficient de corrélation).

- La valeur de R^2 s'établit entre:
- 0 (l'équation de régression estimée n'explique en rien la variabilité de Y)
- 1 (touts les points (X,Y) appartiennent à la droite de régression)

- Les données du tableau précèdent permettent de calculer $R^2 = 0,96$
- Cela voudrait dire que le modèle permet d'expliquer **96%** de la variabilité de Y.

Le tableau suivant donne l'offre d'un bien en volume, Y, à des prix divers, X, toutes choses restant égales par ailleurs.

Y	12	14	10	13	17	12	11	15
X	5	11	7	8	11	7	6	9

a/ Estimer l'équation de régression de Y par rapport à X.
b/ Tester la signification statistique des estimateurs des paramètres (T de Student).
c/ Déterminer R^2, commenter et résumer tous les résultats précédents sous la forme classique de présentation.

Y	X	Y-mY (1)	X-mX (2)	1*2	$(X-mX)^2$	Y estimé	ei	ei^2
12	5	-1	-3	3	9	10,53	1,47	2,16
14	11	1	3	3	9	15,47	-1,47	2,16
10	7	-3	-1	3	1	12,18	-2,18	4,74
13	8	0	0	0	0	13,00	0,00	0,00
17	11	4	3	12	9	15,47	1,53	2,34
12	7	-1	-1	1	1	12,18	-0,18	0,03
11	6	-2	-2	4	4	11,35	-0,35	0,12
15	9	2	1	2	1	13,82	1,18	1,38
13	8	0	0	28	34		0	12,94

a estimé = 28/34= 0,82

b estimé= 13-(8*0,82) = 6,41 } ^Y = 0,82 X+ 6,41

$S^2 = 12,94/(8-2) = 2,16$

$S^2_a = 2,16/34 = 0,063$

$S^2_b = 2,16(1/8+8^2/34) = 4,3$

$t_{\hat{a}} = 0,82/0,25 = 3,28$

$t_{\hat{}b} = 6,41/2,07 = 3,09$

	Yestim-mY (3)	Y-mY (4)	3^2	4^2
	-2,47	-1,00	6,10	1
	2,47	1,00	6,10	1
	-0,82	-3,00	0,68	9
	0,00	0,00	0,00	0
	2,47	4,00	6,10	16
	-0,82	-1,00	0,68	1
	-1,65	-2,00	2,71	4
	0,82	2,00	0,68	4
somme	0,00	0,00	23,06	36

$R^2 = 23,06/36 = 0,64$