Analyse de Données

Par: Houda Benbrahim

Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- Chapitre II: Régression Linéaire Simple et Multiple
- Chapitre III: Analyse Factorielle: ACP
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques

Plan du Cours

- Chapitre I: Rappel: Statistique Descriptive
- **Chapitre II: Régression Linéaire Simple et Multiple**
- Chapitre III: Analyse Factorielle: ACP
- Chapitre IV: Techniques de Classifications
- En parallèle: Travaux Pratiques

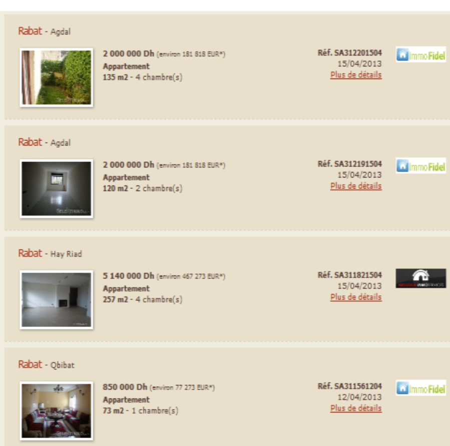
Chapitre II: Régression Linéaire Simple et Multiple

- Section I: Introduction et Définition
- Section II: Régression Linéaire Simple
- Section III: Régression linéaire Multiple
- Section IV: Récapitulatif et Conclusion

Chapitre II: Régression Linéaire Simple et Multiple

Exemple:

Supposant qu'on veut acheter un appartement à Rabat. Une recherche dans les sites immobiliers nous a conduit aux données suivantes:



Exemple:

Table de données:

Quartier	Superficie m^2	Etage	HS	Prix de Vente Dhs
Agdal	135	1	Non	2 000 000
Hay Ryad	257	3	Oui	5 140 000
Hassan	180	5	Oui	3 400 000
Agdal	120	3	Oui	2 000 000
Qbibat	73	2	Non	850 000
Agdal				
Guich				
Hay Ryad				
Agdal				

Exemple: Questions à se poser?

- Comment relier les prix de vente à la superficie des appartements?
- Quelle est le prix de vente espérés si la superficie qu'on désire est 200 m^2 ?
- Est-ce que la surface d'un appartement détermine assez largement son prix?
  - ➢ pas complètement !
  - ➢ autres facteurs à prendre en compte
    - ✓ quartier
    - ✓ étage
    - ✓ Haut standing
    - ✓ présence ascenseur, orientation, parking, gardien, année de construction, etc.

Exemple: Solution:

➔Analyser les données afin d'en dégager des informations nouvelles qui vont fonder des décisions.

➔Prendre des décisions

Exemple: Etude statistique:

- On distingue **2** objectifs:

➔On cherche à savoir s'il existe un lien entre *la superficie et le prix de vente.*

➔ On cherche à savoir si *la superficie a une influence sur le prix de vente et éventuellement* prédire *le prix de vente à partir de la superficie.*

Exemple: Etude statistique:

- On distingue **2** objectifs:

➔On cherche à savoir s'il existe un lien entre *la superficie et le prix de vente.*

➢ Liaison entre *les variables . On définit un indice de liaison : coeff. De* corrélation, statistique du Khi-2,...

➔On cherche à savoir si *la superficie a une influence sur le prix de vente et éventuellement* prédire *le prix de vente à partir de la superficie.*

➢Influence de *variables sur une autre. On modélise cette influence: régression* logistique, **régression linéaire**,...

Régression: Définition

- La régression est une technique de modélisation qui permet de mettre en équation une relation entre une variable endogène (à expliquer) et n variables exogènes (explicatives).

- La régression permet d'analyser la manière dont une variable (dite expliquée) est affectée par les valeurs d'une ou plusieurs autres variables (dites explicatives)

- Un problème de régression consiste à chercher une fonction f telle que pour tout i, Y_i soit approximativement égale à $f(X_i)$.

Régression: Définition

- Un problème de régression consiste à chercher une fonction f telle que pour tout i, Y_i soit approximativement égale à $f(X_i)$.

Variable à expliquer	Variables explicatives	Nom de l'analyse
1 quantitative	1 quantitative	régression simple
1 quantitative	plusieurs quantitatives	régression multiple
1 quantitative	plusieurs qualitatives	analyse de variance
1 quantitative	plusieurs qualitatives et quantitatives	analyse de covariance
1 qualitative	plusieurs quantitatives et qualitatives	régression logistique
1 qualitative	plusieurs quantitatives	analyse discriminante probabiliste

Régression Linéaire: Définition

- Cette technique est couramment utilisée lorsque l'on souhaite prédire la réalisation d'une variable de type continue à l'aide d'un ensemble de variables, dits prédicateurs, du même type.

- La régression linéaire permet de modéliser une relation entre une variable endogène (ou dépendante) Y et p variables exogènes (ou indépendantes) X_1, X_2, …,X_p:

$$Y = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p + \varepsilon$$

- Les β_i sont les coefficients que l'on va chercher à estimer.
- ε est la partie aléatoire que l'on ne peut contrôler. On l'appelle aussi erreur.

Régression Linéaire: Exemple

• On veut représenter la *consommation* d'un agent énergétique en fonction de facteurs explicatifs :

- La température moyenne sur un mois d'un ménage
- L'épaisseur de l'isolation du logement

	Cosommation Gallon/mois	Isolation (en cm)	Température Moyenne (°F)
1	275,30	3,00	40,00
2	363,80	3,00	27,00
3	164,30	10,00	40,00
4	40,80	6,00	73,00
5	94,30	6,00	64,00
6	230,90	6,00	34,00
7	366,70	6,00	9,00
8	300,60	10,00	8,00
9	237,80	10,00	23,00
10	121,40	3,00	63,00
11	31,40	10,00	65,00
12	203,50	6,00	41,00
13	441,10	3,00	21,00
14	323,00	3,00	38,00
15	52,50	10,00	58,00

Régression Linéaire: Exemple

• *consommation* énergétique en fonction de la température moyenne mensuelle et de l'épaisseur de l'isolation du logement.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

Observation i de la Consommation mensuelle

Terme constant

Influence de l'isolation

Influence de la Température

Erreur aléatoire

Régression Linéaire: Applications

- Le modèle de régression linéaire a de nombreuses applications pratiques.

- Il permet de faire des analyses de prédiction:
  - ➢ estimer un modèle de régression linéaire
  - ➢ prédire quel serait le niveau de *y* pour des valeurs particulières de *x*.

- Il permet d'estimer l'effet d'une variable sur une autre contrôlée par d'autres facteurs.
  - ➢ dans le domaine des sciences de l'éducation, on peut évaluer l'effet de la taille des classes sur les performances scolaires des enfants contrôlée par la catégorie socioprofessionnelle des parents ou par l'emplacement géographique de l'établissement.

Régression Linéaire: Applications

- Econométrie:
  - ✓Modèle linéaire pour estimer l'effet du nombre de policiers sur la criminalité.
  - ✓Régression linéaire pour estimer l'effet des institutions sur le développement actuel des pays.
  - ✓Modèle linéaire pour analyser sur des données américaines l'effet des lois autorisant le travail le dimanche sur la participation religieuse.

- Sociologie:
  - ✓ La structure sociale européenne est analysée à l'aide de la régression linéaire entre l' ecart type du niveau de revenu et celui du niveau d'éducation.
  - ✓ La régression linéaire pour évaluer l'estime de soi en fonction du niveau de consommation de cannabis, de l'age et du sexe.

- …

Section I: Introduction et Définition

Section II: Régression Linéaire Simple

Chapitre II: Régression Linéaire Simple et Multiple

Section III: Régression linéaire Multiple

Section IV: Récapitulatif et Conclusion

Régression Linéaire Simple

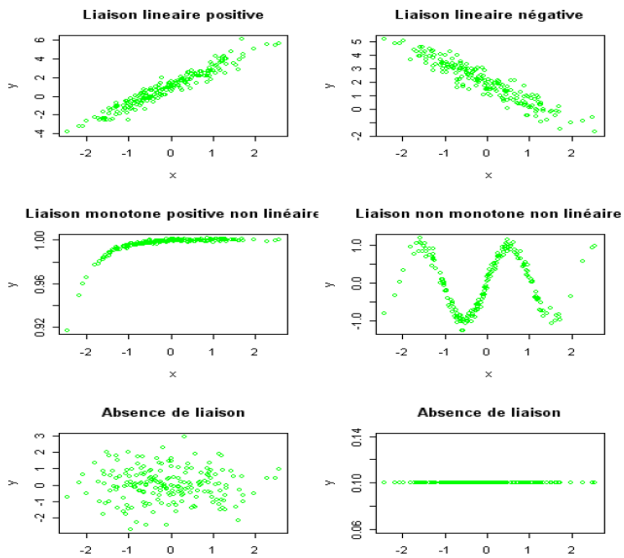
- La relation entre deux variables *x* et *y* est décrite par:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Où β_0 *et* β_1 sont deux constantes que l'on cherche à évaluer et ε est un terme aléatoire que l'on appelle erreur.
Pour estimer β_0 *et* β_1 on dispose d'un échantillon $(x_1,y_1), \ldots, (x_n,y_n)$ supposé vérifier:

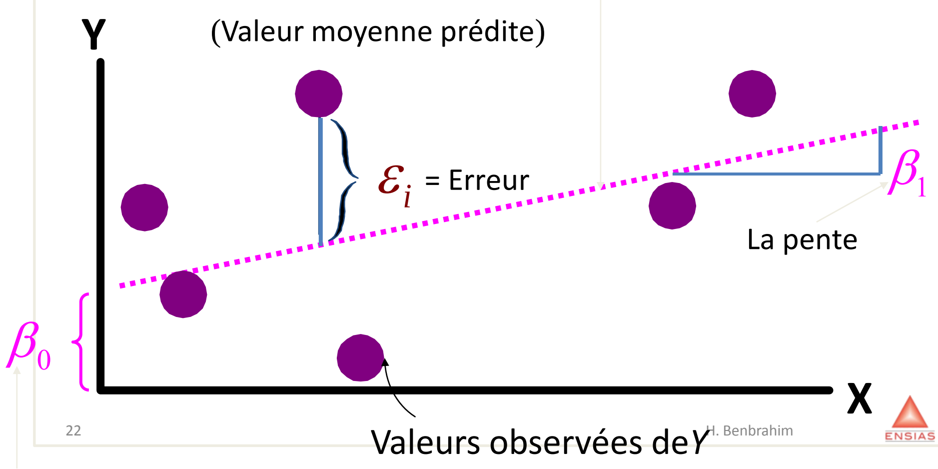
$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad pour\ i = 1,2,\cdots,n.$$

Type de relation

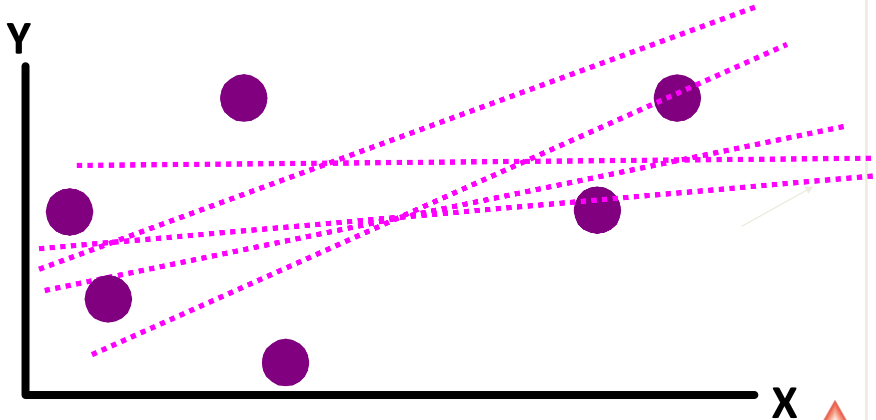


La droite de Régression

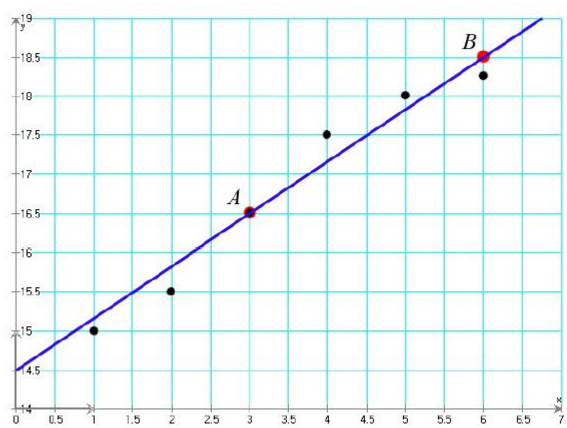
$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$



Mais quelle droite?

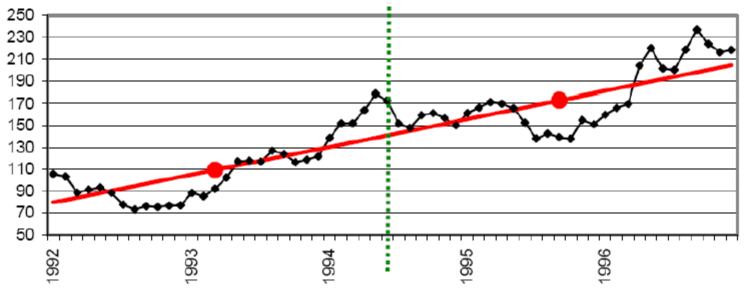


Droite passant par 2 points choisis !



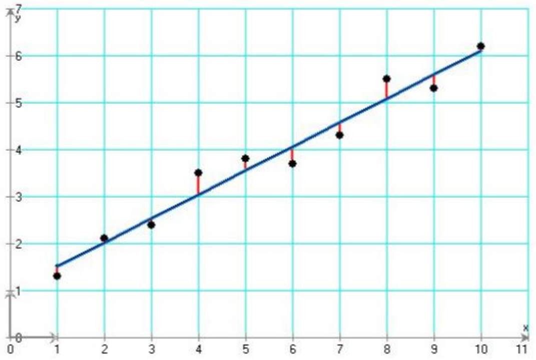
L'équation de la droite passant par les points $A(x_A, y_A)$ et $B(x_B, y_B)$ est donnée par : $y - y_B = \frac{y_B - y_A}{x_B - x_A}(x - x_B)$

Droite de Mayer



On découpe le nuage de points en deux sous-ensembles de même effectif. Pour chacun des deux sous-ensembles, on calcule la moyenne des x_i et la moyenne des y_i. On obtient ainsi deux points $(\bar{x}_1, \bar{y}_1)$ et $(\bar{x}_2, \bar{y}_2)$, appelés **points moyens**. Il reste à tracer la droite passant par ces deux points.

Droite par la méthode des moindres carrés



L'ajustement linéaire par la méthode des moindres carrés consiste à déterminer la droite (que l'on appelle aussi **droite de régression**) telle que la somme des carrés des n valeurs $y_i - \hat{y}_i$ soit minimale (ce qui explique le nom de la méthode).

Modèle de Régression Linéaire Simple

- La relation entre deux variables *x* et *y* est décrite par:

$$y_i = a \times x_i + b + \varepsilon_i, \quad pour\ i \quad = 1, 2, \cdots, n.$$

- ➢ *a et* b sont les paramètres du modèle.
- ➢ a est la pente, b est la constante.
- ➢ ε est l' erreur du modèle.
- ➢ ε résume toute l'information qui n'est pas prise en compte dans la relation linéaire.
- ➢ Les propriétés des estimateurs reposent sur les hypothèses que nous formulons sur ε.

Exemple: Rendement de maïs et quantité d' engrais

- On dispose de *n = 10 observations . On cherche à expliquer Y le rendement* en maïs (en quintal) de parcelles de terrain, à partir de *X la quantité d'engrais (en kg) que l'on y a* épandu.

- L'objectif est de modéliser le lien à travers une relation linéaire.

- si l'on ne met pas d'engrais du tout, il sera quand même possible d'obtenir du maïs, c'est le sens de la constante *b de la régression Sa valeur devrait être positive.*

- *Ensuite, plus on mettra de l'engrais, meilleur sera le* rendement. On suppose que cette relation est linéaire, d'où l'expression *a × x, on imagine à l'avance que a devrait être positif.*

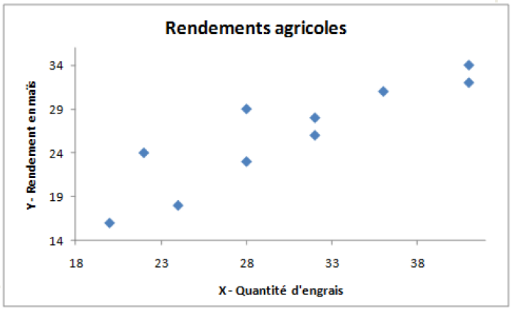
i	Y	X
1	16	20
2	18	24
3	23	28
4	24	22
5	28	32
6	29	28
7	26	32
8	31	36
9	32	41
10	34	41

Exemple: Rendement de maïs et quantité d' engrais

• Le graphique nuage de points associant *X et Y semble confirmer cette première analyse.*

• Dans le cas contraire où les coefficients estimés contredisent les valeurs attendues (*b ou/et a sont négatifs):*

- ➢ une perception faussée du problème,
- ➢ les données utilisées ne sont pas représentatives du phénomène que l'on cherche à mettre en exergue,
- ➢ ou bien...



Hypothèses (1/3)

• Ces hypothèses pèsent sur les propriétés des estimateurs (biais, convergence) et l'inférence statistique (distribution des coefficients estimés).

H1 - Hypothèses sur *Y et X.*

- ✓ *X et Y sont des grandeurs numériques mesurées sans erreur.*
- ✓ *X est une* donnée exogène dans le modèle. Elle est supposée non aléatoire.
- ✓ *Y est aléatoire par l'intermédiaire* de ε *i.e. la seule erreur que l'on a sur Y provient des insuffisances de X à expliquer ses valeurs* dans le modèle.

Hypothèses (2/3)

H2 - Hypothèses sur le terme aléatoire *ε*.

✓ *Les εi sont i.i.d (indépendants et identiquement distribués).*

H2.a *E(εi) = 0*

➢ *en moyenne les erreurs s'annulent c.-à-d. le modèle est bien spécifié.*

H2.b *V (εi) = σ2*

➢ *la variance de l'erreur est constante et ne dépend pas de l'observation.*

➢ *C'est* l'hypothèse d‘ homoscédasticité.

H2.c *COV (xi, εi) = 0*

➢ l'erreur est indépendante de la variable exogène

Hypothèse (3/3)

H2.d *COV (εi, εj) = 0*

➢ Indépendance des erreurs.

➢Les erreurs relatives à 2 observations sont indépendantes.

➢ *On parle de "non auto-corrélation des erreurs".*

H2.e *εi ≡ N(0, σ).*

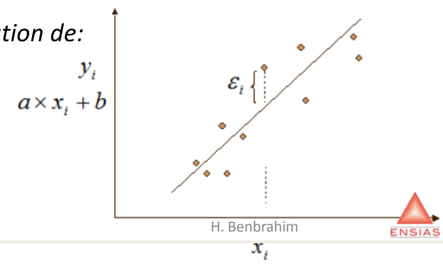
➢ *L'hypothèse de normalité des erreurs est un élément clé pour l'inférence* statistique.

Estimateur des moindres carrés

- Objectif = déterminer les valeurs de *a et b en utilisant les informations apportées par* l'échantillon.
  - ➢ estimation meilleure ➔ la droite de régression doit approcher au mieux le nuage de points.

- Le critère des moindres carrés ➔ minimiser la somme des carrés des écarts (des erreurs) entre les vraies valeurs de Y et les valeurs prédites avec le modèle de prédiction.
- L'estimateur des moindres carrées des paramètres *a et b répond à la minimisation de:*

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



Estimateur des moindres carrés

- S minimum ➔ dérivée première par rapport à a et b sont 0.

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \end{cases}$$

Estimateur des moindres carrés

• Avec les dérivées partielles de S par rapport à a et à b , on obtient le système:

$$\begin{cases} \dfrac{\partial \sum_{i=1}^{n} (y_i - a \cdot x_i - b)}{\partial a} = 0 \\ \dfrac{\partial \sum_{i=1}^{n} (y_i - a \cdot x_i - b)}{\partial b} = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} (y_i - a \cdot x_i - b) \cdot x_i = 0 \\ \sum_{i=1}^{n} (y_i - a \cdot x_i - b) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} x_i \cdot y_i - a \sum_{i=1}^{n} x_i^2 - b \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} a \cdot x_i - \sum_{i=1}^{n} b = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \sum_{i=1}^{n} x_i \cdot y_i = a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} 1 \end{cases}$$

$$\Rightarrow \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i = a \frac{1}{n} \sum_{i=1}^{n} x_i^2 + b \overbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}^{=\bar{x}} \\ \underbrace{\frac{1}{n} \sum_{i=1}^{n} y_i}_{=\bar{y}} = a \underbrace{\frac{1}{n} \sum_{i=1}^{n} x_i}_{=\bar{x}} + b \underbrace{\frac{1}{n} \sum_{i=1}^{n}}_{=1} \end{cases}$$

$$\Rightarrow \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i = a \frac{1}{n} \sum_{i=1}^{n} x_i^2 + b \cdot \bar{x} \\ \bar{y} = a \cdot \bar{x} + b \end{cases}$$

$$\Rightarrow \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i = a \frac{1}{n} \sum_{i=1}^{n} x_i^2 + (\bar{y} - a \cdot \bar{x}) \cdot \bar{x} \\ \bar{y} = a \cdot \bar{x} + b \end{cases}$$

$$\Rightarrow \begin{cases} \frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i = a \frac{1}{n} \sum_{i=1}^{n} x_i^2 + \bar{y} \cdot \bar{x} - a \cdot \bar{x}^2 \\ \bar{y} = a \cdot \bar{x} + b \end{cases}$$

Estimateur des moindres carrés

En appelant ^*a et ^b les solutions de ces équations normales, nous obtenons les estimateurs des* moindres carrés :

$$\Rightarrow \begin{cases} \overbrace{\frac{1}{n} \sum_{i=1}^{n} x_i \cdot y_i - \bar{y} \cdot \bar{x}}^{=cov(x,y)} = a \left(\overbrace{\frac{1}{n} \sum_{i=1}^{n} x_i^2 - \bar{x}^2}^{=\sigma_x^2} \right) \\ \bar{y} = a \cdot \bar{x} + b \end{cases}$$

$$\Rightarrow \begin{cases} \hat{a} = \dfrac{cov(x,y)}{\sigma_x^2} \\ \bar{y} = \hat{a} \cdot \bar{x} + \hat{b} \end{cases}$$

La droite passe donc par le point moyen G de coordonnées : (Ẋ,Ẏ)

Exemple: Rendement Agricole

i	Y	X	(Y-YB)	(X-XB)	(Y-YB)*(X-XB)	(X-XB)²
1	16	20	-10.1	-10.4	105.04	108.16
2	18	24	-8.1	-6.4	51.84	40.96
3	23	28	-3.1	-2.4	7.44	5.76
4	24	22	-2.1	-8.4	17.64	70.56
5	28	32	1.9	1.6	3.04	2.56
6	29	28	2.9	-2.4	-6.96	5.76
7	26	32	-0.1	1.6	-0.16	2.56
8	31	36	4.9	5.6	27.44	31.36
9	32	41	5.9	10.6	62.54	112.36
10	34	41	7.9	10.6	83.74	112.36

Moyenne	26.1	30.4

Somme	351.6	492.4

a^	0.7141
b^	4.3928

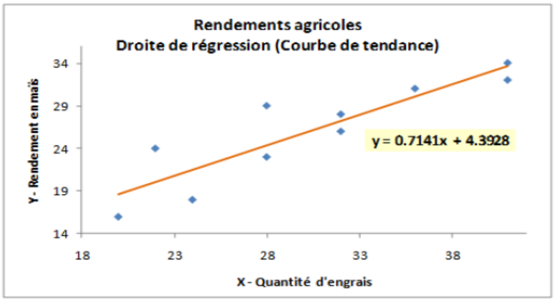
- Nous calculons les moyennes des variables, $\bar{y} = 26.1$ et $\bar{x} = 30.4$.
- Nous formons alors les valeurs de $(y_i - \bar{y})$, $(x_i - \bar{x})$, $(y_i - \bar{y}) \times (x_i - \bar{x})$ et $(x_i - \bar{x})^2$.
- Nous réalisons les sommes $\sum_i (y_i - \bar{y}) \times (x_i - \bar{x}) = 351.6$ et $\sum_i (x_i - \bar{x})^2 = 492.4$.

Exemple: Rendement Agricole

les estimations :

$$\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{351.6}{492.4} = 0.7141$$

$$\hat{b} = \bar{y} - \hat{a}\bar{x} = 26.1 - 0.7141 \times 30.4 = 4.3928$$



Droite de régression - "Rendements agricoles"

Exemple: Rendement Agricole

- On constate que la droite passe plus ou moins au milieu du nuage de points.
- Est-ce que notre modélisation est suffisamment intéressante?
  - Evaluation visuelle ➔ ne suffit pas.
  - Critère quantitatif?

Erreur et Résidu

- *ε est l'erreur inconnue introduite dans la spécification du modèle.*
- la valeur prédite de l'endogène *Y pour l'individu i:*

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

- l'erreur observée, appelée "résidu" de la régression: $\hat{\varepsilon}_i = y_i - \hat{y}_i$
- La somme (et donc la moyenne) des résidus est nulle dans une régression avec constante:

$$\begin{aligned}\sum_i \hat{\varepsilon}_i &= \sum_i [y_i - (\hat{a}x_i + \hat{b})]\\ &= n\bar{y} - n\hat{a}\bar{x} - n\hat{b}\\ &= n\bar{y} - n\hat{a}\bar{x} - n\times(\bar{y} - \hat{a}\bar{x})\\ &= 0\end{aligned}$$

Décomposition de la variance – Equation d'analyse de variance

- L'objectif est de construire des estimateurs qui minimisent la somme des carrés des résidus: $$SCR = \sum_i \hat{\varepsilon}_i^2 = \sum_i (y_i - \hat{y}_i)^2$$
- Prédiction parfaite ➔ *SCR = 0.*
- *Mais dans d'autre cas, qu'est-ce* qu'une bonne régression ?
- A partir de quelle valeur de SCR peut-on dire que la régression est mauvaise ?
- Comparer la SCR avec une valeur de référence ?

Décomposition de la variance – Equation d'analyse de variance

➔décomposer la variance de *Y:*

On appelle *somme des carrés totaux* (SCT) la quantité suivante :

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

Dans la régression avec constante, et uniquement dans ce cas, on montre que

$$2\sum_i (\hat{y}_i - \bar{y})(y_i - \hat{y}_i) = 0$$

Décomposition de la variance – Equation d'analyse de variance

➔décomposer la variance de *Y:*

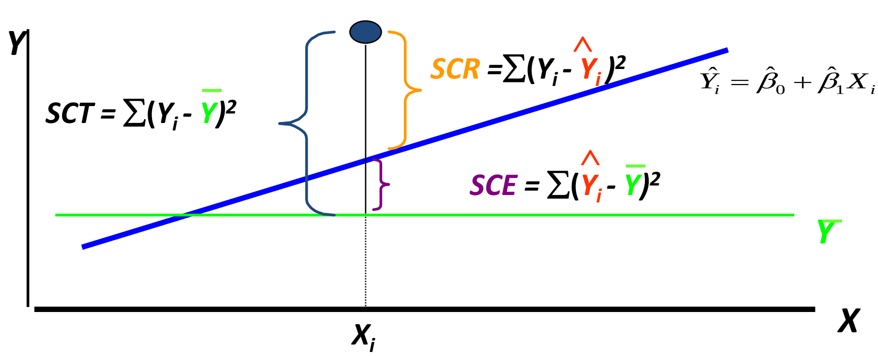
On obtient dès lors l'équation d'analyse de variance :

$$SCT = SCE + SCR$$

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

Décomposition de la variance – Equation d'analyse de variance

➔décomposer la variance de *Y: Interprétation Graphique:*



$$\sum_{i=1}^{N}(Y_i - \overline{Y})^2 = \sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{N}(\hat{Y}_i - \overline{Y})^2$$

Décomposition de la variance – Equation d'analyse de variance

- Comment interpréter ces quantités ?

✓ **SCT** est la somme des carrés totaux. Elle indique la variabilité totale de *Y, i.e. l'information* disponible dans les données.

✓ **SCE** est la somme des carrés expliqués. Elle indique la variabilité expliquée par le modèle, i.e. la variation de *Y expliquée par X.*

✓ **SCR** est la somme des carrés résiduels. Elle indique la variabilité non-expliquée (résiduelle) par le modèle, i.e. l'écart entre les valeurs observées de Y et celles prédites par le modèle.

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | **Modèle** | Propriétés | Inférence | Intervalle de Prédiction | SPSS

Décomposition de la variance – Equation d'analyse de variance

- Situations extrêmes:

✓ Le meilleur des cas: *SCR = 0 → SCT = SCE :*
- *les variations de Y sont complètement* expliquées par celles de *X.*
- *On a un modèle parfait*
- *la droite de régression passe exactement par* tous les points du nuage(*yi^ = yi).*

✓ Le pire des cas: *SCE = 0 :*
- *X n'apporte aucune information sur Y .*
- *yi^ = Ẏ, la meilleure* prédiction de *Y est sa propre moyenne.*

Décomposition de la variance – Equation d'analyse de variance

- Tableau d'analyse de variance:

Source de variation	Somme des carrés
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$

Tableau simplifié d'analyse de variance

Coefficient de détermination

- Coefficient de détermination:

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

- un indicateur synthétique calculé à partir de l'équation d'analyse de variance.
- Il indique la proportion de variance de Y expliquée par le modèle.

- R2 proche de 1:
  → bon modèle,
  → la connaissance des valeurs de *X permet* de deviner avec précision celle de *Y*

- *R2 proche de 0:*
  → *X n'apporte pas d'informations utiles (intéressantes)* sur *Y*
  → *la connaissance des valeurs de X ne nous dit rien sur celles de Y .*

Coefficient de corrélation linéaire multiple

• Le coefficient de corrélation linéaire multiple est la racine carrée du coefficient de détermination: $R = \sqrt{R^2}$

• Dans le cas de la régression simple (et uniquement dans ce cas), R est le coefficient de corrélation *ryx entre Y et X. Son signe est déni par la pente a^ de la régression:* $r_{yx} = signe(\hat{a}) \times R$

Exemple: Rendement Agricole

i	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

Moyenne	26.1

a^	0.71405
b^	4.39277

Y^	epsilon^
18.674	-2.674
21.530	-3.530
24.386	-1.386
20.102	3.898
27.242	0.758
24.386	4.614
27.242	-1.242
30.099	0.901
33.669	-1.669
33.669	0.331

(Y-YB)²	(Y^-YB)²	(Y-Y^)²
102.010	55.148	7.149
65.610	20.884	12.461
9.610	2.937	1.922
4.410	35.977	15.195
3.610	1.305	0.574
8.410	2.937	21.286
0.010	1.305	1.544
24.010	15.990	0.812
34.810	57.289	2.785
62.410	57.289	0.110
314.900	251.061	63.839
SCT	SCE	SCR

R²	0.797273
Racine(R²)	0.892901
Correl(y,x)	0.892901

Propriété des estimateurs

- Deux propriétés importantes pour l'évaluation d'un estimateur:

1. Est-ce qu'il est sans biais ? i.e. est-ce qu'en moyenne nous obtenons la vraie valeur du paramètre ?
2. Est-ce qu'il est convergent? i.e. à mesure que la taille de l'échantillon augmente, l'estimation devient de plus en plus précise ?

Propriété des estimateurs: Biais

- On dit que *ϑ^ est un estimateur sans biais de ϑ si E[ϑ^] = ϑ.*
- Comment procéder à cette vérification pour *a^ et b^ ?*

➔ a^ et b^ sont sans biais, si et seulement si les deux hypothèses suivantes sont respectées :

H1 L'exogène *X n'est pas stochastique (X est non aléatoire) ;*

H2.a *E(εi) = 0, l'espérance de l'erreur est nulle.*

Propriété des estimateurs: variance - convergence

- Un estimateur ϑ^ *sans biais de ϑ est convergent si et seulement si* $V(\hat{\theta}) \underset{n\to\infty}{\longrightarrow} 0$
- La variance est *V (a^) = E[(a^ − a)²]*

➔ a^ et b^ sont convergents, si et seulement si les deux hypothèses suivantes sont respectées :

H2.b *V (εi) = σ² . C'est l'hypothèse d'homoscédasticité.*

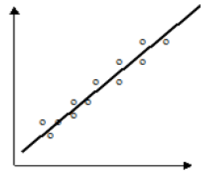
H2.d *COV (εi,εi′) = E(εi,εi′) = 0. C'est l'hypothèse de non-autocorrélation des erreurs.*

Propriété des estimateurs: remarques

- Estimateurs plus précis → les variances plus petites :
  - La variance de l'erreur est faible, i.e. la régression est de bonne qualité.
  - La dispersion des *X est forte , i.e. les points recouvrent bien l'espace de représentation.*
  - Le nombre d'observations *n est élevé.*

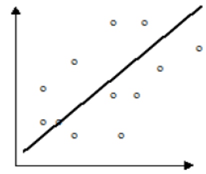
Propriété des estimateurs: remarques

(1)



$E(\varepsilon_i^2) = \sigma_\varepsilon^2$ est faible

$\longrightarrow$ $V(\hat{a})$ est faible, modèle « stable »

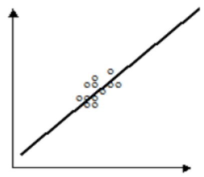


$E(\varepsilon_i^2) = \sigma_\varepsilon^2$ est élevé

$\longrightarrow$ $V(\hat{a})$ est moyennement élevée

Cette élévation est compensée par la valeur élevée de $\sum_i (x_i - \bar{x})^2$

(2)



$E(\varepsilon_i^2) = \sigma_\varepsilon^2$ est faible

$\sum_i (x_i - \bar{x})^2$ est faible

$\Big\}$ $V(\hat{a})$?

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Intervalle de Prédiction | SPSS

Théorème de Gauss - Markov

- Les estimateurs des Moindres carrés de la régression sont sans biais et convergents.

- Parmi les estimateurs linéaires sans biais de la régression, les estimateurs MC sont à variance minimale, i.e. il n'existe pas d'autres estimateurs linéaires sans biais présentant une plus petite variance.

- Les estimateurs des MC sont BLUE (best linear unbiased estimator).

Evaluation globale de la régression

- Evaluation de la qualité de l' ajustement:
  - ✓ la décomposition de la variance
  - ✓ le coefficient de détermination *R2, i.e. dans quelle proportion la variabilité de Y pouvait être expliquée par X.*

- *Est-ce que la régression est* globalement significative ?
  - ✓ Est-ce que *X* emmènent **significativement** de l'information sur *Y*
  - ✓ *Est-ce que R2 est représentative d'une relation* linéaire réelle dans la **population**, ou bien juste une simple fluctuation d'échantillonnage ?

- Ou bien: considérer le test d'évaluation globale comme un test de significativité de *R2 :*
  - ✓ *dans quelle mesure R2 calculé sur un échantillon s'écarte réellement de la valeur 0 ?*

Tableau d'analyse de Variance - Test de significativité globale

Source de variation	Somme des carrés	Degrés de liberté	Carrés moyens
Expliquée	$SCE = \sum_i (\hat{y}_i - \bar{y})^2$	1	$CME = \frac{SCE}{1}$
Résiduelle	$SCR = \sum_i (y_i - \hat{y}_i)^2$	$n-2$	$CMR = \frac{SCR}{n-2}$
Totale	$SCT = \sum_i (y_i - \bar{y})^2$	$n-1$	-

Tableau d'analyse de variance pour la régression simple

- Degrés De Liberté: le nombre de termes impliqués dans les sommes (le nombre d'observations) moins le nombre de paramètres estimés dans cette somme.

  - ➢SCT: estimation de la moyenne *y* → *DDL = n-1.*
  - ➢SCR: coefficients estimés *^a et ^b pour obtenir la projection ^yi* → *DDL = n-2.*
  - ➢SCE: SCT – SCR → DDL = *(n – 1) – (n – 2) = 1.*

Test de significativité globale de la régression

- La statistique **F**:

$$F = \frac{CME}{CMR} = \frac{\frac{SCE}{1}}{\frac{SCR}{n-2}}$$

➢ Cette statistique indique si la variance expliquée est significativement supérieure à la variance résiduelle.

➢ l'explication emmenée par la régression traduit une relation qui existe réellement dans la population.

- La statistique F:

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

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | **Inférence** | Intervalle de Prédiction | SPSS

Test de significativité globale de la régression

- Distribution sous H0*:*
  - ✓ *SCE est distribué selon un χ2(1)*
  - ✓ *SCR est distribué selon un χ2(n – 2)*
  - ➔

$$F \equiv \frac{\frac{\chi^2(1)}{1}}{\frac{\chi^2(n-2)}{n-2}} \equiv \mathcal{F}(1, n-2)$$

  - ✓ Sous H0, F est donc distribué selon une loi de Fisher à (1*, n – 2) degrés de liberté.*

- La région critique du test:
  - ✓ correspondant au rejet de H0
  - ✓ au risque *α est définie pour les valeurs* anormalement élevées de F
  - ➔ *R.C. : F > $F_{1-\alpha}$(1, n – 2)*

- Décision à partir de la p-value:
  - ✓ la probabilité critique (p-value) *α ' = probabilité que la loi de Fisher dépasse la* statistique calculée F.
  - ✓ la règle de décision au risque *α: R.C. : α' < α*

Exemple : les rendements agricoles

i	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

Y^	epsilon^
18.674	-2.674
21.530	-3.530
24.386	-1.386
20.102	3.898
27.242	0.758
24.386	4.614
27.242	-1.242
30.099	0.901
33.669	-1.669
33.669	0.331

(Y-YB)²	(Y^-YB)²	(Y-Y^)²
102.010	55.148	7.149
65.610	20.884	12.461
9.610	2.937	1.922
4.410	35.977	15.195
3.610	1.305	0.574
8.410	2.937	21.286
0.010	1.305	1.544
24.010	15.990	0.812
34.810	57.289	2.785
62.410	57.289	0.110
314.900	251.061	63.839
SCT	SCE	SCR

Moyenne	26.1

a^	0.71405
b^	4.39277

Tableau d'analyse de variance

Source	SC	DDL	Carrés Moyens
Expliquée	251.061	1	251.061
Résiduelle	63.839	8	7.980
Totale	314.900	9	

F	31.462
ddl1	1
ddl2	8
F_0.95	5.318
p-value	0.00050487

Conclusion : Le modèle est globalement significatif au risque 5%

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | **Inférence** | Intervalle de Prédiction | SPSS

Exemple : les rendements agricoles

- Détail des calculs:

✓ $CME = \frac{SCE}{1} = \frac{251.061}{1} = 251.061$

✓ $CMR = \frac{SCR}{n-2} = \frac{63.839}{10-2} = 7.980$

✓ $F = \frac{CME}{CMR} = \frac{251.061}{7.980} = 31.462$

✓ nous comparons au quantile d'ordre $(1 - \alpha)$ *de la loi* $F(1, n-2)$.
✓*Pour* $\alpha = 5\%$*, elle est* égale à $F_{0.95}(1, 8) = 5.318$.
➔*Nous concluons que le modèle est globalement significatif au risque* 5%.
➔La relation linéaire entre *Y et X est représentatif d'un phénomène existant réellement dans la* population.

✓ la probabilité critique: $\alpha' \approx 0.00050$*, inférieure à* $\alpha = 5\%$.
➔*La conclusion* est la même.

➔Il ne peut pas y avoir de contradictions entre ces deux visions.

Distribution des paramètres estimés

- Pour étudier les coefficients estimés:

➔ calculer l'espérance et la variance des paramètres
➔ déterminer la loi de distribution.
➔ statistique inférentielle:
  - la définition des intervalles de variation à un niveau de confiance donné
  - la mise en place des tests d'hypothèses → les tests de significativité.

Distribution de a^

•On a: $\hat{a} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$ et $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$

- *yi = axi +b+εi suit aussi une loi normale, et a^ étant une combinaison linéaire des yi*

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

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

- mais σ_ε^2= ?

→ Pour obtenir une estimation calculable sur un échantillon de données de l'écart-type $\hat{\sigma}_{\hat{a}}$ *du coefficient a^ → produire une estimation de l'écart* type de l'erreur $\hat{\sigma}_\varepsilon$›
La variance estimée: $\hat{\sigma}_{\hat{a}}^2 = \frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}$

Distribution de b^

- de même on a: $\dfrac{\hat{b}-b}{\sigma_{\hat{b}}} \equiv \mathcal{N}(0,1)$
- et: $\sigma_{\hat{b}}^2 = \sigma_\varepsilon^2 \left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_i (x_i-\bar{x})^2}\right]$

Estimateur sans biais de la variance de l'erreur

- on a le résidu: $$\begin{aligned}\hat{\varepsilon}_i &= y_i - \hat{y}_i \\ &= ax_i + b + \varepsilon_i - (\hat{a}x_i + \hat{b}) \\ &= \varepsilon_i - (\hat{a}-a)x_i - (\hat{b}-b)\end{aligned}$$
- on montre que: $E\left[\sum_i \hat{\varepsilon}_i^2\right] = (n-2)\sigma_\varepsilon^2$
- Estimateur sans biais: $\hat{\sigma}_\varepsilon^2 = \dfrac{\sum_i \hat{\varepsilon}_i^2}{n-2} = \dfrac{SCR}{n-2}$

Distribution de la variance de l'erreur

- On a par hypothèse: $\varepsilon_i \equiv \mathcal{N}(0, \sigma_\varepsilon)$

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

- $$\left(\frac{\hat{\varepsilon}_i}{\sigma_\varepsilon}\right)^2 \equiv \chi^2(1)$$

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

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

Distribution de a^ et b^:

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

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

➔ $\frac{\hat{a}-a}{\hat{\sigma}_{\hat{a}}} \equiv \mathcal{T}(n-2)$, de même: $\frac{\hat{b}-b}{\hat{\sigma}_{\hat{b}}} \equiv \mathcal{T}(n-2)$

- loi de Student est définie par un rapport entre une loi normale et la racine carrée d'une loi du $\chi2$ *normalisée par ses degrés de liberté.*

→
$$\frac{\frac{\hat{a}-a}{\sigma_{\hat{a}}}}{\frac{\hat{\sigma}_{\hat{a}}}{\sigma_{\hat{a}}}} \equiv \frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2(n-2)}{n-2}}}$$

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

Test de significativité de a:

- Le test de significativité de la pente a consiste à vérifier l'influence réelle de l'exogène *X sur l'endogène Y .*
- *Les hypothèses à confronter :* $\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$
- Statistique de test: $t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}}$

  suit une loi de Student à (*n − 2) degrés de liberté.*

- La région critique (de rejet de H0) au risque *α s'écrit:* $R.C. : |t_{\hat{a}}| > t_{1-\frac{\alpha}{2}}$

Où $t_{1-\frac{\alpha}{2}}$ est le quantile d'ordre $(1 - \frac{\alpha}{2})$ de la loi de Student. Il s'agit d'un test bilatéral.

Exemple: Rendement Agricole:

i	Y	X	Y^	epsilon^	(epsilon^)²	(X-XB)²
1	16	20	18.674	-2.674	7.149	108.16
2	18	24	21.530	-3.530	12.461	40.96
3	23	28	24.386	-1.386	1.922	5.76
4	24	22	20.102	3.898	15.195	70.56
5	28	32	27.242	0.758	0.574	2.56
6	29	28	24.386	4.614	21.286	5.76
7	26	32	27.242	-1.242	1.544	2.56
8	31	36	30.099	0.901	0.812	31.36
9	32	41	33.669	-1.669	2.785	112.36
10	34	41	33.669	0.331	0.110	112.36

Moyenne	26.1	30.4

a^	0.71405
b^	4.39277

SCR	63.839
(sigma^)²_eps	7.980
sigma^(eps)	2.825

Somme	492.4

Exemple: Rendement Agricole:

- l'estimation de la variance de l'erreur:

$$\hat{\sigma}_\varepsilon^2 = \frac{SCR}{n-2} = \frac{63.839}{8} = 7.980 \qquad \hat{\sigma}_\varepsilon = \sqrt{7.980} = 2.825$$

- $$\begin{aligned}\hat{\sigma}_{\hat{a}} &= \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_i (x_i - \bar{x})^2}} \\ &= \sqrt{\frac{7.980}{492.4}} \\ &= \sqrt{0.01621} \\ &= 0.12730\end{aligned} \qquad t_{\hat{a}} = \frac{\hat{a}}{\hat{\sigma}_{\hat{a}}} = \frac{0.71405}{12730} = 5.60909$$

Au risque $\alpha = 5\%$, le seuil critique pour la loi de Student à $(n-2)$ degrés de liberté pour un test bilatéral[6] est $t_{1-\frac{\alpha}{2}} = 2.30600$. Puisque $|5.60909| > 2.30600$, nous concluons que la pente est significativement non nulle au risque 5%.

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Intervalle de Prédiction | SPSS

Exemple: Rendement Agricole:

ESTIMATION

a	0.714053615
b	4.392770106

		sigma²(epsilon)	7.979843623
sigma²(a^)	0.016206019	sigma(a^)	0.127302862
sigma²(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	acceptation H0

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

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

Puisque $|t_{\hat{a}}| > t_{1-\alpha/2}$

Rejet de H0 : a = 0

Intervalle de confiance de la droite de la régression

- Les coefficients formant le modèle sont entachées d'incertitude

→la droite de régression l' est également.

➔ L'objectif est de produire un intervalle de confiance de la droite de régression.

↔ au calcul de l'intervalle de confiance de la prédiction de la moyenne de *Y conditionnellement X .*

✓*C'est* l'intervalle de confiance de ce que l'on a modélisé avec la droite

✓ à ne pas confondre avec l'intervalle de confiance d'une prédiction lorsque l'on fourni la valeur *xi∗ pour un nouvel individu i∗ n'appartenant* pas à l'échantillon.

- L''intervalle de confiance au niveau (1−*α) de la droite* de régression:

$$\hat{a} \times x_i + \hat{b} \pm t_{1-\frac{\alpha}{2}} \times \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{x_i - \bar{x}}{\sum_j (x_j - \bar{x})^2}}$$

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | **Inférence** | Intervalle de Prédiction | SPSS

Exemple: Rendement Agricole:

i	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
n = 10	34	41
Moyenne		30.4

Y^	epsilon^	(epsilon^)²
18.674	-2.674	7.149
21.530	-3.530	12.461
24.386	-1.386	1.922
20.102	3.898	15.195
27.242	0.758	0.574
24.386	4.614	21.286
27.242	-1.242	1.544
30.099	0.901	0.812
33.669	-1.669	2.785
33.669	0.331	0.110

(X-XB) ^2
108.16
40.96
5.76
70.56
2.56
5.76
2.56
31.36
112.36
112.36
Somme 492.4

b.basse	b.haute
14.99	22.36
18.74	24.32
22.21	26.56
16.89	23.32
25.13	29.36
22.21	26.56
25.13	29.36
27.46	32.73
29.94	37.40
29.94	37.40

SCR	63.8387
sigma^(epsilon)	2.8249

n	10
a^	0.71405
b^	4.39277
t_0.975(8)	2.30600

Exemple: Rendement Agricole:

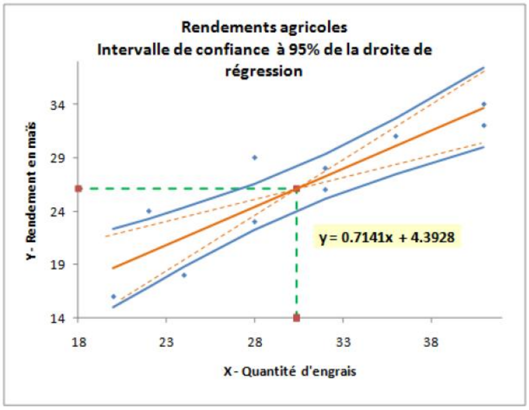
$$b.b.(\mu_{Y/X=x_1}) = 18.674 - 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 14.99$$

$$b.h.(\mu_{Y/X=x_1}) = 18.674 + 2.30600 \times 2.8249 \times \sqrt{\frac{1}{10} + \frac{(20-30.4)^2}{492.4}} = 22.36$$

- **Il y a 95% de chances que la droite soit comprise entre les deux courbes bleues.**

- **La droite ne peut être placée n'importe où dans la zone délimitée.**

- **elle pivote forcément autour du barycentre.**



Prédiction et Intervalle de Prédiction

- Régression?
  - ➢ analyse structurelle
  - ➢ interprétation des coefficient
  - ➢ utilisée pour la prédiction ou prévision:
    - ✓Pour un nouvel individu donné, à partir de la valeur de l'exogène *X, connaître la valeur* que prendrait l'endogène *Y* .

✓ Pour un nouvel individu *i∗, qui n'appartient pas à l'échantillon de données ayant participé à l'élaboration* du modèle, connaissant la valeur de *xi∗, on cherche à obtenir la prédiction ^yi∗.*

Prédiction ponctuelle

• Pour prédire Y a partir d'une valeur connue X:

•*On applique* directement l'équation de régression:

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

•La prédiction est sans biais: $E[\hat{y}_{i*}] = y_{i*}$.

En effet,

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

$\Rightarrow$

$$\begin{aligned}E(\hat{\varepsilon}_{i*}) &= E\left[(\hat{a} - a)x_{i*} + (\hat{b} - b) - \varepsilon_{i*}\right] \\ &= x_{i*}\underbrace{E(\hat{a} - a)}_{0} + \underbrace{E(\hat{b} - b)}_{0} - \underbrace{E(\varepsilon_{i*})}_{0}\end{aligned}$$

Les EMC sont sans biais

L'erreur du modèle est nulle par hypothèse

Prédiction par intervalle

• Prédiction ponctuelle est intéressante, mais avec quel degré de confiance?
→une intervalle de prédiction en lui associant une probabilité de recouvrir la vraie valeur *yi∗.*
➔Connaître
- ➢ la variance de l'erreur de prédiction
- ➢ la loi de distribution de l'erreur.

Prédiction par intervalle – Variance de l'erreur de prédiction

Variance de l'erreur de prévision

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

On montre

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

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. . points atypiques).

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | **Intervalle de Prédiction** | SPSS

Prédiction par intervalle – Variance de l'erreur de prédiction

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.

Prédiction par intervalle – la distribution de la variance de l'erreur de prédiction

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

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

$\Rightarrow$ $\dfrac{\hat{y}_{i*} - y_{i*}}{\hat{\sigma}_{\hat{\varepsilon}_{i*}}} \equiv \Im(n-2)$ 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-α)

Exemple: Rendement Agricole

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**
								sigma²(erreur)	**7.97984362**

ESTIMATION

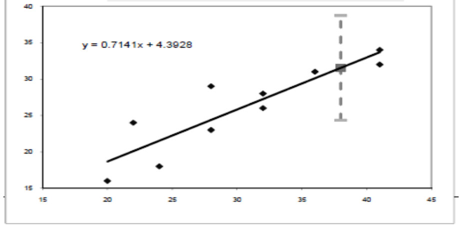
a	0.714053615
b	4.392770106
x*	38
y^	31.52680747
(x*-xb)^2	57.76
sigma²(epsilon^)	9.71389
t (0.975)	2.306004133
borne.basse	**24.33965896**
borne.haute	**38.71395598**

Variance de l'erreur de prédiction

$$\hat{\sigma}_{\hat{\varepsilon}_{i*}}^2 = \hat{\sigma}_\varepsilon^2 \left[1 + \frac{1}{n} + \frac{(x_{i*} - \bar{x})^2}{\sum_i (x_i - \bar{x})^2}\right]$$
$$= 7.9798 \times \left[1 + \frac{1}{10} + \frac{57.76}{492.4}\right]$$
$$= 9.71389$$

$\Rightarrow$ $b.b. = 31.5298 - 2.306 \times \sqrt{9.71389} = 24.3397$
$b.h. = 31.5298 + 2.306 \times \sqrt{9.71389} = 38.7140$

Intervalle de prédiction pour x* = 38



Récapitulatif ?

- Régression linéaire simple:
  - ✓ Nuage de point pour vérifier s' il y a relation linéaire.
  - ✓ Estimation des coefficients
  - ✓Tableau d' analyse de variance et coefficient de détermination
  - ✓ Test de significativité globale de la régression
  - ✓ Test de la significativité de la pente
  - ✓ Prédiction ponctuelle et par intervalle

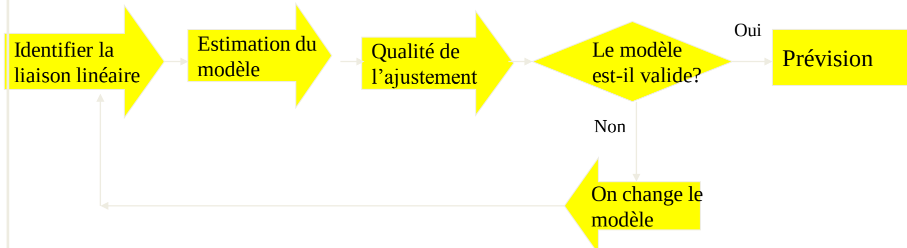
Mise en œuvre sous SPSS de la RLS

Exemple:

Une grande agence immobilière souhaite élaborer un modèle pour prédire le prix de vente d'une maison à partir de son prix d'achat.

Le fichier « House.sav » contient les données relatives aux 30 dernières maisons vendues.

Méthodologie pour l'élaboration de Modèle de Régression



Etape 1 : Identification de la liaison de type linéaire entre Y et les Xi

Etape 2 : Estimation de l'équation de régression (des coefficients)

Etape 3 : Qualité du modèle de régression linéaire

Etape 4 : Validation du modèle (Analyse des résidus)

Etape 5 : Prévision

Identification de la liaison linéaire

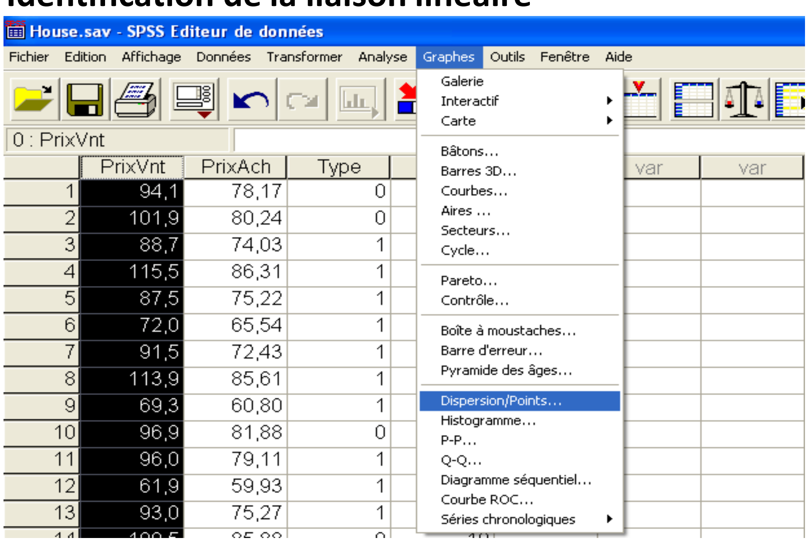


Diagramme de Dispersion

Identification de la liaison linéaire

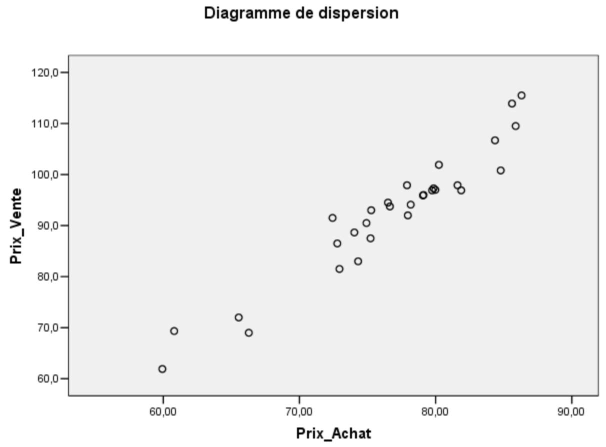


Diagramme de Dispersion

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Intervalle de Prédiction | SPSS

Identification de la liaison linéaire

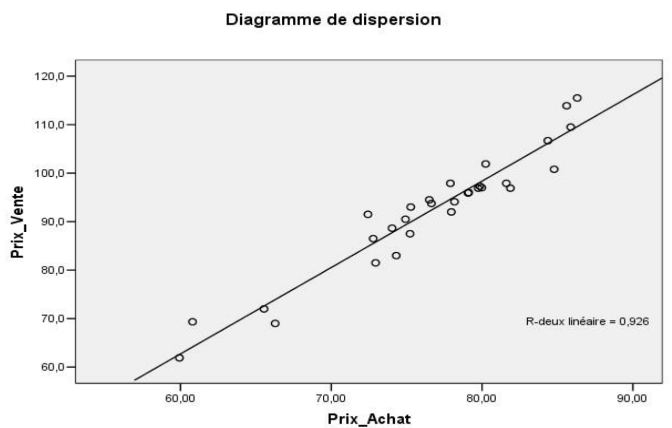
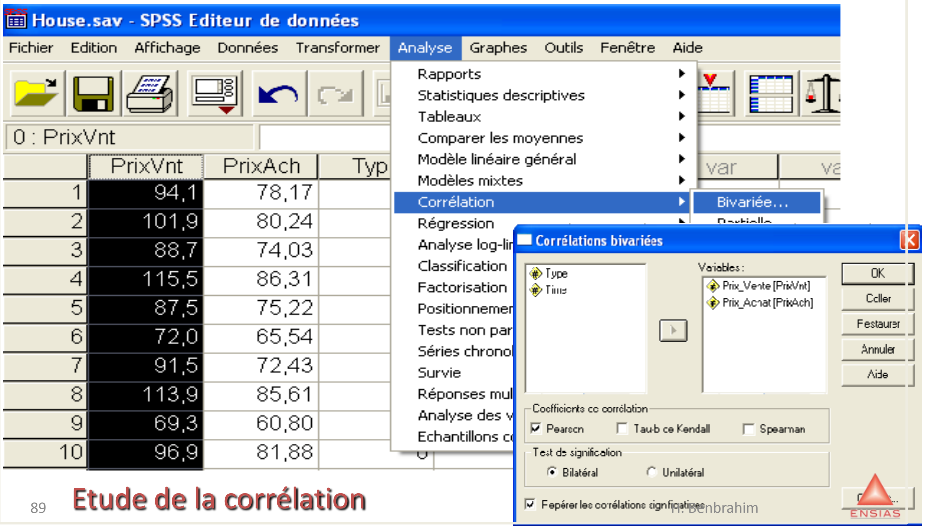


Diagramme de Dispersion

Identification de la liaison linéaire



Etude de la corrélation

Identification de la liaison linéaire

Corrélations

		Prix_Vente	Prix_Achat
Prix_Vente	Corrélation de Pearson	1	,962
	Sig. (bilatérale)		,000
	N	30	30
Prix_Achat	Corrélation de Pearson	,962	1
	Sig. (bilatérale)	,000	
	N	30	30

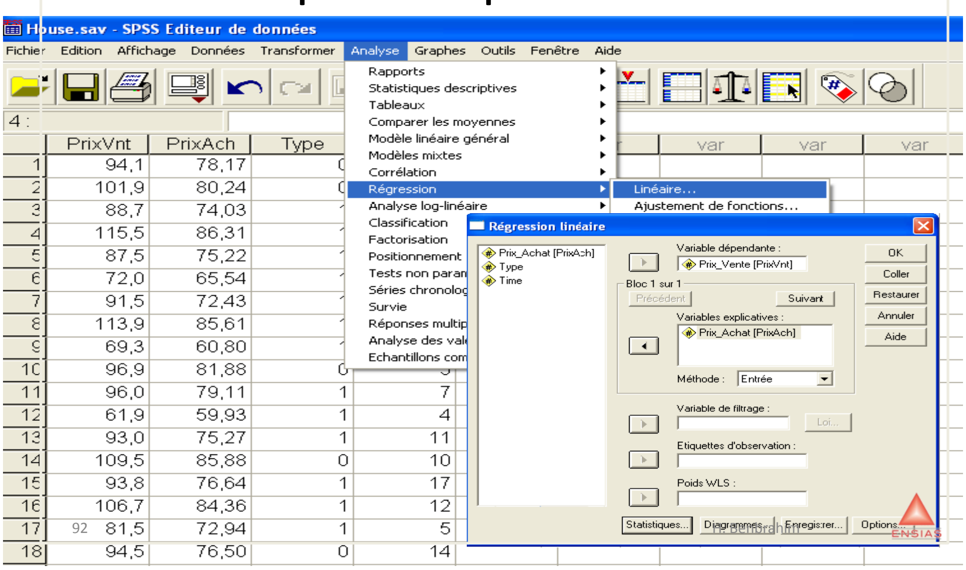
Test de corrélation

Identification du modèle

Aussi bien le diagramme de dispersion que le test de corrélation de Pearson suggèrent une relation linéaire entre le prix de vente (variable dépendante) ***y*** et le prix d'achat (variable indépendante) ***x***:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Estimation des paramètres par les moindres carrés



Estimation par les moindres carrés

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	?
		B	Erreur standard	Bêta	
1	(constante)	-43,615	7,668		
	Prix_A	1,775	,100	,959	

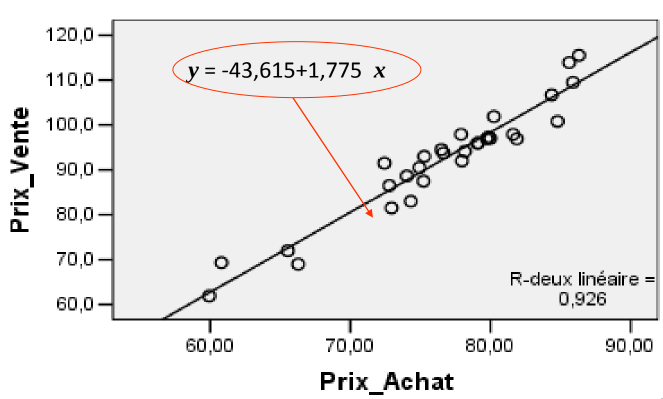
a. Variable dépendante : Prix_V

$Bêta = B_1 * (S_X / S_Y)$

Prix_vente = -43,615 + 1,775 * Prix_achat

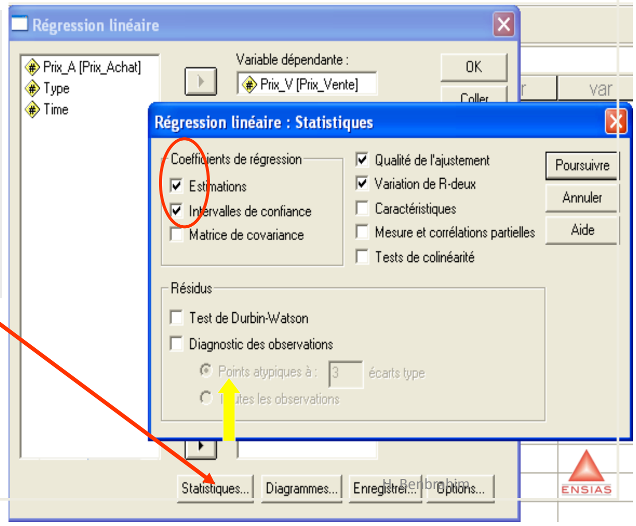
Interprétation

Une pente de 1,775 implique qu'une augmentation d'une unité en X entraînera une augmentation moyenne de 1,775 unités en Y.



Inférence

Statistiques permet d'obtenir l'estimation des coefficients de la régression ainsi que les intervalles de confiance des variables exogènes.



Inférence

Coefficients[a]

Modèle		Coefficients non standardisés: B	Coefficients non standardisés: Erreur standard	Coefficients standardisés: Bêta	t	Signification	Intervalle de confiance à 95% de B: Borne inférieure	Intervalle de confiance à 95% de B: Borne supérieure
1	(constante)	-43,615	7,668		-5,688	,000	-59,323	-27,908
	Prix_A	1,775	,100	,959	17,816	,000	1,571	1,979

a. Variable dépendante : Prix_V

Estimation ponctuelle

Ecart-type estimé de l'estimateur

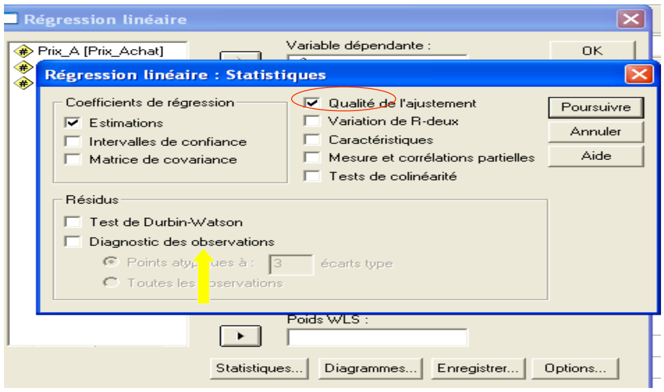
Statistique t pour tester la signification

P-value

Estimation par intervalle

- |T| > 1.96 ⇒ « signification » de la variable explicative, ou
- P-value < 0.05

Qualité d'ajustement



Récapitulatif du modèle[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,959[a]	,919	,916	3,6273

a. Valeurs prédites : (constantes), Prix_A

b. Variable dépendante : Prix_V

Qualité d'ajustement (suite)

Le *R^2-Ajusté*, est davantage utilisé que le R^2 car il ne dépend pas du nombre de variables:

$$R^2_{ajusté} = R^2 - \frac{p(1-R^2)}{N-p-1}$$

où p est le nombre de variables indépendantes et N le nombre d'observations.

Table de l' ANOVA

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0 \text{ pour au moins un i}$$

ANOVA[b]

Modèle			Somme des carrés	ddl	Carré moyen	F	Signification
1	*SSR*	Régression	4206,671	1	4206,671	348,374	,000[a]
	SSE	Résidu	338,105	28	12,075		
	SST	Total	4544,775	29			

a. Valeurs prédites : (constantes), Prix_Achat

b. Variable dépendante : Prix_Vente

$$F = \frac{SSR/k}{SSE/(n-k-1)} \approx F_{k,n-k-1}$$

Valeurs prédites et leurs écarts-types

- Prix de vente moyen de maisons dont la valeur à l'achat est de 67K.
- Prix de vente d'une maison dont la valeur à l'achat est de 67K.

Prix_vente prédit = - 43.615+1.775 * 67 = 75.33 K

Même chose. Ce qui diffère c'est l'erreur standard de la prévision.

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Validation du modèle?

- Relation linéaire entre Y et X ?
- Normalité de Y?
- L'analyse des résidus
  - ✓ Normalité de ε
  - ✓ Heteroscedasticité
  - ✓ Auto-corrélation

102 H. Benbrahim ENSIAS

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Linéarisation: Modèle Linéaire

Modèle linéaire

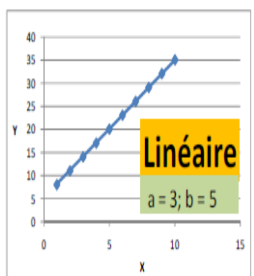
Lecture de la pente

$$Y = aX + b$$

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

103 H. Benbrahim ENSIAS

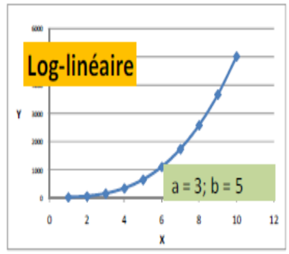
Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Linéarisation: Modèle Log-Linéaire

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)

104 H. Benbrahim ENSIAS

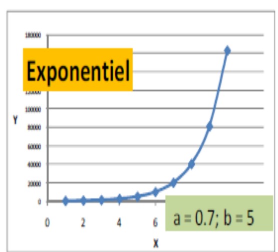
Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Linéarisation: Modèle Exponentiel

Modèle exponentiel (géométrique)

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



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

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

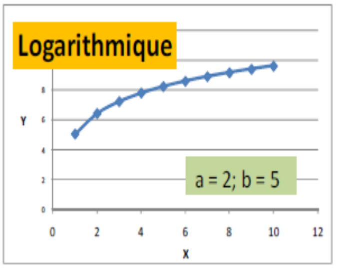
→ Linéarisation : ln(y) = a x + ln(b)

105 H. Benbrahim ENSIAS

Linéarisation: Modèle Logarithmique

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)

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | Validation

Linéarisation: Modèle Logistique

Un modèle particulier
Le modèle logistique

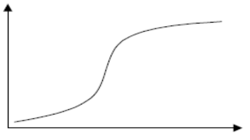
Problème :
Tous les modèles dans ont une concavité constante (dérivée seconde de signe constant), on peut avoir besoin d 'un modèle à plusieurs phases

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

Normalité de Y?

- Méthode graphique:
  - ✓ Histogramme
  - ✓ P-P plot

- Vérifier si Skewness AS ≈ 0
- Vérifier si Kurtosis AP ≈ 3
- Test de Kolmogorov – Smirnov

Analyse des résidus

H2 - Hypothèses sur le terme aléatoire *ε*.

H2.a $E(\varepsilon i) = 0$

H2.b $V(\varepsilon i) = \sigma 2$

➢ *C'est* l'hypothèse d' homoscédasticité.

H2.c $COV(xi, \varepsilon i) = 0$

➢ l'erreur est indépendante de la variable exogène

H2.d $COV(\varepsilon i, \varepsilon j) = 0$

➢ *"non auto-corrélation des erreurs".*

H2.e $\varepsilon i \equiv N(0, \sigma)$.

➢ *L'hypothèse de normalité des erreurs*

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

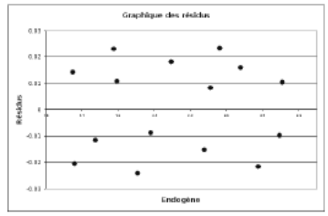
Graphique des résidus

Graphiques de base

Résidus vs. Endogène, vs. Exogènes

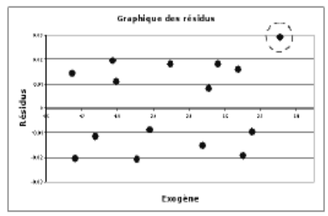
Résidus vs. Endogène

- Vérifier les points atypiques et/ou mal modélisés
- Vérifier si certaines plages de valeurs sont sous ou sur-estimées
- Vérifier la dispersion selon les valeurs de Y



Résidus vs. Exogènes

- Vérifier les points atypiques
- Vérifier les dépendances
- Vérifier la dispersion selon les plages de valeurs de X



Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

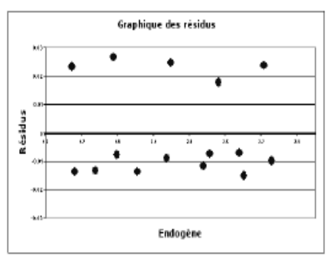
Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Analyse des résidus

Asymétrie, non linéarité et rupture de structure

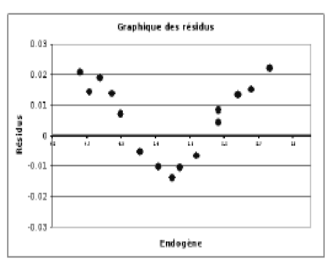
Asymétrie

- Des plages de données de l'endogène mal reconstitués
- Données atypiques
- Mélanges de populations différentes
- Problèmes de spécifications (absence d'exogènes importantes)



Non linéarité

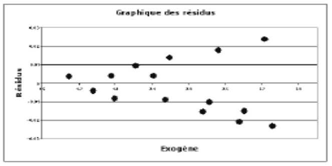
- Modèle linéaire inadapté, utiliser un modèle non linéaire
- Passer par des transformations de variables (log., carré, racine carrée, produit entre variables : interactions, etc.)



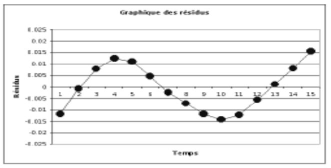
Analyse des résidus

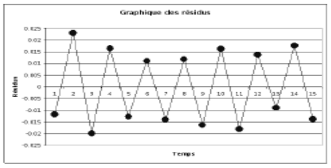
Hétéroscédasticité et autocorrélation des résidus

Hétéroscédasticité
- Variance des résidus non constante
- Exogène en abscisse pour détecter (traiter) dépendance



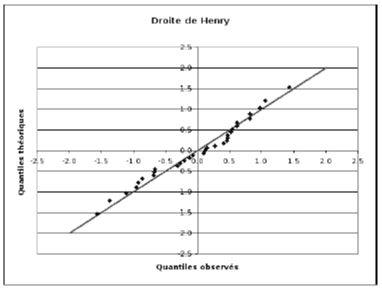
Autocorrélation
- Associée aux données longitudinales
- Processus particulier (régularité) au cours du temps ?
- Positive (blocs +/-) ou négative (alternance +/-)





Analyse des résidus

Test de normalité: Q-Q plot



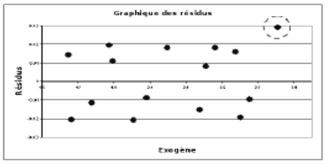
Points atypiques

Points atypiques et points influents

Points atypiques : Points qui s'écartent délibérément des autres
Points influents : Points qui pèsent (exagérément) sur les estimations : si on les enlevait, on obtiendrait des résultats (significativement) différents

Point atypique
Une valeur très différente sur l'endogène et/ou sur une ou combinaison d'exogènes. Elle n'est pas forcément mal modélisée (résidu élevé).

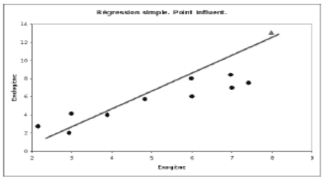
Cf. Endogène atypique O/N x Mal/Bien modélisé



Atypique exogène + Mal modélisé

Point influent
Très difficile à détecter visuellement
→ Peut être atypique ou non
→ Peut être bien modélisé ou non

Cf. Atypique non influent, Non atypique mais influent



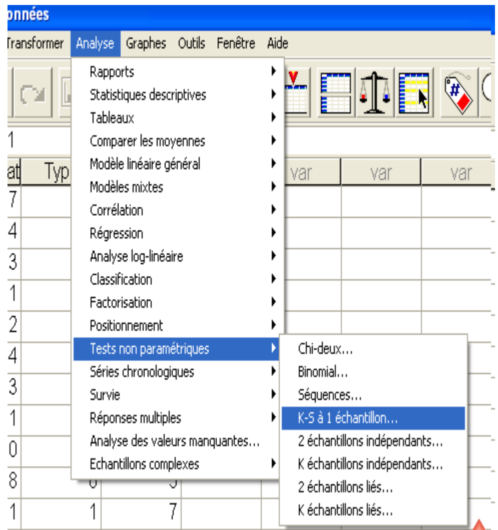
Régression simple : Point manifestement influent

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Validation sous SPSS

Normalité de Y

Le test de *Kolmogorov-Smirnov*



Normalité de Y

Test de Kolmogorov-Smirnov à un échantillon

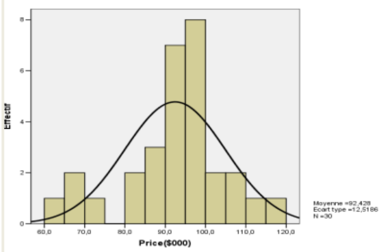
		Prix_V
N		30
Paramètres normaux[a,b]	Moyenne	92,495
	Ecart-type	12,5189
Différences les plus extrêmes	Absolue	,151
	Positive	,133
	Négative	-,151
Z de Kolmogorov-Smirnov		,827
Signification asymptotique (bilatérale)		,502

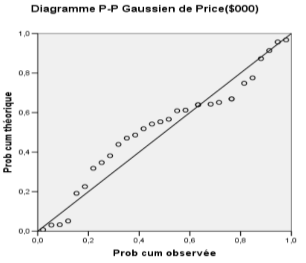
a. La distribution à tester est gaussienne.

b. Calculée à partir des données.

Normalité de Y

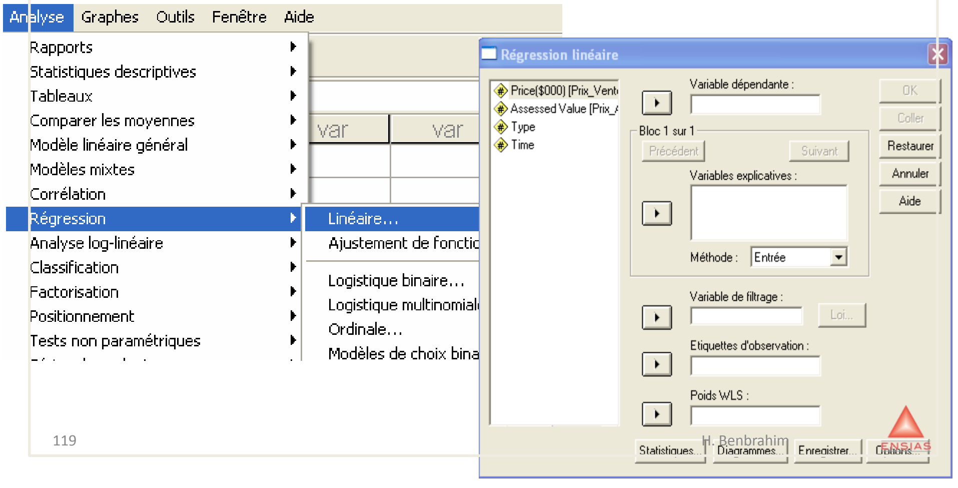
- Graphes>Histogrammes
- Graphes>P-P



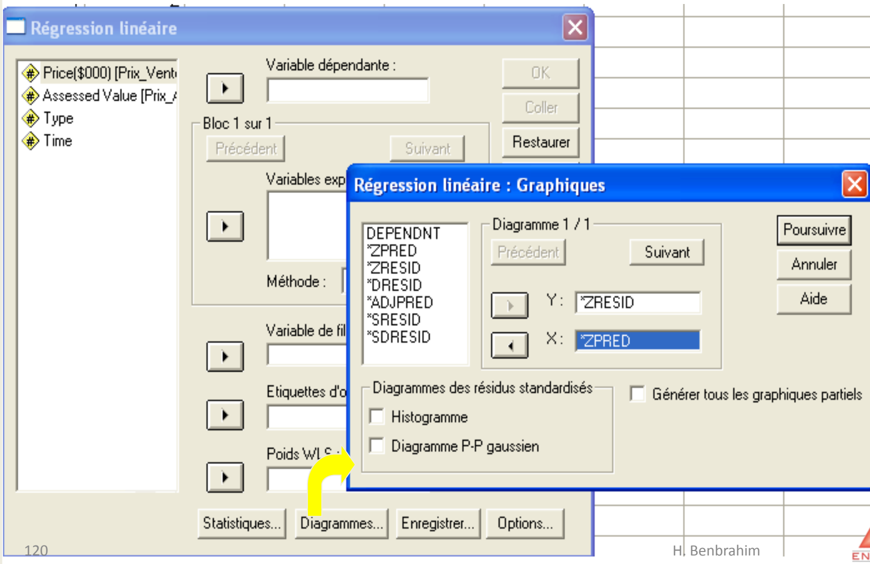


Résidus

En représentant les e_i en fonction des $\hat{Y}$, on peut visualiser les variances:

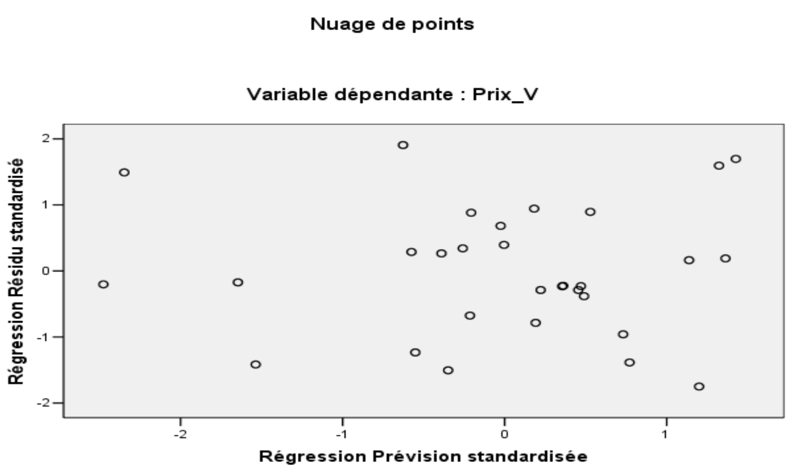


Résidus

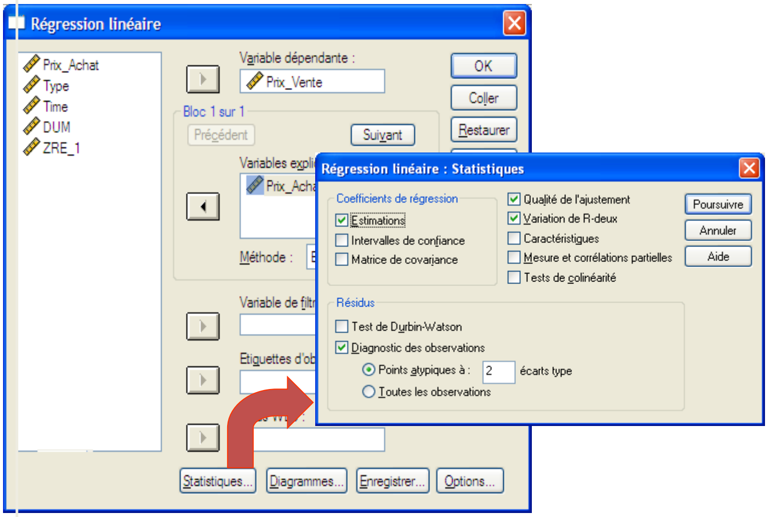


Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité | Modèle | Propriétés | Inférence | Prédiction | SPSS | **Validation**

Résidus



Points aberrants

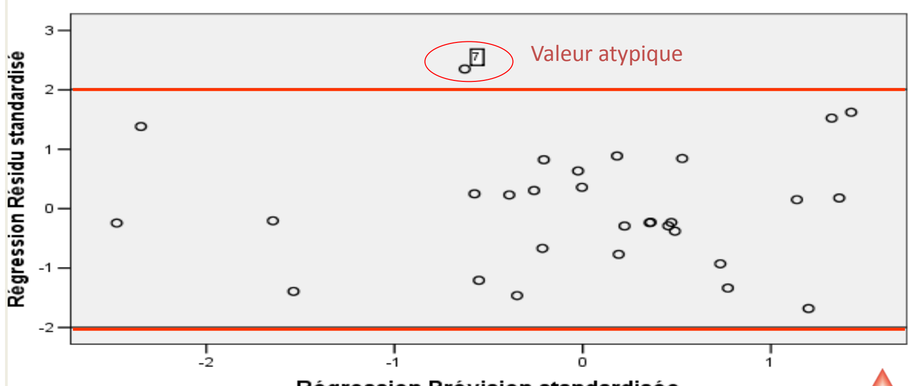


Points aberrants

Diagnostic des observations[a]

Numéro de l'observation	Résidu standardisé	Prix_V	Prévision	Résidu
7	2,351	93,5	84,971	8,5291

a. Variable dépendante : Prix_V



Points aberrants

- Le traitement est très simple. L'objectif est de neutraliser l'effet de cet individu atypique.
- On crée une variable muette qui prendra 1 pour l'observation 7 et 0 ailleurs.
- on introduit cette nouvelle variable en tant que variable explicative. (Dans notre exemple, la variable indicatrice s'appelle DUM).
- L'estimation des nouveaux coefficients est obtenue en introduisant la variable DUM comme deuxième variable exogène.

Points aberrants

Sans Dummy

Récapitulatif du modèle[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,959[a]	,919	,916	3,6273

a. Valeurs prédites : (constantes), Prix_A

b. Variable dépendante : Prix_V

Avec Dummy

Récapitulatif du modèle[b]

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,967[a]	,936	,931	3,2891

a. Valeurs prédites : (constantes), DUM, Prix_A

b. Variable dépendante : Prix_V

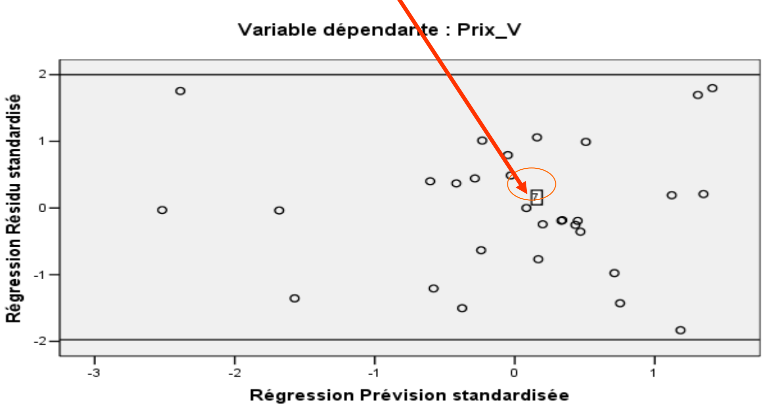
Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Signification
		B	Erreur standard	Bêta		
1	(constante)	-46,108	7,016		-6,572	,000
	Prix_A	1,804	,091	,974	19,825	,000
	DUM	8,949	3,369	,131	2,656	,013

a. Variable dépendante : Prix_V

Points aberrants

L'observation 7 n'est plus atypique

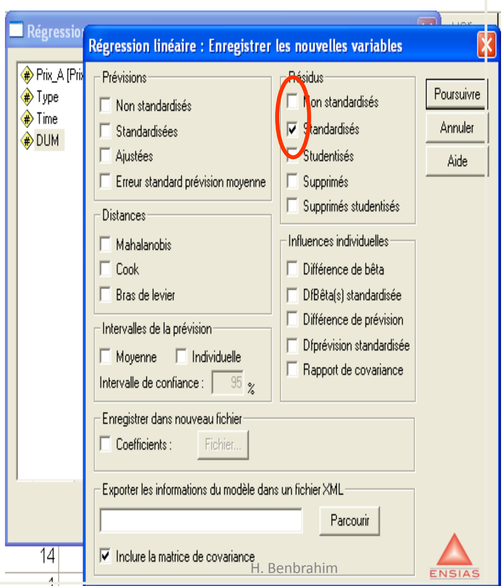


Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion
Généralité Modèle Propriétés Inférence Prédiction SPSS **Validation**

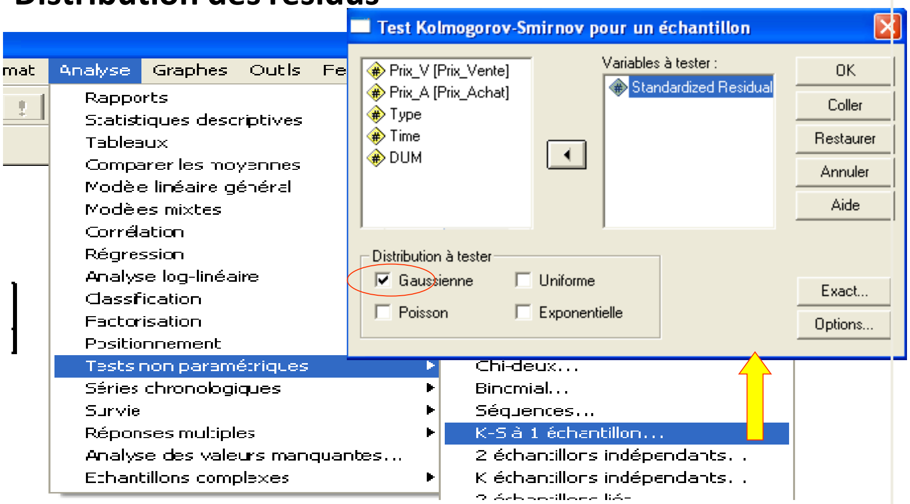
Normalité des résidus

Il est également très important de vérifier la normalité des résidus et de regarder s'ils sont bien aléatoires.

Le test de Kolmogorov-Smirnov permet de tester la normalité des résidus (standardisés) préalablement enregistrés:



Distribution des résidus



Distribution des résidus

Test de Kolmogorov-Smirnov à un échantillon

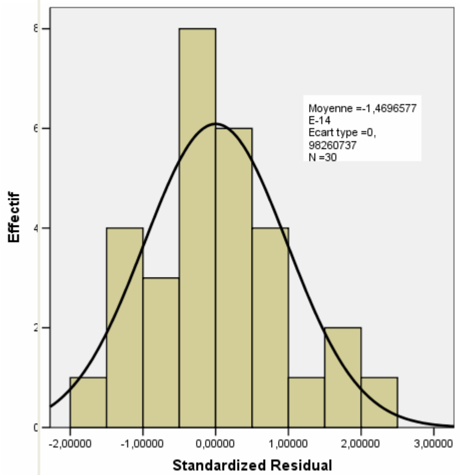
		Standardized Residual
N		30
Paramètres normaux[a,b]	Moyenne	,0000000
	Ecart-type	,96490128
Différences les plus extrêmes	Absolue	,096
	Positive	,073
	Négative	-,096
Z de Kolmogorov-Smirnov		,525
Signification asymptotique (bilatérale)		,946

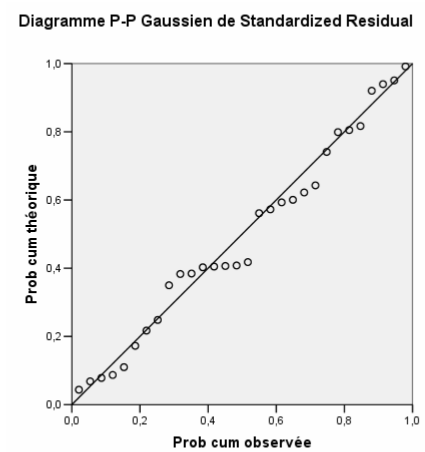
a. La distribution à tester est gaussienne.

b. Calculée à partir des données.

L'hypothèse de normalité est retenue.

Distribution des résidus



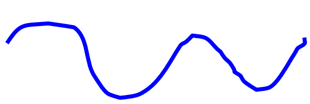


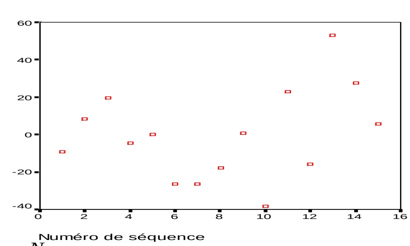
Là encore l'hypothèse de normalité peut être retenue.

L'autocorrélations des erreurs

L'autocorrélations des erreurs : problème souvent rencontré en économétrie: modèle mal spécifié, manque une variable explicative primordiale, etc.

Il n'y a pas de régularité évidente dans la suite des valeurs des résidus.





Durbin et Watson (d)

$$d = \frac{\sum_{t=2}^{N} (e_t - e_{t-1})^2}{\sum_{t=2}^{N} e_t^2}$$

Pour ne pas avoir autocorrélation des erreurs, il est nécessaire que d≈2 .

En résumé : Construction de Modèles de RL

Etape 1 : Identification de la liaison de type linéaire entre Y et les Xi
Etape 2 : Qualité du modèle de régression linéaire
Etape 3 : Estimation de l'équation de régression (des coefficients)
Etape 4 : Validation du modèle (Analyse des résidus)
Etape 5 : Prévision

ENSIAS

**Chapitre II:
Régression Linéaire Simple et Multiple**

Principe de la régression multiple

Population Ω $\begin{cases} Y & \text{variable à prédire (endogène), quantitative} \\ X & \text{variables exogènes (quelconques)} \end{cases}$

Objet de l 'étude

Une série de variables $X=(X_1|\ldots|X_p)$

On veut construire une fonction de prédiction (explication) telle que

$$Y = f(X, \alpha)$$

Objectif de l 'apprentissage

Utiliser un échantillon Ωa (extraite de la population) pour choisir la fonction f et ses paramètres α telle que l 'on minimise la somme des carrés des erreurs

$$S = \sum_{\Omega} [Y - \hat{f}(X, \hat{\alpha})]^2$$

Problèmes :
- il faut choisir une famille de fonction
- il faut estimer les paramètres α
- on utilise un échantillon pour optimiser sur la population

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

La régression linéaire multiple

- Se restreindre à une famille de fonction de prédiction linéaire
- Et à des exogènes continues (éventuellement des qualitatives recodées)

$$y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i \ ; i = 1, \ldots, n$$

Le terme aléatoire ε cristallise toutes les « insuffisances » du modèle :

- le modèle n 'est qu'une caricature de la réalité, la spécification (linéaire notamment) n 'est pas toujours rigoureusement exacte
- les variables qui ne sont pas prises en compte dans le modèle
- les fluctuations liées à l 'échantillonnage (si on change d 'échantillon, on peut obtenir un résultat différent)

ε quantifie les écarts entre les valeurs réellement observées et les valeurs prédites par le modèle

$(a_0, a_1, \ldots, a_p)$ Sont les paramètres du modèle que 'l'on veut estimer à l'aide des données

La régression linéaire multiple

Écriture matricielle

Pour une meilleure concision …

$$\begin{pmatrix} y_1 \\ \\ y_i \\ \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & & x_{1p} \\ 1 & & & \\ 1 & x_{i1} & x_{ij} & x_{ip} \\ 1 & & & \\ 1 & x_{n1} & & x_{np} \end{pmatrix} \begin{pmatrix} a_0 \\ a_1 \\ \\ \\ a_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \varepsilon_i \\ \\ \varepsilon_n \end{pmatrix}$$

N.B. Noter la colonne représentant la constante

$$Y = Xa + \varepsilon$$

$$(n,1) = (n, p+1) \times (p+1,1) + (n,1)$$

ENSIAS

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

La régression linéaire multiple

Démarche de modélisation

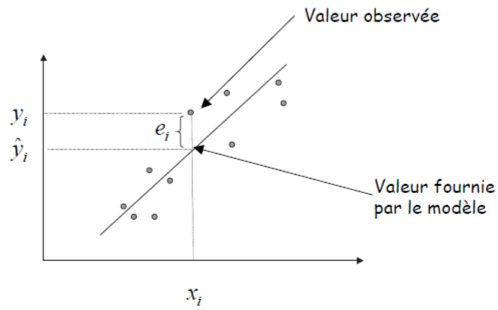
La démarche de modélisation est toujours la même

- estimer les paramètres « a » en exploitant les données
- évaluer la précision de ces estimateurs
- mesurer le pouvoir explicatif du modèle
- évaluer l'influence des variables dans le modèle
    - globalement (toutes les *p* variables)
    - individuellement (chaque variable)
    - un bloc de variables (q variables, q < p)
- sélectionner les variables les plus « pertinentes »
- évaluer la qualité du modèle lors de la prédiction (intervalle de prédiction)
- détecter les observations qui peuvent influencer exagérément les résultats (points atypiques).

ENSIAS

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

La méthode des moindres carrés



La méthode des moindres carrés cherche la meilleure estimation des paramètres « a » en minimisant la quantité

$$SCR = \sum_i e_i^2$$

$$avec\ e_i = Y - X\hat{a}$$

« e », l 'erreur observée est une évaluation du terme résiduel ε

Définition Régression Linéaire Simple Régression Linéaire Multiple Conclusion

Les hypothèses des moindres carrés

« â » deviennent les EMCO (estimateurs des moindres carrés ordinaires)

<u>Hypothèses probabilistes</u>

- *le modèle est linéaire en X*
- *les X sont observés sans erreur*
- *$E(\varepsilon) = 0$, en moyenne le modèle est bien spécifié*
- *$E(\varepsilon^2) = \sigma^2_\varepsilon$ la variance de l 'erreur est constante (hétéroscédasticité)*
- *$E(\varepsilon_i, \varepsilon_j)=0$, les erreurs sont non-corrélés*
- *$Cov(\varepsilon,x)=0$, l 'erreur est indépendante de la variable explicative*
- *$\varepsilon \equiv$ Normale(0, σ^2_ε)*

<u>Hypothèses structurelles</u>

- *Rang(X 'X)=p+1 càd $(X'X)^{-1}$ existe*
- *(X 'X)/n tend vers une matrice finie non singulière*
- *n>p+1, le nombre d 'observations est supérieur au nombre de variables explicatives*

Idée : rendre les calculs possibles et délimiter les propriétés des estimateurs

Les estimateurs des MCO

Pour trouver les paramètres « a » qui minimise S :

$$S = \sum_i \varepsilon_i^{\,2} = \sum_i [y_i - (a_0 + a_{i,1}x_1 + \ldots + a_{i,p}x_p)]^2$$

On doit résoudre $\frac{\partial S}{\partial a} = 0$ Il y a (p+1) équations dites « équations normales » à résoudre

L'estimateur des moindres carrés ordinaires s'écrit : $\hat{a} = (X'X)^{-1}X'Y$

N.B. Compte tenu des hypothèses ci-dessus :
- â est sans biais
- â est convergent
- â est BLUE (c.-à-d. il n'existe pas d'estimateur linéaire sans biais de variance plus petite)

ENSIAS

Définition | Régression Linéaire Simple | Régression Linéaire Multiple | Conclusion

Evaluation 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 traduite par le modèle

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

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

Test associé à l'évaluation globale du modèle

Test de Fisher

Quelques formulations du test de « signification » globale :
- le modèle est-il pertinent pour expliquer les valeurs de Y ?
- la liaison linéaire Y / X1,…,Xp est-elle licite ?
- test d'hypothèse

$$\begin{cases} H_0 : a_1 = a_2 = \cdots = a_p = 0 \\ H_1 : \text{il en existe au moins } \neq 0 \end{cases}$$

Attention, on n'inclut pas la constante dans le test
c.-à-d. la question est donc, y a-t-il au moins une variable qui emmène de l'information ? Ou il n'est pas possible d'effectuer une prédiction/explication meilleure que la simple constante ?

Statistique du test

$$F = \frac{R^2 / p}{(1-R^2) / (n-p-1)} \equiv Fisher(p, n-p-1)$$

A un niveau de signification donné (ex. 10%, 5%, 1%…)
>> Comparer le F-calculé avec le F-théorique fourni par la table
>> Comparer la p-value avec le niveau de signification

Évaluation individuelle des coefficients

Une variable contribue-t-elle de manière significative dans la régression ?

Test d'hypothèse associé

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

H0 vérifiée signifie que la variable peut être supprimée du modèle sans en détériorer le pouvoir explicatif

Statistique du test : il s'appuie sur â et l'estimation de son écart-type

$$t = \frac{\hat{a}_j}{\hat{\sigma}_{\hat{a}_j}} \equiv Student(n-p-1)$$

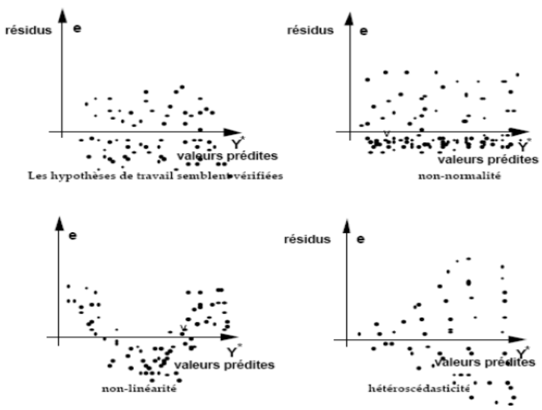
A un niveau de signification donné (ex. 10%, 5%, 1%…)
>> Comparer le t-calculé avec le t-théorique fourni par la table de Student
>> Comparer la p-value avec le niveau de signification

Diagnostiquer une régression avec les graphiques des résidus

Les graphiques des résidus est un excellent outil de diagnostic de la régression, elles permettent notamment de vérifier l'adéquation aux hypothèses initiales et détecter l'existence éventuelle de points atypiques
>> résidus vs. Y
>> résidus vs. chaque variable X

→

On veut vérifier :
>> problème de spécification (non-linéarité du modèle)
>> normalité des résidus (distribués aléatoirement et symétriquement autour de 0)
>> variance constante (homoscédasticité)
>> absence de « structure » dans l'évolution de l'erreur (non auto-corrélation – indépendance)



ENSIAS

Chapitre II: Régression Linéaire Simple et Multiple

RECAPUTILATIF?

CONCLUSION?