

Régression linéaire simple

Prédire / expliquer les valeurs d'une variable quantitative Y à partir d'une autre variable X

Ricco Rakotomalala
Université Lumière Lyon 2



1. Modèle de Régression linéaire simple



Position du problème

Exemple de régression simple (Bourbonnais, page 12)

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 : points atypiques, etc.)

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 :

$$y_i = a \times x_i + b + \varepsilon_i$$

→ Nous disposons donc d'un échantillon de n couples de points (x_i, y_i) i.i.d (indépendants et identiquement distribués), et **on veut expliquer (prédire) les valeurs de Y en fonction des valeurs prises par X.**

→ **Le terme aléatoire** 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, résumer les variables qui sont absentes, etc.)



Hypothèses

Permettent de déterminer les propriétés des estimateurs

Et de mettre en place les outils de statistique inférentielle (tests d'hypothèses, intervalle de confiance)

H1 : Hypothèses sur X et Y. Ce sont des grandeurs numériques mesurées sans erreur. X est une donnée (exogène) dans le modèle, Y est aléatoire par l'intermédiaire de ε (c.-à-d. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs dans le modèle).

H2 : Hypothèses sur le terme aléatoire. Les ε_i sont i.i.d. (indépendants et identiquement distribués)

(H2.a) En moyenne les erreurs s'annulent, le modèle est bien spécifié $E(\varepsilon_i) = 0$

(H2.b) La variance de l'erreur est constante et ne dépend pas de l'observation : homoscedasticité $V(\varepsilon_i) = \sigma_\varepsilon^2$

(H2.c) En particulier, l'erreur est indépendante de la variable exogène $COV(x_i, \varepsilon_i) = 0$

(H2.d) Indépendance des erreurs, les erreurs relatives à 2 observations sont indépendantes (on dit aussi que les erreurs « ne sont pas corrélées ») $COV(\varepsilon_i, \varepsilon_j) = 0$

(H2.e) Loi normale $\varepsilon_i \equiv N(0, \sigma_\varepsilon)$

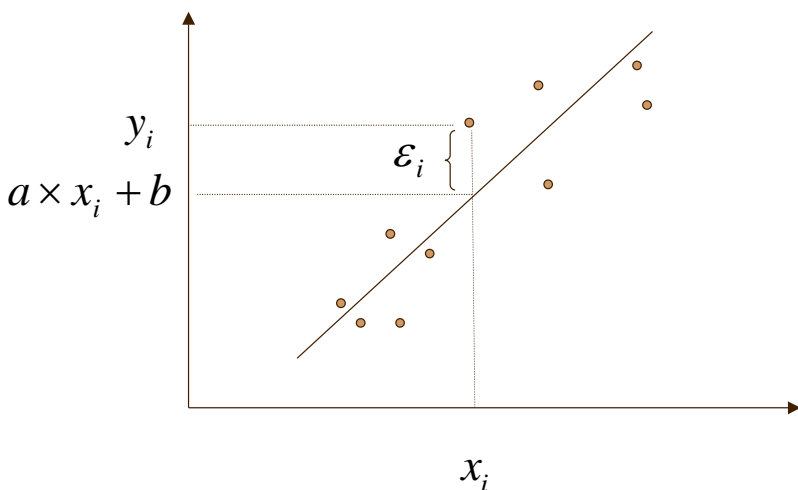


2. Principe de l'ajustement des moindres carrés



Estimateur des MCO (Moindres carrés ordinaires)

Critère numérique



Critère des moindres carrés : trouver les valeurs de **a** et **b** qui **minimise** la somme des carrés des écarts entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.

$$S = \sum_{i=1}^n \epsilon_i^2$$

$$S = \sum_{i=1}^n [y_i - (ax_i + b)]^2$$

$$S = \sum_{i=1}^n [y_i - ax_i - b]^2$$

Remarque : Pourquoi pas la somme des erreurs ? Ou la somme des écarts absolus ?

SOLUTION

$$\begin{cases} \frac{\partial S}{\partial a} = 0 \\ \frac{\partial S}{\partial b} = 0 \end{cases} \quad \rightarrow \quad \begin{cases} \sum_i x_i y_i - a \sum_i x_i^2 - b \sum_i x_i = 0 \\ \bar{y} - a\bar{x} - b = 0 \end{cases}$$

Equations normales

$$\begin{cases} \hat{a} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2} \\ \hat{b} = \bar{y} - \hat{a}\bar{x} \end{cases}$$

Estimateurs des moindres carrés

Voir détail des calculs...

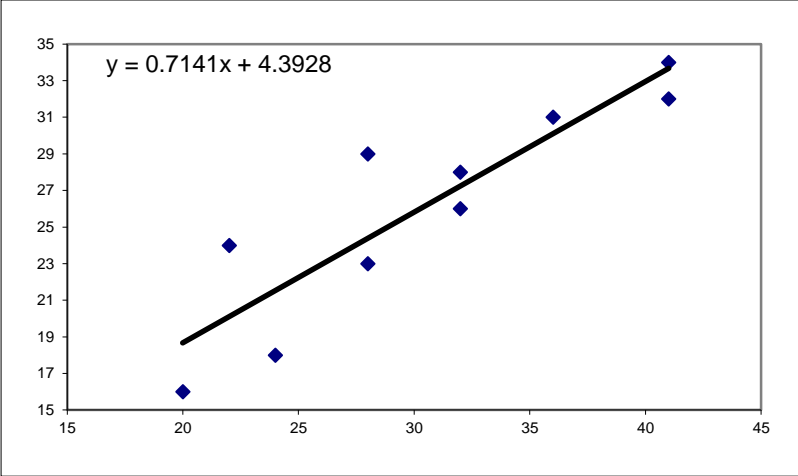
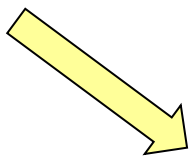


Exemple des rendements agricoles

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB) 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



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



Quelques commentaires

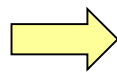
Autre écriture de la pente « a »

$$\hat{a} = \frac{C\hat{O}V(X, Y)}{\hat{\sigma}_X^2} = \hat{r} \frac{\hat{\sigma}_Y}{\hat{\sigma}_X}$$

Relation entre la pente et le coefficient de corrélation linéaire !!!

Erreur et résidus : « **erreur** » = définie dans la spécification du modèle ; « **résidus** », erreurs observées sur les données

$$\begin{aligned} \hat{y}_i &= \hat{y}(x_i) \\ &= \hat{a}x_i + \hat{b} \end{aligned}$$



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

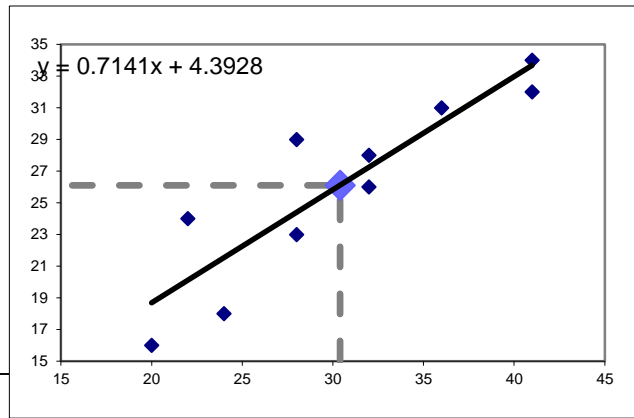
Résidus de la régression

$$\sum_i \hat{\epsilon}_i = 0$$

*Pour la régression avec constante !
Voir détail des calculs...*

Centre de gravité du nuage de points : la droite de régression passe forcément par le barycentre du nuage de points.

$$\begin{aligned} \hat{y}(\bar{x}) &= \hat{a}\bar{x} + \hat{b} \\ &= \hat{a}\bar{x} + (\bar{y} - \hat{a}\bar{x}) \\ &= \bar{y} \end{aligned}$$



3. Evaluation : Analyse de variance et coefficient de détermination



Equation d'analyse de variance

Décomposition de la variance

Objectif de la régression : minimiser S .

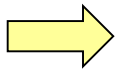
Mais $0 \leq S \leq +\infty$; à partir de quand peut-on dire que la régression est de « bonne qualité » ?

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

Somme des écarts à la moyenne

$$\begin{aligned} \sum_i (y_i - \bar{y})^2 &= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + \underbrace{2 \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y})}_{= 0} \end{aligned}$$

Voir détail des calculs...



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



Coefficient de détermination Et coefficient de corrélation linéaire multiple

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


$$R^2 = \frac{SCE}{SCT} = \frac{\sum_i (\hat{y}_i - \bar{y})^2}{\sum_i (y_i - \bar{y})^2}$$

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

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Coefficient de corrélation
linéaire multiple R

$$R = \sqrt{R^2}$$

On montre que  $r_{Y,X} = \text{sgn}(\hat{a}) \times R$

Lien entre le coefficient de corrélation linéaire (de Pearson) et le coefficient de corrélation linéaire multiple de la régression linéaire simple



Exemple des rendements agricoles

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

$$= 0.714x_i + 4.39$$

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

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB) ²	(Y-YB) ²	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

R² = 0.79727295

R = 0.89290142



4. Evaluation : Propriétés des estimateurs



Les estimateurs sont sans biais si...

$$\begin{aligned} E[\hat{a}] &= a \\ E[\hat{b}] &= b \end{aligned} \quad ?$$

Etape 1 : Exprimer \hat{a} en fonction de a

Voir détail des calculs...

$$\hat{a} = a + \sum_i \omega_i \varepsilon_i$$

où

$$\omega_i = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$$

Etape 2 : Déterminer $E(\hat{a})$ en fonction de a

$$E(\hat{a}) = a + E\left(\sum_i \omega_i \varepsilon_i\right)$$

Etape 3 : Identifier sous quelles conditions $E(\hat{a}) = a$



$$E(\hat{a}) = a + \sum_i \omega_i E(\varepsilon_i)$$

X n'est pas aléatoire par hypothèse, donc ω_i ne l'est pas

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

*$E(\varepsilon_i) = E(\varepsilon)$; les ε_i sont i.i.d.
 $E(\varepsilon) = 0$ par hypothèse*

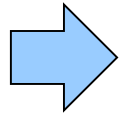


Biais (suite)

Pour « b » $\hat{b} = b + \bar{\varepsilon} - (\hat{a} - a)\bar{x}$

Avec les mêmes hypothèses, on aboutit à

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



Conclusion : Les EMCO (estimateurs des moindres carrés ordinaires) sont sans biais, si...

- Les X ne sont pas stochastiques (non aléatoires)
- $E(\varepsilon) = 0$ c.-à-d. le modèle est bien spécifié



Variance

$$V(\hat{a}) = E[(\hat{a} - a)^2]$$


$$= E\left[\left(\sum_i \omega_i \varepsilon_i\right)^2\right]$$

puisque


$$\hat{a} = a + \sum_i \omega_i \varepsilon_i$$

$$= E\left[\sum_i \omega_i^2 \varepsilon_i^2 + 2 \sum_{i < i'} \omega_i \omega_{i'} \varepsilon_i \varepsilon_{i'}\right]$$

$$= \sum_i \omega_i^2 E(\varepsilon_i^2) + 2 \sum_{i < i'} \omega_i \omega_{i'} E(\varepsilon_i \varepsilon_{i'})$$


$$V(\varepsilon_i) = E(\varepsilon_i^2) = \sigma_\varepsilon^2$$

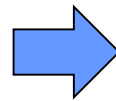
Homoscédasticité


$$E(\varepsilon_i \varepsilon_{i'}) = 0$$

Non-autocorrélation des résidus

avec

$$\omega_i = \frac{(x_i - \bar{x})}{\sum_j (x_j - \bar{x})^2}$$



$$V(\hat{a}) = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$$

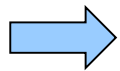


Convergence

$$V(\hat{a}) = \frac{\sigma_{\varepsilon}^2}{\sum_i (x_i - \bar{x})^2}$$

σ_{ε}^2 Est une valeur qui ne dépend pas des effectifs (variance de l'erreur théorique)

$$\sum_i (x_i - \bar{x})^2 \xrightarrow{n \rightarrow +\infty} +\infty$$



\hat{a} est convergent

$$V(\hat{a}) \xrightarrow{n \rightarrow +\infty} 0$$

De même, pour « b »

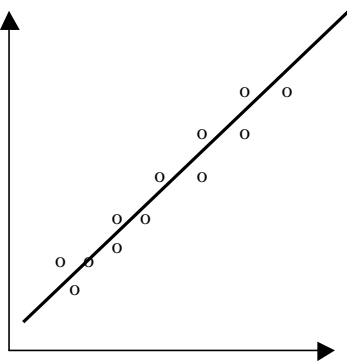
$$V(\hat{b}) = \sigma_{\varepsilon}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i (x_i - \bar{x})^2} \right]$$

$$V(\hat{b}) \xrightarrow{n \rightarrow +\infty} 0$$

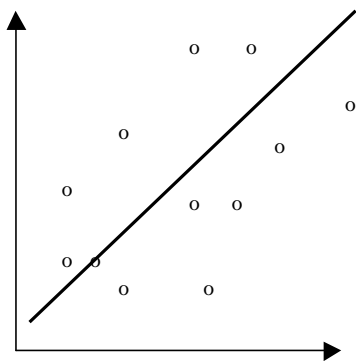


Caractérisation graphique

(1)

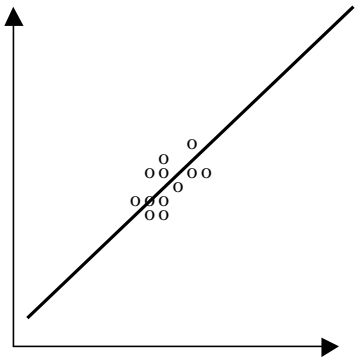


$E(\epsilon_i^2) = \sigma_\epsilon^2$ est faible
 —————> $V(\hat{a})$ est faible, modèle «stable»



$E(\epsilon_i^2) = \sigma_\epsilon^2$ est élevé
 —————> $V(\hat{a})$ est moyennement élevée
 Cette élévation est compensée par $\sum_i (x_i - \bar{x})^2$
 la valeur élevée de

(2)



$E(\epsilon_i^2) = \sigma_\epsilon^2$ est faible } $V(\hat{a})$?
 $\sum_i (x_i - \bar{x})^2$ est faible }

L'adjonction d'un point supplémentaire dans la régression fait « bouger » la droite
 Le modèle est instable également

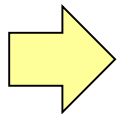
Les estimateurs sont d'autant plus précis que :

- (1) La variance de l'erreur est faible (la droite de régression passe bien au milieu des points.)
- (2) La dispersion des X est forte (les X couvrent bien l'espace de représentation)



Théorème de GAUSS-MARKOV

Les EMCO de la régression sont **sans biais** et **convergent**s.



Parmi les estimateurs sans biais, ils sont à variance minimale c.-à-d. il est impossible de trouver un autre estimateur sans biais à plus petite variance

→ On dit qu'ils sont BLUE (best linear unbiased estimator)

→ Ce sont des « **estimateurs efficaces** »

Cf. démonstration C. Labrousse (1983), page 26



Estimation de la variance de l'erreur

σ_ε^2 Joue un rôle très important. Comment l'estimer à partir des données ?

Le résidu est tel que

$$\begin{aligned}\hat{\varepsilon}_i &= y_i - \hat{y}_i = ax_i + b + \varepsilon_i - (\hat{a}x_i + \hat{b}) \\ &= \varepsilon_i - (\hat{a} - a)x_i - (\hat{b} - b)\end{aligned}$$

On montre que
Giraud & Chaix (1994), page 31

$$E\left(\sum_i \hat{\varepsilon}_i^2\right) = (n-2)\sigma_\varepsilon^2$$

On en déduit un estimateur sans biais



$$\hat{\sigma}_\varepsilon^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n-2} = \frac{SCR}{n-2}$$

Remarque : A propos du degré de liberté (n-2)

Parce 2 contraintes avec les équations normale $\begin{cases} \sum_i x_i \hat{\varepsilon}_i = 0 \\ \sum_i \hat{\varepsilon}_i = 0 \end{cases}$

Parce que (simplement), on estimé 2 paramètres « a » et « b » dans le modèle pour obtenir les prédictions, et donc les résidus



Rendements agricoles

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2	(Y-YB)^2	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	63.83874898
							SCT			SCR

ESTIMATION

a	0.714053615
b	4.392770106

sigma^2(epsilon) 7.979843623

sigma^2(a^)	0.01620602	sigma(a^)	0.127302862
sigma^2(b^)	15.7749386	sigma(b^)	3.971767696

$$\hat{V}(\hat{a}) = \hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_{\epsilon}^2}{\sum_i (x_i - \bar{x})^2} = \frac{SCR / (n - 2)}{492.4}$$

$$= \frac{7.9798}{492.4} = 0.0162$$

$$\hat{\sigma}_{\hat{a}} = \sqrt{\hat{\sigma}_{\hat{a}}^2} = \sqrt{0.0162} = 0.127$$



5. Distribution des coefficients estimés – Inférence statistique



Distribution de « \hat{a} » – Variance de l'erreur connue

$$\hat{a} = \frac{\sum_i (y_i - \bar{y})(x_i - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

X est non aléatoire

Y l'est par l'entremise de ε

$$\varepsilon \equiv N(0, \sigma_\varepsilon)$$

Et « \hat{a} » est issue d'une
combinaison linéaire de Y



$$\frac{\hat{a} - a}{\sigma_{\hat{a}}} \equiv N(0,1)$$

Distribution de l'estimation de la variance de l'erreur

$$\sigma_{\hat{a}}^2 = \frac{\sigma_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$$

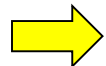


$$\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$$

→ on a besoin de connaître la distribution de $\hat{\sigma}_\varepsilon^2$

Par hypothèse
 $\varepsilon \equiv N(0, \sigma_\varepsilon)$

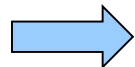
Le résidu étant une réalisation de ε ,
elle suit aussi une loi normale



$$\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \equiv N(0,1)$$



$$\sum_i \left(\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon} \right)^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2)$$



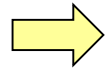
$$(n-2) \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2)$$



Distribution de « \hat{a} » – Variance de l'erreur estimée

On vérifie facilement

$$(n-2) \frac{\hat{\sigma}_{\hat{a}}^2}{\sigma_{\hat{a}}^2} = (n-2) \frac{\hat{\sigma}_{\varepsilon}^2}{\sigma_{\varepsilon}^2}$$



$$(n-2) \frac{\hat{\sigma}_{\hat{a}}^2}{\sigma_{\hat{a}}^2} \equiv \chi^2(n-2)$$

On en déduit dès lors que

$$\frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}} \equiv \mathfrak{T}(n-2)$$

De la même manière, on montre

$$\frac{\hat{b} - b}{\hat{\sigma}_{\hat{b}}} \equiv \mathfrak{T}(n-2)$$

A partir de ces éléments, on peut mettre en place l'inférence statistique



- Intervalle de confiance au niveau $(1 - \alpha)$ $[\hat{a} \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{a}}]$
- Tests d'hypothèses au risque α $\begin{cases} H_0 : a = a_0 \\ H_1 : a \neq a_0 \end{cases}$
- Avec, en particulier le **test de significativité** (*mesurer l'impact de X dans l'explication de Y via le modèle*) $\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$



Rendements agricoles – Tests de significativité des coefficients

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2	(Y-YB)^2	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	63.83874898
							SCT			SCR

sigma^2(epsilon) **7.979843623**

ESTIMATION

a	0.714053615
b	4.392770106

sigma^2(a^)	0.016206019	sigma(a^)	0.127302862
sigma^2(b^)	15.77493863	sigma(b^)	3.971767696

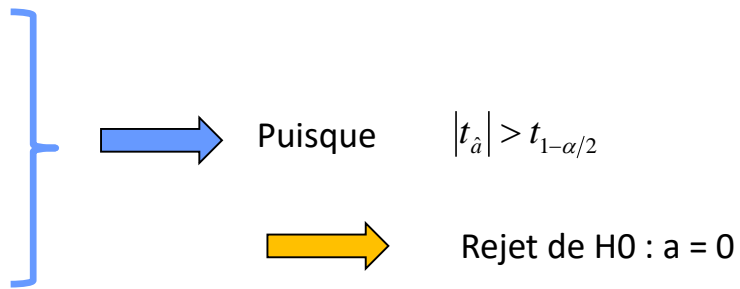
ddl 8

t théorique (bilatéral à 5%) **2.306004133**

t(a^)	5.609093169	rejet H0
t(b^)	1.10599875	

$$t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0.714}{0.127} = 5.609$$

$$t_{1-\alpha/2}(8) = t_{1-0.05/2}(8) = t_{0.975}(8) = 2.306$$



Test de significativité globale du modèle

H0 : Le modèle n'amène rien dans l'explication de Y

H1 : Le modèle est pertinent (globalement significatif)

Tableau d'analyse de variance

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyen
Régression (expliqués)	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$\frac{SCE}{1}$
Résidus	$SCR = \sum_i (\hat{y}_i - y_i)^2$ $= \sum_i \hat{\epsilon}_i^2$	n - 2	$\frac{SCR}{n - 2}$
Total	$SCT = \sum_i (y_i - \bar{y})^2$	n - 1	

Statistique de test

$$F = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}} \equiv F(1, n-2)$$

Remarque : Ecriture de F à partir du R²

$$F = \frac{R^2}{\frac{(1-R^2)}{(n-2)}}$$

Région critique au risque α

$$F > F_{1-\alpha}(1, n-2)$$

Remarque : Tester la significativité de la régression et tester la significativité de la pente sont équivalents dans la régression simple.



Rendements agricoles – Tests de significativité globale

	Y	X	(Y-YB)	(X-XB)	(Y-YB)(X-XB)	(X-XB)^2	(Y-YB)^2	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		63.83874898

SCT

SCR

ESTIMATION

a	0.714053615
b	4.392770106

Tableau d'analyse de variance			
Source de variation	SC	DDL	CM
Expliqués (Régression)	251.061251	1	251.061251
Résidus	63.83874898	8	7.979843623
Total	314.9	9	

F calculé	31.46192618
-----------	-------------

rejet de H0

DDL1	1
DDL2	8
F théorique (à 5%)	5.317655063

$$F = \frac{\frac{SCE}{n-2}}{\frac{SCR}{n-2}} = \frac{251.06}{7.9798} = 31.4619$$

$$F_{1-\alpha}(1,8) = F_{0.95}(1,8) = 5.37655$$

Puisque

$$F > F_{1-\alpha}$$



Rejet de H0 c.-à-d. on conclut que le modèle est globalement significatif

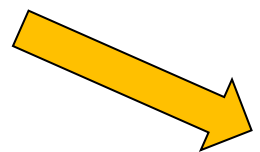
Remarque :

$$\sqrt{F} = \sqrt{31.4619} = 5.609 = t_{\hat{a}}$$



Rendements agricoles – La fonction DROITEREG d'EXCEL

Y	X
16	20
18	24
23	28
24	22
28	32
29	28
26	32
31	36
32	41
34	41



DROITEREG		
\hat{a}	0.71405361	4.392770106 \hat{b}
$\hat{\sigma}_a$	0.12730286	3.971767696 $\hat{\sigma}_b$
R^2	0.79727295	2.8248617 $\hat{\sigma}_\varepsilon$
F	31.4619262	8 $n-2$
SCE	251.061251	63.83874898 SCR

Intervalle de confiance à 5%

t théorique	2.30600413	2.306004133
Borne basse	0.42049269	-4.76614262
Borne haute	1.00761454	13.55168283

Test de significativité des coefficients

t de Student	5.60909317	1.10599875
p-value	0.00050487	0.30087418

Test de la régression globale

F-calculé	31.4619262
DDL numérateur	1
DDL dénominateur	8
p-value	0.00050487



6. Pr evision et intervalle de pr evision



Prévision ponctuelle

A prédire d'une valeur connue de X, prédire la valeur de Y

Pour un individu i^* , la prédiction ponctuelle s'écrit

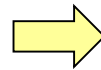
$$\hat{y}_{i^*} = \hat{y}(x_{i^*}) = \hat{a}x_{i^*} + \hat{b}$$

La prédiction est sans biais c.-à-d.

$$E(\hat{y}_{i^*}) = y_{i^*}$$

En effet,

$$\begin{aligned}\hat{\varepsilon}_{i^*} &= \hat{y}_{i^*} - y_{i^*} \\ &= \hat{a}x_{i^*} + \hat{b} - (ax_{i^*} + b + \varepsilon_{i^*}) \\ &= (\hat{a} - a)x_{i^*} + (\hat{b} - b) - \varepsilon_{i^*}\end{aligned}$$



$$\begin{aligned}E(\hat{\varepsilon}_{i^*}) &= E[(\hat{a} - a)x_{i^*} + (\hat{b} - b) - \varepsilon_{i^*}] \\ &= x_{i^*}E(\hat{a} - a) + E(\hat{b} - b) - E(\varepsilon_{i^*})\end{aligned}$$

0 0

Les EMCO sont sans biais

L'erreur du modèle est nulle par hypothèse



Prévision par intervalle

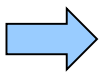
Variance de l'erreur de prévision

Puisque $\hat{\varepsilon}_{i^*} = \hat{y}_{i^*} - y_{i^*}$
 $E(\hat{\varepsilon}_{i^*}) = 0$

On montre
 Giraud & Chaix (1994), page 30

$$V(\hat{\varepsilon}_{i^*}) = E(\hat{\varepsilon}_{i^*}^2) = \sigma_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_{i^*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right] = \sigma_{\hat{\varepsilon}_{i^*}}^2$$

D'où la variance estimée de l'erreur de prévision



$$\hat{\sigma}_{\hat{\varepsilon}_{i^*}}^2 = \hat{\sigma}_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_{i^*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

Remarque :

$$h_{i^*} = \frac{1}{n} + \frac{(x_{i^*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}$$

est le **LEVIER** de l'observation i^*
 (Il joue un rôle très important dans la régression. Cf. points atypiques).

La variance de l'erreur sera d'autant plus faible que :

- (1) $\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-2}$ est petit c.-à-d. la droite ajuste bien le nuage de points .
- (2) $(x_{i^*} - \bar{x})^2$ est petit c.-à-d. le point est proche du centre de gravité du nuage.
- (3) $\sum_i (x_i - \bar{x})^2$ est grand c.-à-d. la dispersion des points est grande.
- (4) n est grand c.-à-d. le nombre d'observations ayant servi à la construction du modèle est élevé.



Prévision par intervalle

Distribution – Définition de l'intervalle

Puisque... $\varepsilon \equiv N(0, \sigma_\varepsilon)$ $\Rightarrow \hat{\varepsilon}_{i^*} = \hat{y}_{i^*} - y_{i^*} \equiv N(0, \sigma_\varepsilon \sqrt{1+h_{i^*}})$

$\Rightarrow (n-2) \frac{\hat{\sigma}_\varepsilon^2}{\sigma_\varepsilon^2} \equiv \chi^2(n-2)$

$\Rightarrow \frac{\hat{y}_{i^*} - y_{i^*}}{\hat{\sigma}_{\hat{\varepsilon}_{i^*}}} \equiv \mathfrak{T}(n-2)$ Rapport d'une loi normale avec un KHI-2 normalisé

$\Rightarrow \hat{y}_{i^*} \pm t_{1-\alpha/2} \times \hat{\sigma}_{\hat{\varepsilon}_{i^*}}$ Intervalle de confiance au niveau $(1-\alpha)$



Rendements agricoles – $x^* = 38$

Prédiction ponctuelle →

$$\hat{y}_{i^*} = \hat{a}x_{i^*} + \hat{b}$$

$$= 0.714 \times 38 + 4.39$$

$$= 31.5268$$

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

ESTIMATION

a	0.714053615
b	4.392770106

x*	38
y^	31.52680747

(x*-xb)^2	57.76
-----------	-------

sigma^2(epsilon^)	9.71389
-------------------	---------

t (0.975)	2.306004133
-----------	-------------

borne.basse	24.33965896
borne.haute	38.71395598

sigma^2(erreur) 7.97984362

Variance de l'erreur de prédiction

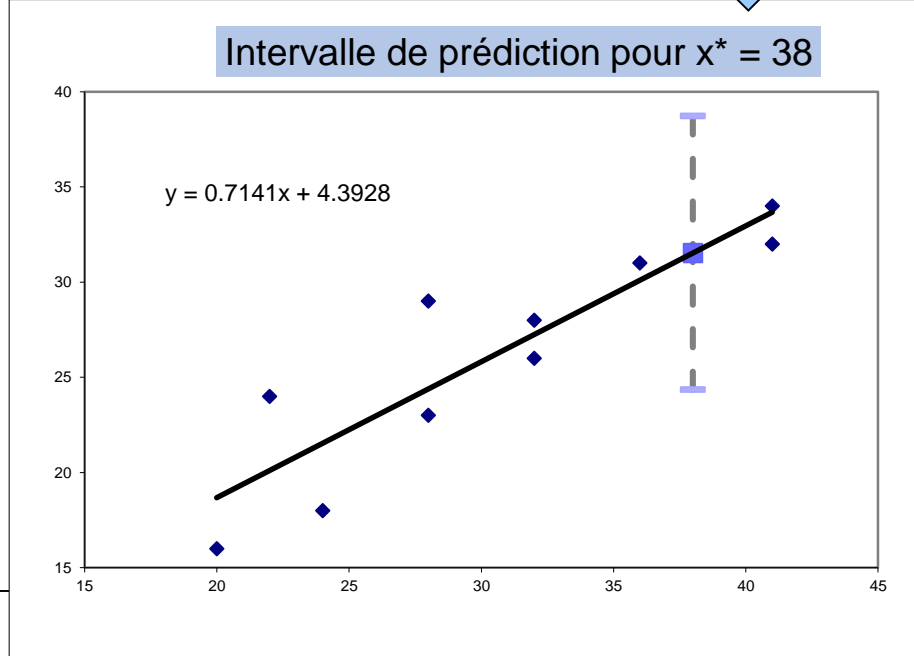
$$\hat{\sigma}_{\hat{y}_{i^*}}^2 = \hat{\sigma}_{\epsilon}^2 \left[1 + \frac{1}{n} + \frac{(x_{i^*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2} \right]$$

$$= 7.9798 \times \left[1 + \frac{1}{10} + \frac{57.76}{492.4} \right]$$

$$= 9.71389$$

$$b.b. = 31.5298 - 2.306 \times \sqrt{9.71389} = 24.3397$$

$$b.h. = 31.5298 + 2.306 \times \sqrt{9.71389} = 38.7140$$



7. Modèles dérivés et interprétation des coefficients

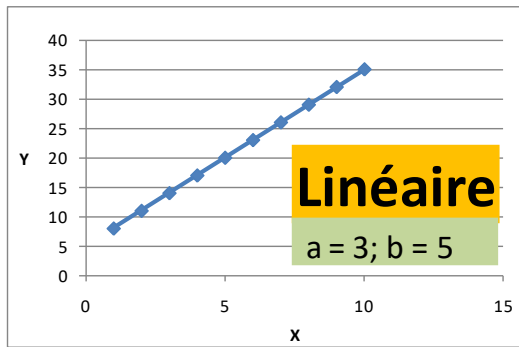


Modèle linéaire

Lecture de la pente

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.

$$Y = aX + b$$



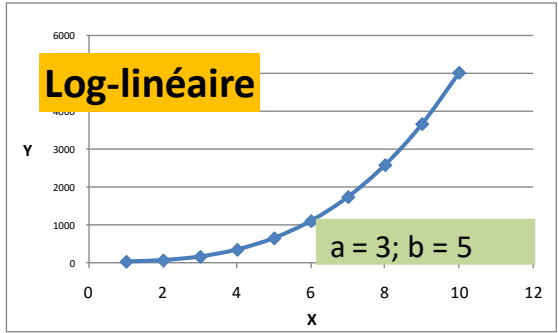
$$\rightarrow a = \frac{dy}{dx}$$

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

$$Y = bX^a$$



$$\rightarrow a = \frac{\frac{dy}{y}}{\frac{dx}{x}}$$

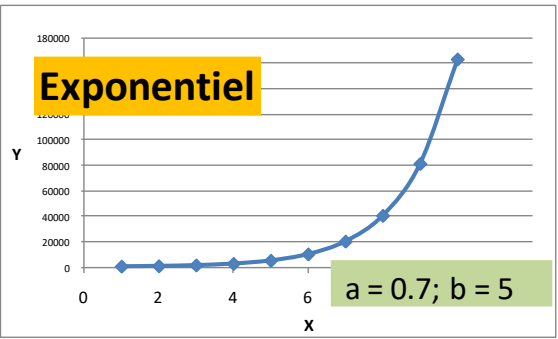
Le taux de variation de Y est proportionnelle au taux de variation de X

- Avantages**
- Modèle à **élasticité** constante : favori des économistes
 - Ex. emploi = f(production), demande = f(prix)
 - Linéarisation : $\ln(y) = a \ln(x) + \ln(b)$



Modèle exponentiel (géométrique)

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



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

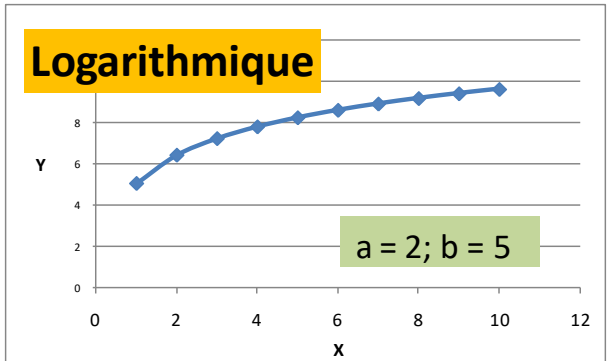
Avantages

- 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 : $\ln(y) = a x + \ln(b)$

➔ $a = \frac{dy/y}{dx}$

Modèle logarithmique

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



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)

➔ $a = \frac{dy}{dx/x}$



Un modèle particulier

Le modèle logistique

3) Un modèle particulier : le modèle logistique

Problème :

Tous les modèles dans (2) ont une concavité constante (dérivée seconde de signe constant), on peut avoir besoin d'un modèle à plusieurs phases

ex : lancement d'un produit dans le temps

Décollage

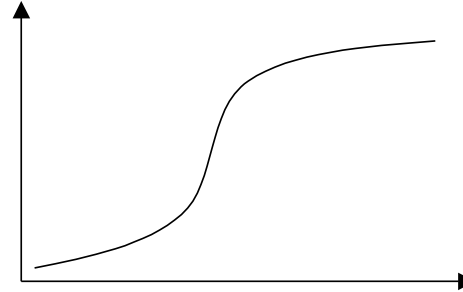
- produit inconnu
- positionnement sur le marché

Croissance accélérée

- large diffusion

Freinage

- saturation du marché
- concurrence



Equation

$$y = y_{\min} + \frac{y_{\max} - y_{\min}}{1 + e^{ax+b}}$$

Linéarisation

$$\ln\left(\frac{y_{\max} - y}{y - y_{\min}}\right) = ax + b$$



8. Régression sans constante



Cas des données centrées

Lorsque les données sont centrées

$$\begin{cases} y_i = y_i - \bar{y} \\ x_i = x_i - \bar{x} \end{cases}$$



La constante est nulle par construction

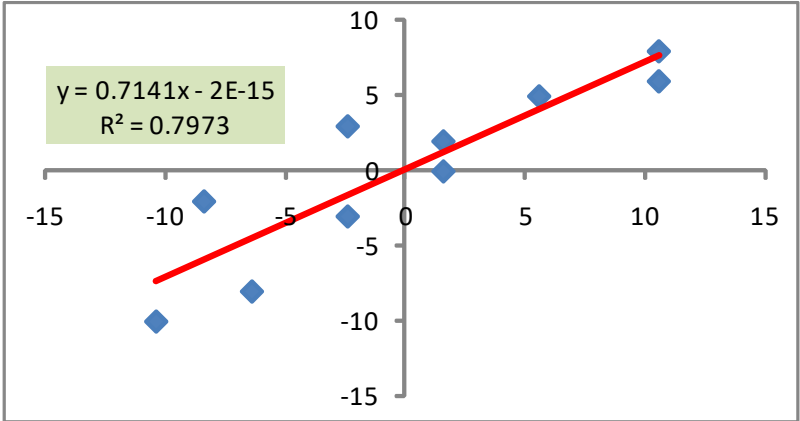
$$\hat{b} = \bar{y} - \hat{a} \times \bar{x} = 0$$

Parce que le barycentre du nuage de points est l'origine du repère c.-à-d.

$$\bar{y} = \bar{x} = 0$$

Y	X	(Y-YB)	(X-XB)
16	20	-10.1	-10.4
18	24	-8.1	-6.4
23	28	-3.1	-2.4
24	22	-2.1	-8.4
28	32	1.9	1.6
29	28	2.9	-2.4
26	32	-0.1	1.6
31	36	4.9	5.6
32	41	5.9	10.6
34	41	7.9	10.6

Moyenne 26.1 30.4



La droite passe forcément par le barycentre, qui se trouve être l'origine (0, 0) du repère.



Cas des données non-centrées

$b = 0 \rightarrow$ on **force** le modèle à passer par l'origine (0,0) du repère

\Rightarrow $y_i = a \times x_i + \varepsilon_i$

On veut minimiser

$$S = \sum_i \varepsilon_i^2 = \sum_i (y_i - a \times x_i)^2$$



Une équation normale

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



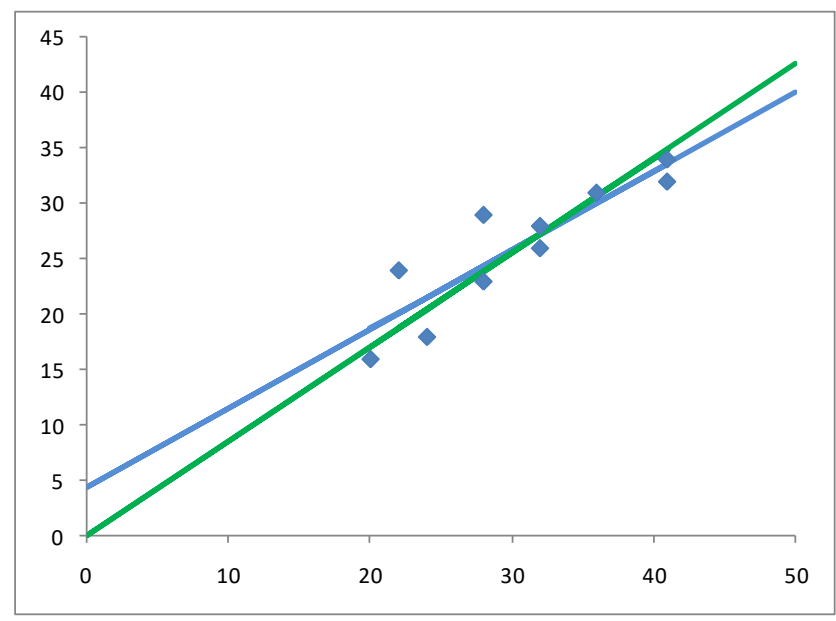
Estimation de la pente

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

Y	X	Y^1	Y^2
16	20	18.6738424	17.0248613
18	24	21.5300569	20.4298336
23	28	24.3862713	23.8348058
24	22	20.1019496	18.7273474
28	32	27.2424858	27.2397781
29	28	24.3862713	23.8348058
26	32	27.2424858	27.2397781
31	36	30.0987002	30.6447504
32	41	33.6689683	34.9009657
34	41	33.6689683	34.9009657
	0	4.39277011	0
	50	40.0954509	42.5621533

Rég.1 - Avec constante
0.71405361 4.39277011
a **b**

Rég.2 - Sans constante
0.85124307
a



Cas des données non-centrées (suite)

A propos du R^2

Le coefficient de détermination R^2 n'a plus de sens parce que : $SCT \neq SCE + SCR$
→ Ca ne sert à rien de le calculer

A propos des degrés de liberté

Un seul paramètre « a » estimé à partir des données → **ddl = n - 1**

→ $\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-1}$ Estimateur sans biais de la variance de l'erreur

→ $\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i)^2}$ Variance de la pente estimée

→ $\frac{\hat{a} - a}{\hat{\sigma}_{\hat{a}}} \equiv \mathfrak{T}(n-1)$ Sert pour les intervalles de confiance
Pour les tests de significativité
Pour les autres tests

A propos des performances

SCR (modèle avec constante) ≤ SCR (modèle sans constante)
→ Pourquoi s'embêter avec un modèle sans constante alors ???
→ Pour les possibilités d'interprétations



Cas des données non-centrées – Un exemple

Comparaison des salaires à l'intérieur des ménages

Numero	Sal.Homme	Sal.Femme
1	7.43	7.20
2	6.83	7.06
3	6.97	7.10
4	7.85	7.39
5	7.48	6.97
6	7.86	7.50
7	7.44	7.16
8	7.83	7.77
9	7.36	7.78
10	7.28	7.47
11	7.53	7.51
12	8.40	8.07
13	7.48	7.25
14	7.46	6.79
15	7.33	7.14
16	7.80	7.38
17	7.57	7.53
18	6.02	6.03
19	7.28	7.05
20	8.42	8.01
21	7.42	7.25
22	7.47	7.59
23	7.14	7.20
24	7.29	6.93
25	8.28	7.85
26	6.98	7.29
27	8.03	7.94
28	7.69	7.11
29	6.67	6.76
30	7.92	7.72

En termes de régression linéaire simple (Y : Sal.H ; X : Sal.F)

$$y_i = a \times x_i + \varepsilon_i$$

Test d'hypothèses
(Attention : test unilatéral) !

$$\begin{cases} H_0 : a = 1 \\ H_1 : a > 1 \end{cases}$$

a^	1.021323921	0
sigma(a)	0.006821202	#N/A
	0.998708093	0.27418841
	22418.42983	29 ddl
	1685.401501	2.18019923

a^-1 0.021323921

t calculé 3.126123666

t-théorique (95%) 1.699126996

Conclusion Rejet de H0

$$t = \frac{\hat{a} - 1}{\hat{\sigma}_{\hat{a}}} = \frac{1.0213 - 1}{0.00682}$$

En moyenne, l'homme a-t-il un salaire plus élevé que sa conjointe dans les ménages (lorsque les deux sont salariés ?)

$$t = 3.126 > t_{1-\alpha}(n-1) = t_{0.95}(29) = 1.699$$



Bibliographique

- R. Bourbonnais, « Économétrie », Dunod, 1998.
- Y.Dodge, V.Rousson, « Analyse de régression appliquée », Dunod, 2004.
- M. Tenenhaus, « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2007.

