

Généralités sur l'Econométrie

Ricco Rakotomalala
Ricco.Rakotomalala@univ-lyon2.fr

PLAN

1. Econométrie : Origine(s), définition(s) et objectif(s)
2. La démarche économétrique
3. Analyse de régression - L'hypothèse de linéarité
4. Domaines d'application
5. Types de données
6. Bibliographie

Econométrie : Origine(s), Définition(s), Objectif(s)

Quelques définitions

Définition 1. Etudes des relations quantitatives de la vie économique faisant appel à l'analyse statistique et à la formulation mathématique.

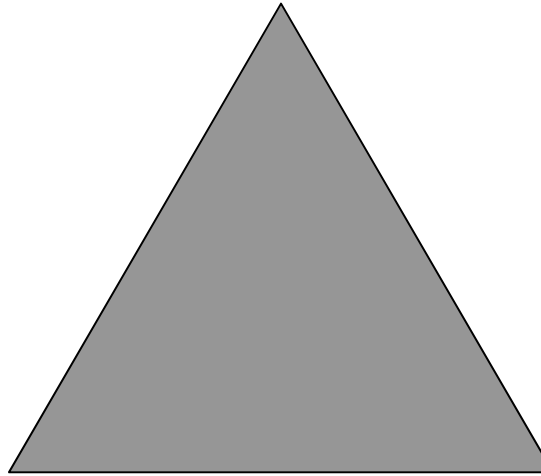
Définition 2. L'économétrie exprime quantitativement les corrélations pouvant exister entre des phénomènes économiques dont la théorie affirme l'existence. La théorie économique fournit des idées sur les processus qui déterminent les grandeurs économiques, l'économétrie apporte une vérification empirique et établit quantitativement les corrélations qui apparaissent valides.

Définition 3. L'objectif de l'économétrie est de confronter un modèle économique à un ensemble de données (données de panel, série temporelle, etc.) et ainsi d'en vérifier la validité.

Définition 4. L'économétrie est une branche de l'économie qui traite de l'estimation pratique des relations économiques.

Carrefour de 3 disciplines

Economiste (Expert du domaine)
Exprime une théorie sur un phénomène économique
Ex. La demande dépend du prix



Mathématicien (Modélisation)
Propose une formulation
algébrique de la théorie.
*Ex. Demande = $a * \text{prix} + b$*

Statisticien (Estimation)
Estime les paramètres du
modèle à partir de données.
Validation statistique.
Ex. $a = -0.5 ; b = 10$

Sous le contrôle de l'Economiste
Validation de l'Expert du domaine (ex. a est forcément négatif)

Notions clés – Modèle Economique

Un modèle consiste en une présentation formalisée d'un phénomène sous forme d'équations mathématiques.

Comme toutes les variables économiques sont interdépendantes (notion de système), il n'est pas suffisant de construire des équations isolées : il faut établir un système complet d'équations.

Exemple :

$O = f(p)$	}	Equations de comportement.	Théorie économique
$D = g(p)$			
$O = D + \Delta$	} Identité		

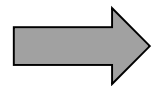
➔

$O = a \times p + b$	Modélisation (Introduction d'hypothèses simplificatrices sur la forme de la relation)
$D = \alpha \times p + \beta$	

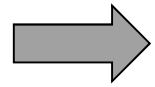
➔ Estimation de a , b , α et β à partir des données disponibles

Notions clés – Modèle Econométrique

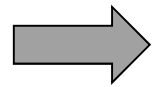
Faire intervenir l'aléatoire dans l'équation économique.
Parce que la relation n'est pas déterministe.



La spécification retenue est une simplification, il est évident qu'il ne résume pas toute la teneur de la relation (ex. dans les équations, la relation est vraiment linéaire ?)



Il y a d'autres facteurs dont on ne tient pas compte (ex. le prix des autres de biens qui peuvent se substituer au bien étudié)



Les erreurs de mesure sur les grandeurs étudiées, soit lors du processus de récolte des informations, soit tout simplement parce que la donnée récoltée représente peu ou prou le concept que l'on veut étudier.

**Introduction du facteur « aléatoire »
Résumé de toute l'information non prise
en compte dans le modèle**

$$O = a \times p + b + \varepsilon_O$$

$$D = \alpha \times p + \beta + \varepsilon_D$$

Notions clés – Variable

Les variables représentent des grandeurs économiques observées ou mesurées. Ex. les quantités vendues d'un bien, le prix d'un bien, des taux d'intérêt, le solde d'une balance commerciale, le taux de change, etc.

La variable doit être représentative du phénomène que l'on étudie, de sa qualité dépend la validité des résultats obtenus

Problèmes sur les variables

→ Problèmes d'inadéquation (étudier les ventes de pain, et utiliser des données mesurant les ventes de biscottes)

→ Erreur de mesures (problèmes lors du recueil des données ou des transmissions des données), d'unités (compter en nombre de pain vendu, ou en chiffre d'affaires)

→ Problème de représentativité (mesurer uniquement des ventes des boulangeries, et ne pas tenir compte des ventes en grande surface)

Notions clés – Variable aléatoire

Une variable aléatoire est une grandeur mesurable dont les valeurs sont soumises à une certaine dispersion lors de la répétition d'un processus donné.

La dispersion d'une variable aléatoire est régie par une **loi de probabilité** .

Ex. le résultat du jet d'une pièce de monnaie est une variable aléatoire, il prend deux valeurs possibles « pile » ou « face », il suit une loi de Bernouilli de paramètre $p = 0.5$.

Remarque : à chaque phénomène étudié sa loi de probabilité.

Ex. Durée entre deux phénomènes, nombre d'occurrence d'un phénomène dans un laps de temps, nombre d'essais avant d'obtenir un résultat, etc.

Notions clés – Types de variables

Quantitative

Qualitative nominale

Qualitative ordinale

Success	Wages	Job	Refunding
Y	0	Unemployed	Slow
N	2000	Skilled Worker	Slow
N	1400	Worker	Slow
N	1573	Retired	Slow
Y	2776	Skilled Worker	Slow
N	2439	Retired	Fast
N	862	Office employee	Slow
Y	1400	Salesman	Slow
N	1700	Skilled Worker	Slow
Y	785	Employee	Fast
Y	1274	Worker	Slow
N	960	Employee	Fast
N	1656	Worker	Fast
N	0	Unemployed	Slow

Le critère le plus important pour distinguer les variables est de déterminer si l'écart entre deux valeurs a un sens, et qu'elles sont comparables deux à deux.

Ex. Age, Salaires, Satisfaction, Type d'études suivies,...

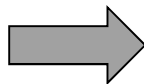
Notions clés – Population et échantillon

La population définit l'ensemble d'individus sur lesquels nous voulons travailler : on parle alors de population de référence ou de population parente ou population mère (ex. les véhicules vendus en France en 2005, etc.). Tous les résultats obtenus sont toujours relatifs à (circonscrites à) une population.

Les enquêtes exhaustives consiste à observer tous les individus qui composent la population. Opération très coûteuse.

On procède alors à un échantillonnage, on prélève une fraction de la population en veillant à ce qu'il soit représentatif de la population c.-à.-d refléter la composition et la complexité de la population.

Le taux de sondage correspond au rapport entre la taille de l'échantillon et la taille de la population.

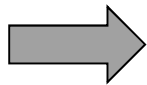


**Attention au mauvais échantillonnage.
Comment s'assurer que l'échantillon est représentatif ?
Rôle des variables de contrôle.**

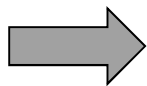
Notions clés – Inférence statistique

Inférence statistique. Elle consiste alors à effectuer des études sur l'échantillon et transposer les résultats sur la population.

Cette transposition n'est pas stricte, elle attache toujours une probabilité aux résultats et aux conclusions émises.

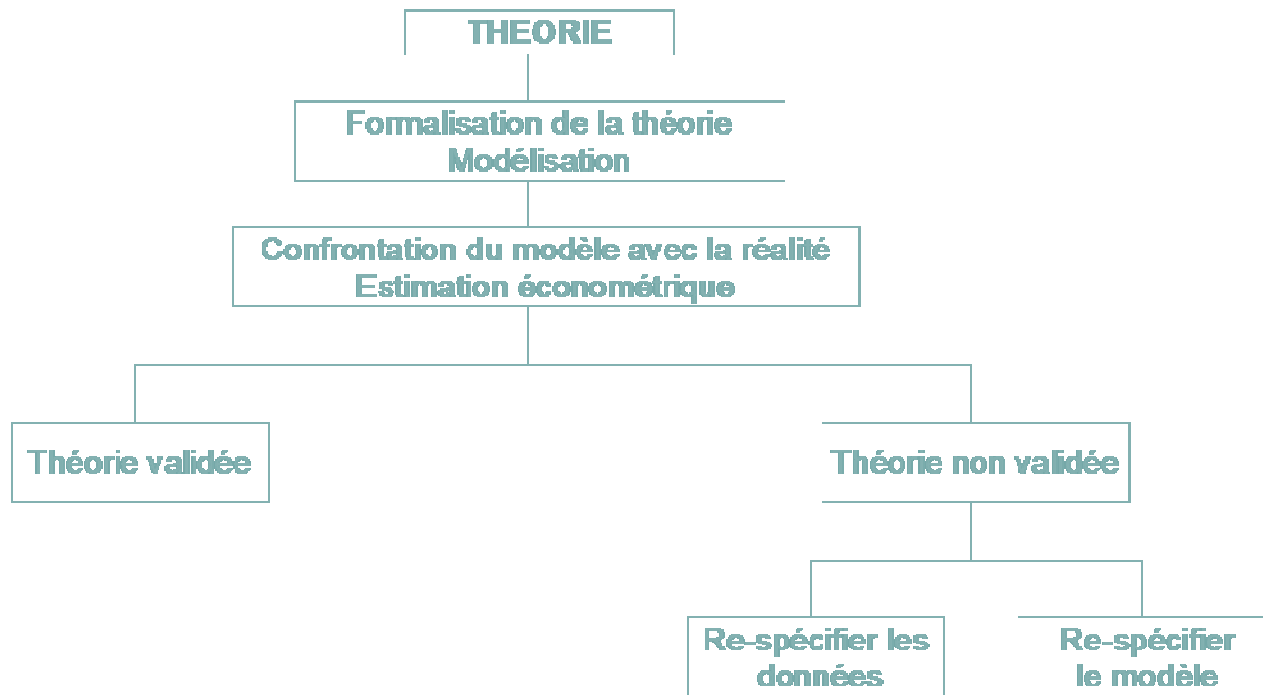


Tirer des conclusions sur l'existence ou non d'un phénomène (test d'hypothèses - ex. l'augmentation du prix du tabac réduit-t-il vraiment la consommation de cigarettes ?)

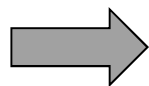


Estimer les paramètres d'un phénomène (estimation de paramètres - ex. une augmentation de 1 euro du prix du paquet de cigarette réduit de combien le nombre de paquets vendus ?)

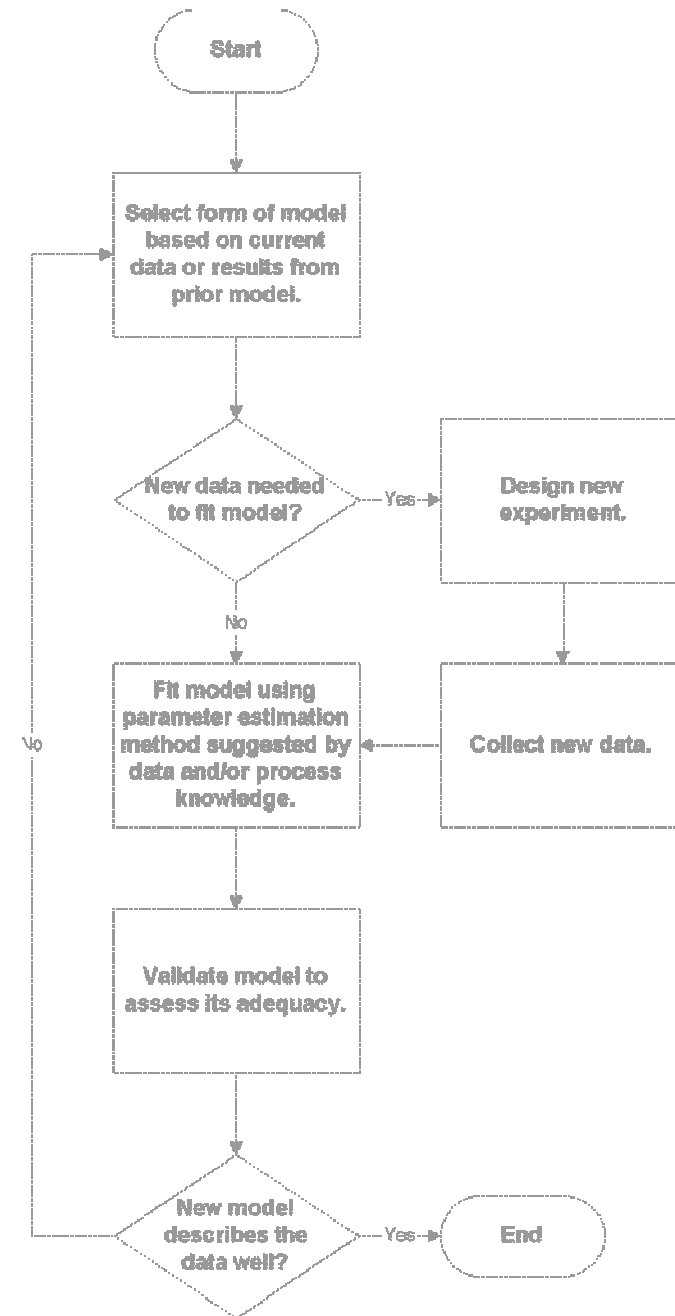
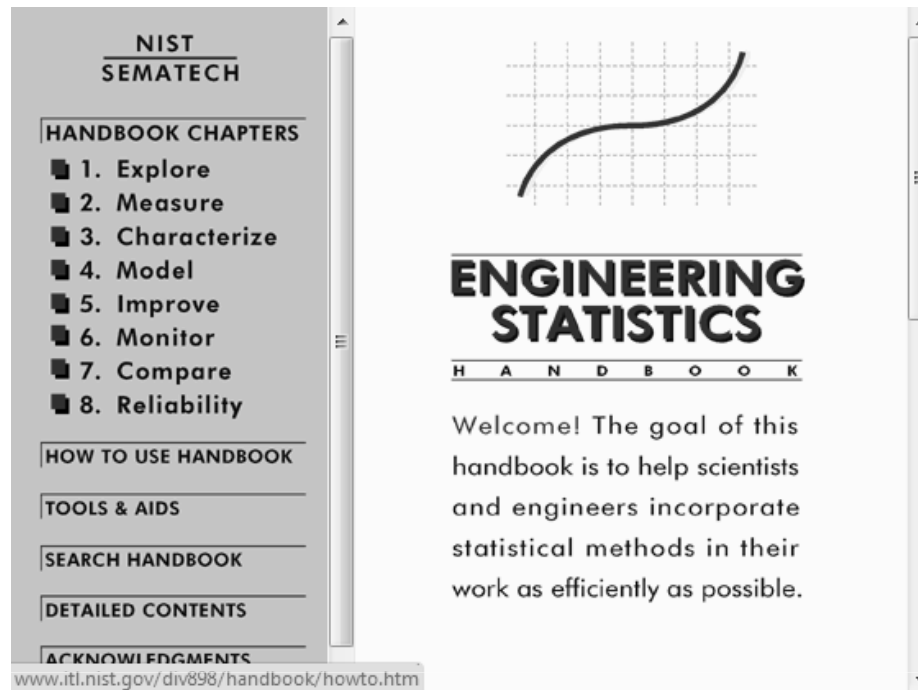
La démarche économétrique



Attention : Distinguer ce qui relève de la simple régularité statistique (artefact) de ce qui représente une causalité économique.



La théorie économique (la connaissance du domaine) est un garde-fou indispensable.



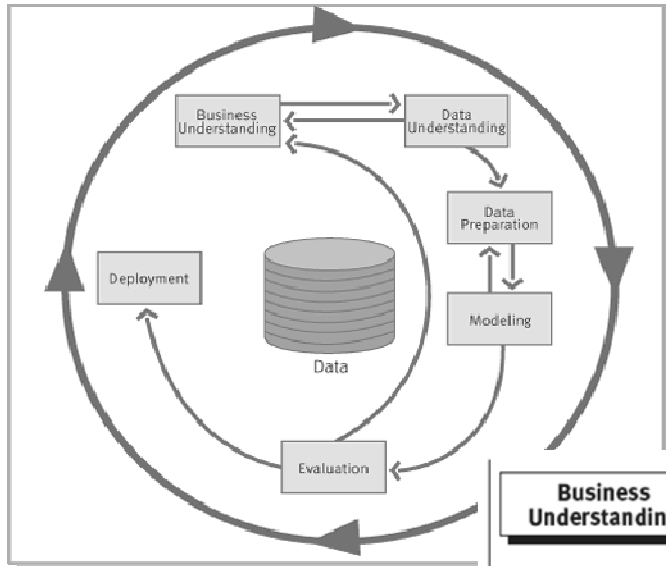
4.4.1. What are the basic steps for developing an effective process model?

Basic Steps Provide Universal Framework The basic steps used for model-building are the same across all modeling methods. The details vary somewhat from method to method, but an understanding of the common steps, combined with the typical underlying assumptions needed for the analysis, provides a framework in which the results from almost any method can be interpreted and understood.

Basic Steps of Model Building The basic steps of the model-building process are:

1. model selection
2. model fitting, and
3. model validation.

Source: CRISP-DM 1.0, Step-by-step Data Mining Guide, SPSS Publication



Business Understanding	Data Understanding	Data Preparation	Modeling	Evaluation	Deployment
<p>Determine Business Objectives Background Business Objectives Business Success Criteria</p> <p>Assess Situation Inventory of Resources Requirements, Assumptions, and Constraints Risks and Contingencies Terminology Costs and Benefits</p> <p>Determine Data Mining Goals Data Mining Goals Data Mining Success Criteria</p> <p>Produce Project Plan Project Plan Initial Assessment of Tools and Techniques</p>	<p>Collect Initial Data Initial Data Collection Report</p> <p>Describe Data Data Description Report</p> <p>Explore Data Data Exploration Report</p> <p>Verify Data Quality Data Quality Report</p>	<p>Select Data Rationale for Inclusion/ Exclusion</p> <p>Clean Data Data Cleaning Report</p> <p>Construct Data Derived Attributes Generated Records</p> <p>Integrate Data Merged Data</p> <p>Format Data Reformatted Data Dataset Dataset Description</p>	<p>Select Modeling Techniques Modeling Technique Modeling Assumptions</p> <p>Generate Test Design Test Design</p> <p>Build Model Parameter Settings Models Model Descriptions</p> <p>Assess Model Model Assessment Revised Parameter Settings</p>	<p>Evaluate Results Assessment of Data Mining Results w.r.t. Business Success Criteria Approved Models</p> <p>Review Process Review of Process</p> <p>Determine Next Steps List of Possible Actions Decision</p>	<p>Plan Deployment Deployment Plan</p> <p>Plan Monitoring and Maintenance Monitoring and Maintenance Plan</p> <p>Produce Final Report Final Report Final Presentation</p> <p>Review Project Experience Documentation</p>

L'analyse de régression

L'hypothèse de linéarité (spécification)

Analyse de régression – Schéma de régression

Modèle à une équation :

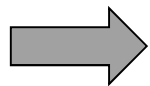
$$Y = f(X_1, X_2, \dots, X_p) + \varepsilon$$

Prédiction / Explication : Prédire/expliquer les valeurs de Y à partir des valeurs de X1, X2, ..., Xp.

Y est dite « variable endogène », c'est la variable donc on essaie de prédire les valeurs (variable à prédire, variable dépendante, expliquée) ;

X1...Xp sont les « variables exogènes », ce sont les variables qui servent à prédire les valeurs de Y (variables prédictives, variables indépendantes, explicatives).

Les valeurs des X sont donc connues (ou mesurées rapidement, facilement), elles servent à prédire les valeurs des Y qui sont inconnues (ou connues avec retard).



Ex 1. Prédire les ventes nationales de pain sur l'année (connu uniquement à la fin de l'année) à partir de son prix (connu instantanément).

Ex 2. Expliquer la consommation des pays européens à partir du revenu et du taux de chômage.

Régression linéaire multiple

Le modèle parfait n'existe pas. On procède très souvent à une simplification supplémentaire en considérant que la liaison est linéaire, ou encore on procède à des transformations (de variables) de manière à se ramener à combinaison linéaire des variables exogènes.

→ Il faut pouvoir estimer les paramètres, il faut pouvoir les interpréter !!!

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \dots + a_p X_p + \varepsilon$$

Y quantitative (forcément).
X quantitative ou qualitative
recodée (0/1).

ε est le terme d'erreur. C'est une variable aléatoire. Elle résume tout ce que le modèle n'explique pas.

X est supposé non aléatoire.

Y est aléatoire à cause de ε .

Linéarité par rapport aux paramètres

$$Y = a_0 + a_1 \ln(X_1) + a_2 X_1 + a_3 X_1^2$$

C'est un modèle linéaire.
Cf. Transformation de variables.

$$Y = \frac{\alpha_0 + \alpha_1 X_1}{\beta_0 + \beta_1 X_2}$$

Ce n'est pas un modèle linéaire.

$$Y = b \times e^{aX}$$
$$\ln(Y) = \ln(b) + aX$$

Linéaire après transformation.

Evaluation de la régression linéaire

Quel est le pouvoir explicatif du modèle ? Est-ce la liaison découverte entre Y et les X est significative ? (c.-à-d. transposable dans la population et non pas propre à l'échantillon observé)

Quel est l'apport marginal de chaque variable X dans l'explication des valeurs de Y ? (c.-à-d. un paramètre est-il significativement différent de 0 ?)

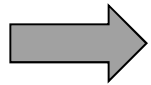
Quelle sont les propriétés (notamment la précision) des paramètres « a » obtenus ? (biais, variance)

Quelle sera la qualité de la prédiction des valeurs de Y à partir des valeurs de X ? (intervalle de prédiction, fourchettes)

Usages et domaines d'application

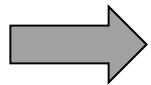
Usage de la régression linéaire

L'explication. Comprendre la nature des liaisons entre les variables. On parle également d'**analyse structurelle**.



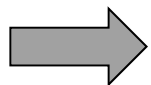
Outil privilégié pour valider les théories émises par les économistes.

Ex. $consommation = a * revenu + b$: $b > 0$, c'est la consommation incompressible, a est positif et sûrement inférieur à 1, $[1-a]$ correspond alors au taux d'épargne des ménages)



La prédiction. Premier usage opérationnel de la régression. Pour l'anticipation et la prise de décision.

Ex. La consommation des ménages va augmenter l'année prochaine ?



La simulation et la définition des politiques économiques. Second usage opérationnel de la régression. Permet de définir (1) les bonnes politiques économiques et (2) d'en mesurer à l'avance les conséquences.

Ex. Fixer la bonne valeur de la « prime à la casse ».

Autres domaines d'application

Tous les domaines où on essaie de détecter des relations de causalité

Economiste \Leftrightarrow Expert du domaine

Marketing. Evaluer le budget publicitaire nécessaire à une augmentation significative des ventes.

Sociologie. Prédire le niveau des notes des étudiants à partir de leur âge ou du nombre de redoublements. Expliquer le niveau d'études atteint par les étudiants à partir de la profession et des revenus des parents...

Agriculture. Evaluer les rendements des parcelles de terrains à partir de la quantité d'engrais utilisés ou du nombre de jours de pluie dans l'année.

Ecologie. Estimer la mortalité des poissons à partir de la quantité de résidus rejetés par les usines dans les cours d'eau.

Santé. Evaluer l'influence des compléments alimentaires sur la fréquence des maladies cardio-vasculaires (cf. par exemple les oméga 3 et les maladies cardio-vasculaires).

Types de données

Recueil des données

Problème récurrent : le manque de données pertinentes.

Ex. Analyse des processus de blanchiment d'argent

Données brutes vs. données corrigés normalisées

Données brutes : recueillies directement sur le terrain, très bonne qualité si précautions de recueil prises.

Données corrigées (institut de sondages) : + normalisation des définitions ; - déjà manipulées et corrigées, attention.

Données expérimentales vs. données non-expérimentales

Données expérimentales : contrôlées dans une expérimentation (ex. doses de médicaments pour un cobaye).

Données non expérimentales : directement observées.

→ X peut être expérimental ; Y est toujours observé.

Données transversales, longitudinales, de panel

Coupes transversales

Ligne = individu

Ex. Personne, véhicule, client, parcelle de terrain, etc.

Parcelle	Rendement (quintal)	Engrais (kilo)
A	16	20
B	18	24
C	23	28
D	24	22
E	28	32
F	29	28
G	26	32
H	31	36
I	32	41
J	34	41

Données temporelles (longitudinales)

Ligne = date

« Stock », définie sur une date

« Flux » définie sur une période

Stock → Flux facile (ex. somme, moyenne)

Flux → Stock pas évident (Mars = 5000 euros de CA, comment définir la valeur pour la date du 15 mars ?)

Mois	CA (K-euros)	Prospectus distribués
janv-04	1250	156
févr-04	1456	178
mars-04	4863	293

Données de panel

Faire des coupes transversales sur plusieurs dates.

Si on observe spécifiquement les mêmes individus, on parle de « cohorte ».



Ex. Recueillir les ventes d'un échantillon de concessionnaires. Renouveler l'opération sur plusieurs mois.

Bibliographique

http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

Régis BOURBONNAIS, « Econométrie – Manuel et exercices corrigés », Dunod, 1998.

Y.Dodge, V.Rousson, « Analyse de régression appliquée », Dunod, 2004.

M. Tenenhaus, « Statistique : Méthodes pour décrire, expliquer et prévoir », Dunod, 2007.

René GIRAUD, Nicole CHAIX, « Econométrie », PUF, 1994. (il existe une version QSJ, plus accessible)

Jack JOHNSTON, John DINARDO, « Méthodes économétriques », ECONOMICA, 1997.