

Liaison et dépendance entre deux variables quantitatives

Régression linéaire simple

Position du problème

X et Y sont deux grandeurs statistiques observées

ex: en macro-économie

PIB, Revenu des ménages, Importations...

en micro-économie

Revenu d'un ménage, bénéfices d'une entreprise...

en médecine ou autres sujets

taille d'un individu, poids...

Hypothèse de base : X et Y sont des grandeurs continues, théoriquement définies dans \mathbb{R} . Dans la réalité on se contente du fait que la différence entre deux valeurs de X (resp. Y) ait un sens (par ex. l'âge d'une personne n'est pas définie dans \mathbb{R} , les valeurs sont discrètes, en revanche la différence d'âge entre deux personnes a un sens)

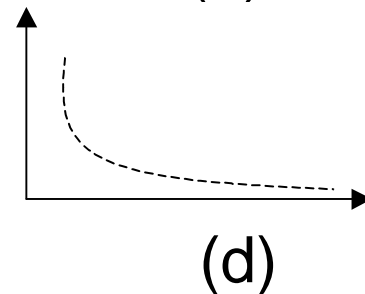
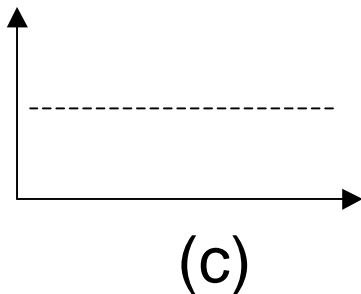
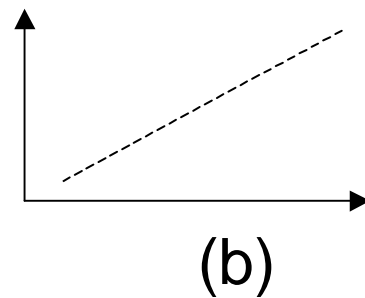
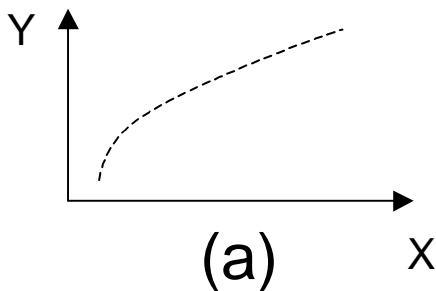
Dans l'étude de la dépendance entre X et Y, on se pose trois questions fondamentales :

- 👉 **X et Y sont-ils liés, comment mesurer cette liaison ?**
- 👉 **Trouver une fonction qui permet de déterminer Y à partir de X**
- 👉 **Estimer les paramètres de cette fonction ?**

Position du problème

Evaluer la liaison entre X et Y, i.e répondre à la question X et Y ont-ils une évolution commune ?

1) Etude graphique



Plusieurs points de vue :

- ☞ en terme d'évolution - quand X augmente, Y augmente (diminue) ?
- ☞ en terme de niveaux - quand X est faible (fort), Y est faible (fort) ?

Comment quantifier ces évaluations graphiques ?

2) Etude numérique : le coefficient de corrélation

Objectif : Quantifier la liaison entre X et Y de manière à mettre en évidence

- ☞ le « **sens** » de la liaison;
- ☞ la « **force** » de la liaison.

Le coefficient de corrélation

$$r = \frac{\text{cov}(X, Y)}{\sigma_X \cdot \sigma_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}}$$

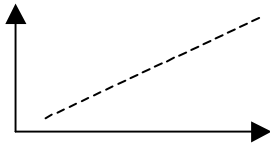
Tableau de données

i	X	Y
...	2.5	3.5
	4.5	5.5
	3.5	4.5
	6.5	7.8
	4.6	8.5

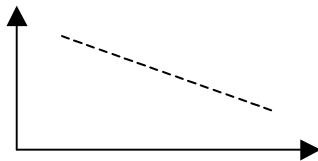
i est le numéro d'observation

- *si i est une date, on parle de données « temporelles » ou encore « longitudinales »*
- *si i représente un individu statistique (un ménage, une voiture...), on parle de données transversales ou encore de coupe instantanée*

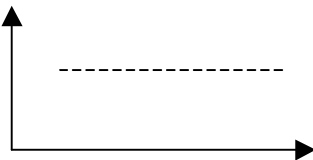
Interprétation : sens de la corrélation



$r > 0$, corrélation positive



$r < 0$, corrélation négative



$r = 0$, absence de corrélation

Interprétation : force de la corrélation

$|r| \neq 1$, corrélation forte

$|r| \neq 0$, corrélation faible

3) Evaluation statistique : test d'hypothèses

Problème :

- on travaille souvent sur un échantillon (de taille limitée) issu de la population
- on veut inférer les résultats obtenus sur la population originelle

ex: pour connaître les résultats des élections (plusieurs millions de votants en France), on pose la question à un échantillon de 1000 personnes choisis au hasard et on en extrait une conclusion sur l'ensemble des votants

Principe du test statistique :

Concernant la population totale, une hypothèse est formulée, la question qui se pose est : **dans quelle mesure cette hypothèse est confirmée / infirmée par les données observées**

ex: « 50% des électeurs voteront pour Duchemol », ceci est-il confirmé par les données observées ?

Attention :

↑↑ *on ne peut pas décider avec certitude puisque l'on ne connaît pas la population totale*

☞ *mais le degré de confiance que l'on accorde à la conclusion peut être exprimé en terme de probabilité*

Hypothèses à tester :

On oppose généralement une hypothèse dite nulle (H_0) avec une hypothèse dite alternative (H_1), les risques associés à la prise de décision sont les suivants :

Décision fondée sur les données

	Décider que H_0 est vrai	Décider que H_0 est faux
<i>Etat de la nature (réalité)</i>	Décision correcte	<i>Risque de première espèce (α)</i>
H_0 est faux (H_1 est vrai)	<i>Risque de 2^{ème} espèce (β)</i>	Décision correcte

Application au coefficient de corrélation :

On ne dispose pas directement de r mais de \hat{r} qui est estimé sur un échantillon. On veut tester

$$H_0 : r = 0$$

$$H_1 : r \neq 0$$

Une règle de décision simple serait :

$$\text{Accepter } H_0 \text{ ssi } |\hat{r}| < r_\alpha$$

$$\text{Rejeter } H_0 \text{ ssi } |\hat{r}| \geq r_\alpha$$

Le seuil critique du test (celle qui permet de définir la région de rejet et la région d'acceptation de l'hypothèse H_0) est défini par :

$$\alpha = P(\text{Rejeter } H_0 / H_0 \text{ est vrai})$$

$$\alpha = P(|\hat{r}| > r_\alpha / r = 0)$$

Si l'on connaît la loi de distribution statistique P , on peut calculer le seuil à partir du risque de première espèce

Dans la pratique : on ne connaît pas la loi de \hat{r} , on connaît en revanche la loi de distribution de

$$t = \frac{\hat{r}}{\sqrt{\frac{1 - \hat{r}^2}{n - 2}}} \equiv \text{Student} (n - 2)$$

Loi de distribution de t, une loi de Student à (n-2) degrés de liberté; n est le nombre d'observations du tableau statistique.

Règle de décision

Accepter $H_0 (r = 0)$ ssi $|t| < t_\alpha$

Rejeter $H_0 (r \neq 0)$ ssi $|t| \geq t_\alpha$

Attention :

Les logiciels donnent rarement la valeur de t_α afin que l'on puisse le comparer avec t

Ils fournissent en général directement la valeur α' telle que

$$\alpha' = P(|\text{Student}(n - 2)| \geq |t|)$$

La règle de décision devient ainsi

Accepter $H_0 (r = 0)$ ssi $\alpha' > \alpha$

Rejeter $H_0 (r \neq 0)$ ssi $\alpha' \leq \alpha$

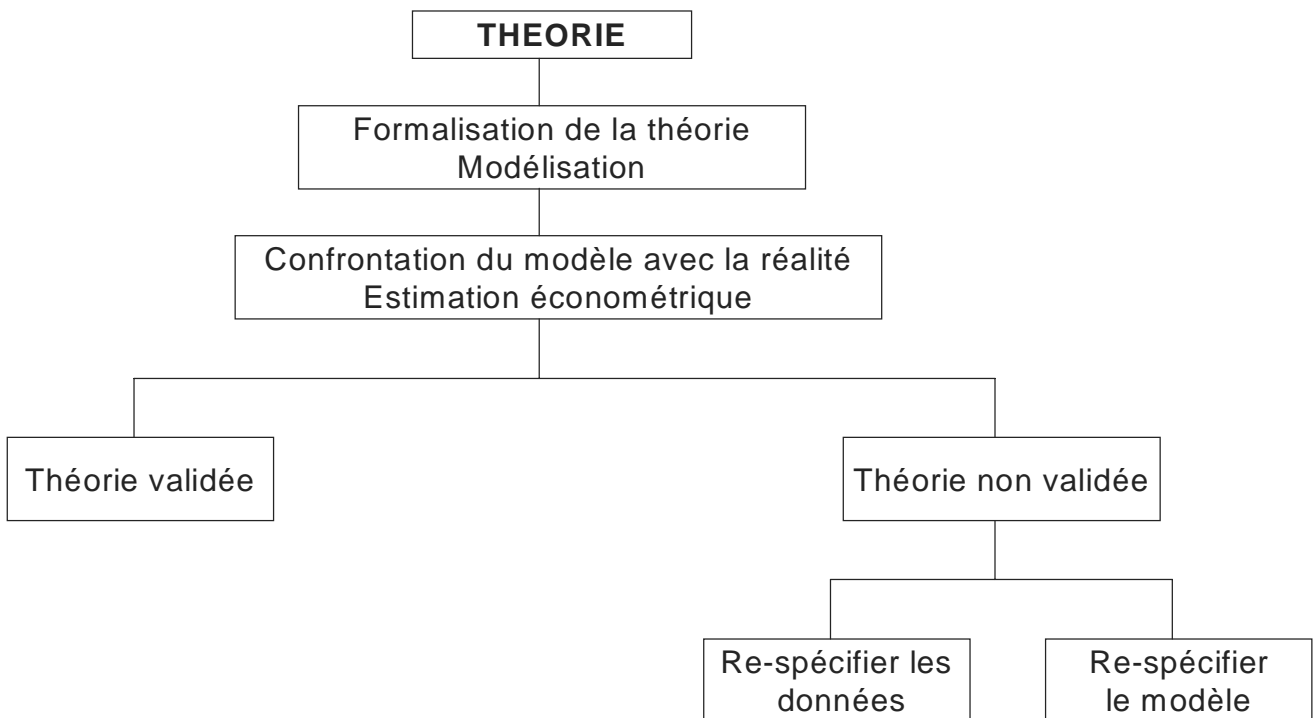
α' est connu sous le terme « p-value » ou « significance »

Choix de la spécification permettant de déterminer Y à partir de X

Position du problème

On cherche une fonction f telle que $Y=f(X,\alpha)$
Comment spécifier cette fonction f ?

1) Généralités sur la démarche économétrique



Attention :

Seule la théorie (économique,...) doit nous guider pour la spécification du modèle, les données ne doivent servir qu'à valider ou invalider les hypothèses que l'on émet...

Il est donc nécessaire de bien comprendre les hypothèses sous-jacentes à chacune des fonctions proposées

2) Quelques modèles de base

Modèle	Formule	Propriété fondamentale
<i>Linéaire</i>	$Y = aX + b$	la variation de Y est proportionnelle à la variation de X
<i>Log-linéaire</i>	$Y = B X^a$	le taux de variation de Y est proportionnel au taux de variation de X
<i>Exponentiel</i>	$Y = e^{aX + b}$	le taux de variation de Y est proportionnel à la variation de X
<i>Logarithmique</i>	$Y = a \ln(X) + b$	la variation de Y est proportionnelle au taux de variation de X

a) Propriétés du modèle linéaire

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

- simplicité
- peut être appliqué directement dans un premier temps pour vérifier l'existence d'une relation
- estimation directe des paramètres par la méthode des moindres carrés

b) Propriétés du modèle log-linéaire

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

- favori des économistes - modèle à élasticité constante
- ex: emploi=f(production), demande=f(prix)...
- linéarisation par $\ln(y) = a \ln(x) + \ln(b)$

c) Propriétés du modèle exponentiel (géométrique)

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

- surtout utilisé quand x =temps (ainsi $dx=1$)
- dans ce cas, la croissance (décroissance) de y est constant dans le temps
- ex : évolution du nombre de pages web dans le monde
- ce type d'évolution ne dure pas longtemps
- linéarisation : $\ln(y) = a x + b$

d) Propriétés du modèle logarithmique

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

- archétype de la croissance (décroissance) qui s'épuise
- ex : salaire = f (ancienneté) ou vente= f (publicité)
- linéarisation : $y = a \ln(x) + b$

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

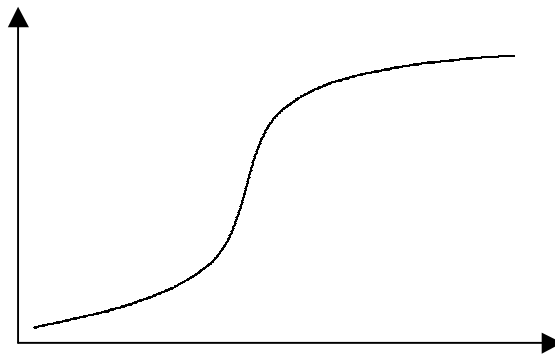
- produit inconnu
- positionnement sur le marché

Croissance accélérée

- large diffusion

Freinage

- saturation du marché
- concurrence



Equation

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

Linéarisation

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

Estimation des paramètres de la fonction spécifiée

La régression simple et multiple

Position du problème

On veut estimer les paramètres « a » et « b » de la fonction

$$y = a x + b$$

en utilisant les données issues de l'échantillonnage

Position du problème (2)

On se place dans un cadre plus général de l'estimation de l'équation de régression

$$y = a_0 + a_1 x_1 + a_2 x_2 + \dots + a_p x_p$$

- y est la variable à prédire, dite endogène
- $x_1 \dots x_p$ sont les variables prédictives, dites exogènes

Les estimations issues de l'échantillon seront notés

$$\hat{a}_0, \hat{a}_1, \dots, \hat{a}_p$$

Notation matricielle et spécification statistique

On dispose de n observations, l'écriture du modèle pour chaque observations ($i=1, \dots, n$) passe par une forme matricielle

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

$$Y = Xa$$

L'estimation statistique passe par le rajout d'un terme aléatoire ε qui tient un rôle essentiel

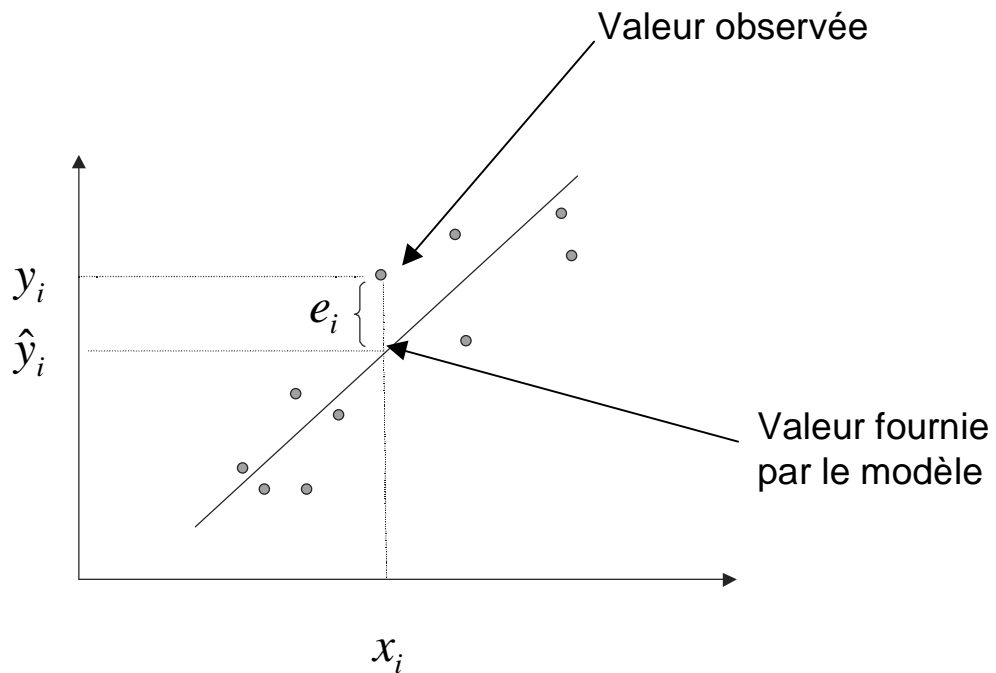
$$Y = Xa + \varepsilon$$

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

- *le modèle n'est qu'une caricature de la réalité, la spécification n'est pas toujours rigoureusement exacte*
- *les erreurs de mesure sur les données*
- *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

Méthode des moindres carrés



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

$$SCR = \sum_i e_i^2$$

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

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

1) Hypothèses

L'estimation des moindres carrés et son évaluation ne prend effet que si certaines hypothèses sont respectées

Hypothèses probabilistes

- le modèle est linéaire en X
- les X sont observés sans erreur
- $E(\varepsilon) = 0$, en moyenne le modèle est bien spécifié
- $E(\varepsilon^2) = \sigma_\varepsilon^2$ 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
- $\varepsilon \equiv Normale(0, \sigma_\varepsilon^2)$

Hypothèses structurelles

- $Rang(X'X) = p+1$ 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

2) Estimation des moindres carrés

L'estimation des moindres carrés du vecteur « a » s'écrit

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

3) Interprétation des coefficients

$$y = a_0 + a_1x_1 + \dots + a_jx_j + \dots + a_px_p$$

Toutes choses égales par ailleurs i.e toutes les autres variables sont constantes, alors

$$\frac{\partial y}{\partial x_j} = a_j$$

4) Evaluation globale du modèle, tableau d 'analyse de variance et coefficient de détermination

La qualité de l 'estimation est traduite par l 'équation d 'analyse de variance

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

Endogène observée
Endogène estimée

SCT

Variabilité totale

SCE

Variabilité expliquée par le modèle

SCR

Variabilité non- expliquée (Variabilité résiduelle)

Les logiciels présente très souvent le tableau d 'analyse de variance

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

Un indicateur synthétique issu du tableau d 'analyse de variance permet d 'évaluer globalement le modèle construit : le coefficient de détermination

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

R²>0, le modèle est intéressant
R²=0, le modèle est mauvais

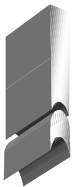
La qualité du modèle étant évaluée sur un échantillon, le R^2 calculé est soumis à une certaine variabilité.

Si on veut s'assurer que le modèle est intéressant, on procédera au test d'hypothèses

$$H_0 : R^2 = 0$$

$$H_1 : R^2 > 0$$

Sachant que l'on dispose du coefficient \hat{R}^2 estimé sur l'échantillon.



Pour ce faire, on a besoin du risque critique et de la loi de distribution du coefficient calculé. Comme pour le coefficient de corrélation, il est nécessaire de passer par un indicateur intermédiaire

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

Loi de distribution de Fischer, à p et $n-p-1$ degrés de liberté

On obtient la p-value par $\alpha' = P(\text{Fischer}(p, n - p - 1) > F)$

Pour un risque α , la règle de décision devient

- *Accepter H_0 ssi $\alpha' > \alpha$*
- *Rejeter H_0 ssi $\alpha' < \alpha$*

5) Evaluation individuelle des coefficients de régression

On cherche à savoir si la variable X_j a une influence significative sur Y ?

⇒ Test d'hypothèses

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

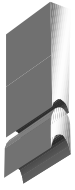
en utilisant les estimations \hat{a}_j

A cet effet, on a besoin de connaître la variance de ces estimations....

$$\hat{\Omega}_{\hat{a}} = \hat{\sigma}_{\varepsilon}^2 (X'X)^{-1} = \begin{pmatrix} \hat{\sigma}_{a_0} & & & & \\ & \hat{\sigma}_{a_1} & & & \\ & & \ddots & & \\ & & & \ddots & \\ & & & & \hat{\sigma}_{a_p} \end{pmatrix}$$

Variance estimée de l'erreur

$$\hat{\sigma}_{\varepsilon}^2 = \frac{\sum_i e_i^2}{n - p - 1}$$



L'indicateur que l'on utilisera pour le test est

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

La p-value est obtenue avec $\alpha' = P(|Student(n - p - 1)| > |t|)$

Pour un risque α , la règle de décision est

- *Accepter H_0 ssi $\alpha' > \alpha$*
- *Rejeter H_0 ssi $\alpha' < \alpha$*

6) Estimation par intervalle des coefficients de régression

\hat{a}_j sont entachés de variabilité: si on change d'échantillon, l'estimation pourrait être légèrement différente



Peut-on produire un intervalle qui, avec une certaine probabilité α , va contenir la « vraie » valeur de a_j ?

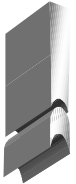
On sait que $\frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \equiv Student(n - p - 1)$



Pour une probabilité α , on peut définir les bornes de variation de la loi de Student $t_{\alpha/2}$

$$-t_{\alpha/2} \leq \frac{\hat{a}_j - a_j}{\hat{\sigma}_{\hat{a}_j}} \leq +t_{\alpha/2}$$

L'intervalle de variation du coefficient estimé s'écrit alors



$$\hat{a}_j - t_{\alpha/2} \times \hat{\sigma}_{\hat{a}_j} \leq a_j \leq \hat{a}_j + t_{\alpha/2} \times \hat{\sigma}_{\hat{a}_j}$$

Un exemple simple
Effet de l'engrais sur le rendement en maïs

Rendement (quintal)	Engrais (kilo)
16	20
18	24
23	28
24	22
28	32
29	28
26	32
31	36
32	41
34	41

$$\hat{R}^2 = 0.99$$

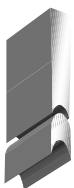
$$F = 862.509 \quad (p\text{-value} = 0.000)$$

$$Rendement = 0.851 \times Engrais$$

$$\hat{\sigma}_{\hat{a}} = 0.029$$

$$t = 29.36 \quad (p\text{-value} = 0.000)$$

Pour un risque d'erreur $\alpha = 0.05$, les bornes de variation de la loi de Student $t_{\alpha/2} = 2.262 \Rightarrow 0.785 \leq a \leq 0.916$



Avec un risque d'erreur de 5% : on peut dire que dans le pire des cas, une injection supplémentaire de 1 kilo d'engrais, fera augmenter la production de 0.785 quintaux

7) Comparaison et Sélection de modèles

Position du problème

On ne dispose pas de théorie précise pour guider la spécification, de fait on est face à des modèles concurrents avec 2 difficultés :

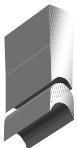
☞ ils n'ont pas le même nombre de variables

☞ ils n'utilisent pas les mêmes variables

A) Comparaison de deux modèles



Intuitivement, le R^2 indiquant la qualité globale du modèle, on préférera le modèle qui a le plus fort R^2



NON ! Le R^2 est un indicateur inapproprié ici car il augmente de manière mécanique avec le nombre de variable

$$R^2_{Y \times x_1, x_2} < R^2_{Y \times x_1, x_2, x_3}$$

↑
Même si x_3 est une variable qui n'apporte aucune information



R^2 corrigé des degrés de liberté

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

↑
Les modèles qui introduisent beaucoup de variables seront pénalisés

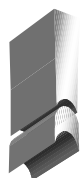
B) Méthodes de sélection automatique des modèles

Principe : chercher

- ☞ *les variables les plus corrélées avec l'endogène*
- ☞ *les moins corrélées entre elles*

C'est notre objectif de départ

Eviter la redondance de l'information, du fait de la colinéarité des variables peuvent être éjectés à tort par d'autres, et les coefficients estimés sont généralement très instables



Nous sommes dans le cadre de la statistique exploratoire ici, l'interprétation des résultats peut être très difficile, voire impossible. En revanche, pour des fins de prévision, ces techniques peuvent être très utiles.

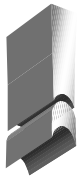
Quelques techniques :

- ☞ tester toutes les régressions possibles et choisir celui qui a le meilleur R^2 corrigé avec tous les coefficients significatifs
- ☞ élimination progressive (backward elimination) : démarrer avec toutes les variables et éliminer un à un les variables dont les coefficients ne sont pas significatifs
- ☞ sélection progressive (forward regression) : prendre la variable la plus corrélée avec Y, puis prendre la 2ème variable la plus corrélée avec Y en éliminant l'influence des variables déjà introduites (cf. Notion de corrélation partielle). On s'arrête quand le t de Student d'une variable introduite est non significatif.
- ☞ stepwise regression : combinaison de forward et backward

Sachant que : $\frac{e_h}{\hat{\sigma}_{e_h}} \equiv Student(n - p - 1)$

L 'intervalle de prévision au niveau de confiance $(1-\alpha)$
(au risque α) s 'écrit :

$$y_h = \hat{y}_h \pm t_{\alpha/2} \times \hat{\sigma}_\varepsilon \left[X'_h (X'X)^{-1} X_h + 1 \right]^{1/2}$$



L 'intervalle de prévision
sera d 'autant plus large

On prend un risque
très faible

La variance résiduelle
est forte (standard error
of estimate)

2) Autocorrélation des résidus

(surtout pour les modèles temporelles)

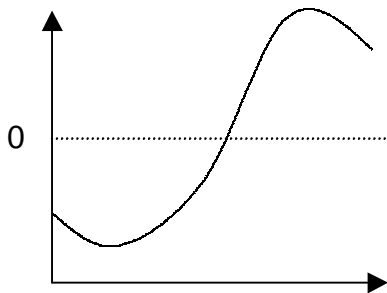
Une des hypothèses probabilistes est violée

$$E(\varepsilon_t, \varepsilon_{t'}) \neq 0$$

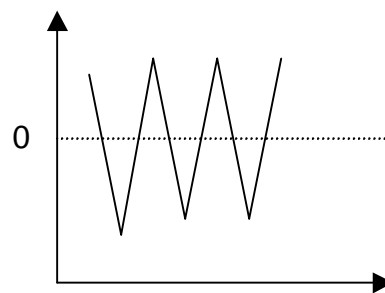
Causes probables :

- ☞ *une variable explicative manque*
- ☞ *le modèle est mal spécifié*
- ☞ *les données ont été « travaillées » au préalable*

Détection graphique



Autocorrélation positive



Autocorrélation négative

Test statistique : test de Durbin-Watson

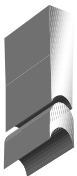
Permet de tester l'autocorrélation d'ordre 1, de la forme

$$\varepsilon_t = \rho \varepsilon_{t-1} + v_t, \text{ avec } v_t \equiv \text{Normale}(0, \sigma_v^2)$$

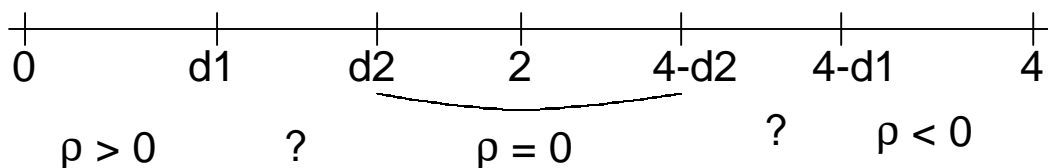
Le test porte sur $\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho \neq 0 \end{cases}$

L'indicateur utilisé est

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{i=1}^n e_i^2}$$



On ne dispose pas de p-value ici, il faut comparer le DW calculé avec les seuils d1 et d2 fournis par la table de Durbin et Watson



Estimation

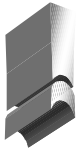
L'estimation s'effectue sur l'équation en différences :

$$y_t - \hat{\rho}y_{t-1} = \hat{a}_0(1 - \hat{\rho}) + \hat{a}_1(x_{t,1} - \hat{\rho}x_{t-1,1}) + \dots + \hat{a}_p(x_{t,p} - \hat{\rho}x_{t-1,p})$$

estimation conjointe des \hat{a}_j et de ρ

- *Cochrane-Orcutt (méthode d'itérations successives)*
- *Hildreth-Lu (méthode de balayage)*

Prévision en présence d'autocorrélation



Il faut utiliser l'équation en différences

Ex : prévoir à la période $t+1$

$$\hat{y}_{t+1} = \hat{\rho}y_t + \hat{a}_0(1 - \hat{\rho}) + \hat{a}_1(x_{t+1,1} - \hat{\rho}x_{t,1}) + \dots + \hat{a}_p(x_{t+1,p} - \hat{\rho}x_{t,p})$$

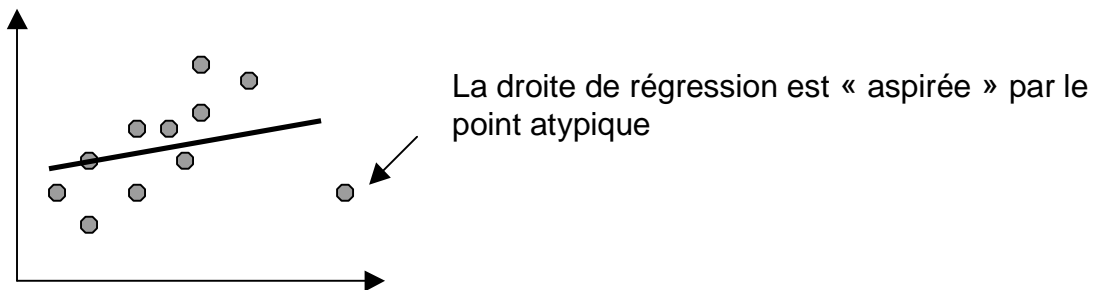
3) Utilisation des variables indicatrices

Une variable indicatrice est une variable qui

☞ *prend ses valeurs dans l'ensemble $\{0,1\}$*

☞ *sert à indiquer la survenance d'un événement ou de l'appartenance à un groupe*

A) Traitement des données aberrantes (atypiques)



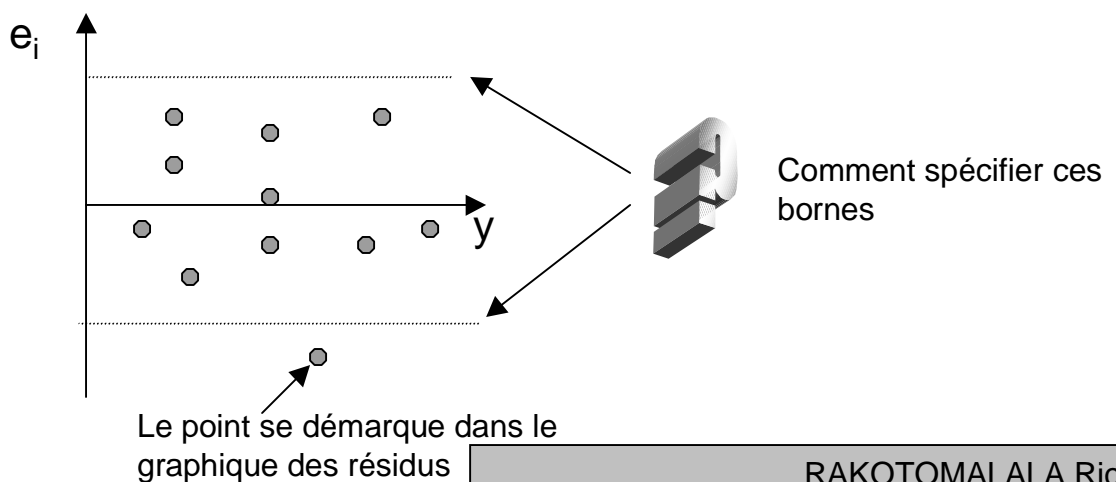
Causes probables :

☞ erreur de mesure ou de saisie

☞ événement exceptionnel

☞ l'observation n'appartient pas à la population étudiée

Détection : graphique des résidus



On sait par hypothèse que

$$\varepsilon \equiv \text{Normale}(0, \sigma_\varepsilon^2)$$

Estimé à l'aide de $\hat{\sigma}_\varepsilon^2$



Règle des 3 sigmas : 99,9% des observations sont situés à $\pm 3\sigma$ de part et d'autre de la moyenne

Traitement : utilisation d'une variable indicatrice, dite variable muette (dummy variable)

Y	X	d
5	3	0
6	1	0
6.5	2.5	0
5.2	2	0
14.5	3.5	1
3	1.5	0

Afin de supprimer l'effet levier dû à l'observation atypique

$$y = \hat{a}_0 + \hat{a}_1 x + \hat{a}_2 d$$

Indique le « décalage » par rapport au modèle

B) Régression sur variables qualitatives

Dans le modèle linéaire général, jusqu'ici tous les x_j étaient quantitatifs

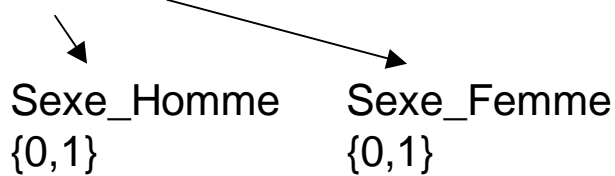


Comment faire pour introduire des variables qualitatives (ex: sexe {homme, femme}, statut marital {marié, célibataire, veuf})

Traitement : codage disjonctif complet des variables qualitatives

Sexe

{Homme,Femme}



Attention, danger de colinéarité :

$$\text{Sexe_Homme} + \text{Sexe_Femme} = 1$$

☞ *colinéarité avec le terme constant de la régression*

Solution : omettre une des modalités des variables
(attention par la suite dans l'interprétation des résultats)

sexe



Sexe_Homme
{0,1}

Sexe_Homme = 1 => Homme

Sexe_Homme = 0 => Femme

C) Analyse de la saisonnalité (cadre des observations temporelles)



Certaines grandeurs économiques (ventes, affluence, trafic ferroviaire...) sont influencés par les phénomènes saisonniers



Comment introduire cette information dans les modèles



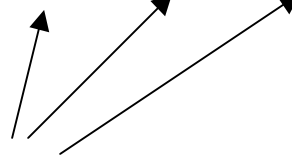
Utilisation des variables indicatrices, une pour chaque saison

Ex : données trimestrielles

☞ 4 variables indicatrices (T1,T2,T3,T4)

à cause de la colinéarité, on n 'en introduit que 3 dans la régression

$$ventes = \hat{a}_0 + \hat{a}_1 pub + \hat{a}_2 T_1 + \hat{a}_3 T_2 + \hat{a}_4 T_3$$



Indique le décalage moyen
par rapport au 4ème trimestre

D) Comparaisons de moyennes (introduction à l'analyse de variance)



Y a-t-il des facteurs de variations systématiques pesant sur une variable d'intérêt



On constitue des groupes, un pour chaque occurrence du facteur contrôlé, et on compare par la suite les moyennes respectives de la variable d'intérêt

Ex : poids des personnes x sexe

$$poids = \hat{a}_0 + \hat{a}_1 \times \text{sexe_homme}$$

Variable indicatrice

Poids moyen chez les femmes

Décalage moyen du poids de l'homme par rapport à celui de la femme