Université Paris I – Panthéon Sorbonne
DESS Conseil en Organisation et Stratégie

Économétrie

Le modèle de régression linéaire

Corinne Perraudin[1]

Année 2004-2005

[1]Merci de me faire part de vos remarques et corrections à apporter à ce document (Corinne.Perraudin@univ-paris1.fr)

Introduction

L'objectif de ce cours est de présenter les méthodes économétriques appropriées pour étudier un phénomène économique, à partir de données relatives à ce phénomène ainsi qu'à ses éventuels facteurs explicatifs. Il s'agit donc d'utiliser des observations, soit dans le temps, soit pour différents individus (qui peuvent être des entreprises, des individus, des pays, etc) pour une date donnée, pour mieux comprendre ce qui explique les variations du phénomène étudié, et mesurer l'influence de ses facteurs explicatifs. Ceci permet alors de tester la validité des théories économiques qui prédisent tel type de relations, mais aussi de construire des prévisions du phénomène étudié.

On peut alors chercher à étudier, pour ne citer que quelques exemples, ce qui détermine le volume de production d'un pays dans le temps, ou son niveau d'emploi dans le temps, ou encore le chiffre d'affaires des entreprises à une date donnée, ou les déterminants du salaire des individus.

L'économétrie a joué un rôle important dans le développement de la théorie économique, mais les méthodes développées en économétrie jouent aussi un rôle important dans d'autres sciences sociales où l'on ressent le besoin de construire et d'estimer des modèles formalisant des relations entre un phénomène et des variables explicatives de ce phénomène. Ainsi, le champ d'application des méthodes économétriques est très important.

L'étude économétrique d'un phénomène repose sur le choix d'un modèle théorique, décrivant le phénomène que l'on cherche à étudier. On définit les variables d'intérêt et on recueille (ou on récupère) les données appropriées à l'étude. Les méthodes économétriques sont alors utilisées pour réaliser le "mariage" entre théorie et données. Il s'agit par exemple de donner des valeurs aux paramètres qui définissent les fonctions de comportement, ceci relève de l'estimation (statistique inférentielle) qu'il faut compléter par l'évaluation de la fiabilité des résultats et la précision de l'estimation de ces paramètres.

Les méthodes économétriques relèvent de la théorie statistique. L'analyse statistique traditionnelle permet d'extraire de l'information d'un ensemble de données et de fournir des indications sur la qualité de cette information. Ainsi, les techniques statistiques peuvent être considérées comme une aide importante dans le processus de décision, permettant de combiner les théories économiques (ou l'intuition) et l'expérience avec une bonne compréhension des faits résumés par les données disponibles.

Les techniques de statistiques dites descriptives permettent d'étudier les caractéristiques élémentaires du phénomène considéré. Il s'agit de décrire les données à l'aide de graphiques (histogramme, etc) et d'indicateurs statistiques élémentaires (moyenne, médiane, variance, quartile). On pourra alors étudier, par exemple, les ventes des entreprises sur différents sites suite à l'introduction d'un nouveau produit, en donnant des réponses à des questions de type "quelles sont les ventes moyennes?" ou "Est-ce que certains sites sont en retard sur d'autres?"

Mais on peut aussi rechercher à expliquer le phénomène étudié, par exemple chercher pourquoi les ventes sont élevées sur certains sites alors qu'elles sont faibles sur d'autres. Il s'agit de trouver des variables explicatives des différences dans les

ventes des entreprises. Ceci nécessite le recours aux modèles de régression, qui cherchent à expliquer les variations d'une variable (dépendante ou à expliquer) en fonction des variations d'autres variables (dites explicatives). Ce type d'études relève plutôt de l'économétrie, méthodes qui utilisent bien évidemment les méthodes statistiques, et qui consistent à mettre à l'épreuve des théories économiques formalisant des relations entre les variables, en les confrontant, grâce à des méthodes statistiques, aux observations des phénomènes étudiés.

L'économétrie est donc un ensemble de méthodes statistiques appliquées à l'économie, qui consiste à utiliser les données disponibles pour un certain nombre de variables pour estimer les relations économiques.

Ainsi, parmi les principaux objectifs de l'économétrie, on peut citer :

- Fournir des estimations des paramètres inconnus qui interviennent dans les modèles économiques, comme par exemple l'élasticité-prix de la demande ou des paramètres des fonctions de comportements, ainsi que de mesurer la validité des théories à l'aide de données observables.

- Trouver les déterminants significatifs d'une variable, par exemple de quoi dépendent les ventes d'une entreprise.

- Mesurer l'intensité des relations entre les variables, par exemple mesurer la valeur de l'influence des prix de vente sur les ventes d'une entreprise ou le lien entre la production et l'investissement.

- Faire des prévisions pour les variables d'intérêt.

L'objectif de ce cours est de comprendre les principes des méthodes économétriques relatives au modèle de régression linéaire. Outre le fait que ce type de modèles rend compte d'une large palette de phénomènes économiques, l'acquisition des principes de base de l'économétrie des modèles linéaires est une étape essentielle pour aborder des méthodes plus complexes.

Avant de présenter le modèle de régression linéaire étudié dans ce cours et les méthodes économétriques qui y sont relatives, nous présentons dans un premier chapitre préliminaire quelques éléments de statistiques permettant d'introduire des notions essentielles pour l'économétrie.

Bibliographie indicative

- Bourbonnais R., Econométrie, Dunod 2000 (3rd edition).

- Cadoret I., Benjamin C., Martin F., Herrard N., Tanguy S., Econométrie appliquée, DeBoeck, 2004.

- Dormont B., Introduction à l'économétrie, Montchrestien 1999.

- Greene W. H., Econometric analysis, Printice Hall 2000 (3rd edition)

- Johnston J. et Dinardo J., Méthodes économétriques (traduit par B. Guerrien), Economica, 1999 (4eme édition).
- Maddala G.S., Introduction to Econometrics, MacMillan.
- Pindyck R.S. et Rubinfeld D.L., Econometric Models and economic forecasts, McGraw-Hill, 1984.
- Siegel A.F., Practical Business Statistics, IRWIN, 1997 (3rd edition).

Chapitre 1 :
Quelques éléments de statistiques

Dans ce cours, nous allons chercher à étudier les facteurs explicatifs d'un phénomène. Nous considérerons ce phénomène comme une variable aléatoire, dont on cherchera à déterminer sa loi, ou plutôt le meilleur modèle pour le décrire, en fonction de variables explicatives. L'influence de ces variables explicatives est donnée par des paramètres, qui sont bien sur inconnus. L'objectif de l'économétrie est de fournir des valeurs pour ces paramètres inconnus, ou plus précisément de calculer des estimations. Ce premier chapitre introduit les notions de variables aléatoires, de loi, d'estimation, notions que l'on retrouvera dans l'étude du modèle de régression.

1 Variable aléatoire, espérance et variance

Une variable aléatoire est une variable qui peut prendre différentes valeurs, chacune avec une certaine probabilité. C'est donc une variable dont on ne sait pas avec certitude la valeur qu'elle va prendre (même si, après avoir réalisé l'expérience, la valeur qu'elle prendra sera un nombre unique). Ainsi, la notion de variable aléatoire formalise l'incertitude des situations.

Exemple : résultat du lancer d'une pièce de monnaie, d'un dé, mais aussi le sexe d'un enfant à naître, le résultat du loto, la température de demain à un certain endroit, les ventes d'une certaine entreprise en 2004, le salaire des gens en 2004 ou le PIB de la France en 2004.

On distingue les variables aléatoires **discrètes**, qui ne prennent que des valeurs isolées, parfois en nombre fini (par exemple, résultat du lancer de dé) et les variables aléatoires **continues**, qui prennent des valeurs dans un intervalle de $\mathbb{R}$ (ensemble des nombres réels) (par exemple, la taille d'une personne).

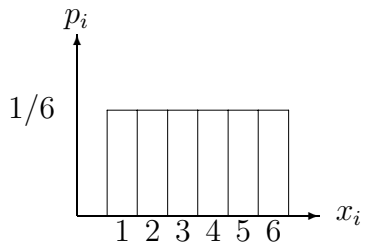
1.1 Variable aléatoire discrète

Une variable aléatoire discrète, notée X, est caractérisée par sa **loi**, ou **distribution de probabilité**, c'est-à-dire les différentes valeurs possibles x_i, $i = 1, \ldots, N$ qu'elle peut prendre et les probabilités d'apparition associées p_i, les probabilités mesurant les chances d'apparition de ces différentes valeurs, et vérifiant les propriétés suivantes : $p_i \in [0; 1]$ et $\sum_{i=1}^{N} p_i = 1$.

Par exemple, si X représente ce que l'on obtient quand on lance une pièce de monnaie, X est une variable aléatoire discrète, les valeurs possibles sont Pile et Face et les probabilités d'apparition, si la pièce est équilibrée, sont 0.5 et 0.5. Si X représente le résultat du lancer d'un dé, X est une variable aléatoire discrète pouvant prendre les valeurs $x_i = \{1, 2, 3, 4, 5, 6\}$ et les probabilités associées sont $p_i = \frac{1}{6} \ \ \forall i \in [1; 6]$.

On représente la loi d'une variable aléatoire discrète X par un histogramme, diagramme des probabilités, représentant les probabilités d'apparition p_i (en or-

données) pour chaque valeur possible x_i (en abscisses). Voir graphique ci-dessous pour la variable aléatoire correspondant au lancé d'un dé.



Une variable aléatoire peut aussi être caractérisée par des indicateurs. Les plus utilisés sont sa moyenne et sa variance, qui sont définis en termes de l'opérateur espérance E.

La **moyenne** d'une variable aléatoire X, ou **espérance**, est notée $E(X)$, et est définie, pour une variable aléatoire discrète, par :

$$E(X) = \sum_{i=1}^{N} p_i x_i$$

Ainsi, l'espérance de la variable aléatoire X correspondant au résultat d'un lancer de dé est donnée par $E(X) = \frac{1}{6}(1+2+3+4+5+6) = 21/6 = 3.5$

La **variance** donne une mesure de la dispersion des valeurs prises autour de la moyenne. La variance d'une variable aléatoire discrète est donnée par :

$$Var(X) = \sigma_X^2 = E[X - E(X)]^2 = E(X^2) - E(X)^2 = \sum_{i=1}^{N} p_i (x_i - E(X))^2$$

L'écart-type est la racine carrée de la variance. On le note $\sigma_X = \sqrt{Var(X)}$

La variance de la variable X correspondant au lancer de dé est 2.92 et son écart-type est 1.71

Propriétés:
Si X est une variable aléatoire et a et b sont des paramètres réels, alors :

$$E(aX+b) = aE(X) + b \quad E[(aX)^2] = a^2 E[X^2] \quad Var[aX+b] = a^2 Var(X)$$

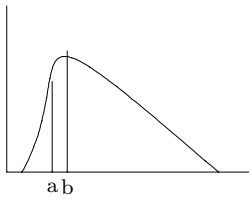
1.2 Variables aléatoires continues

Certaines variables peuvent prendre toutes les valeurs possibles qu'on trouve dans un intervalle (fini ou infini). Il s'agit par exemple de la taille d'un homme, de la durée d'un appel téléphonique, etc. On parle de variables aléatoires continues.

L'ensemble des valeurs possibles que peut prendre une variable aléatoire continue étant indénombrable, la définition des probabilités associées à chaque valeur n'est donc plus possible (elle n'a plus de sens).

Ainsi, la loi de probabilité d'une variable aléatoire continue est donnée, non pas par les probabilités d'apparition de chaque valeur possible, mais par sa fonction de densité (de probabilité) et par sa fonction de répartition.

La **fonction de densité**, souvent notée f, associe aux valeurs possibles x une valeur $f(x)$. On la représente par une courbe de densité, qui peut être considérée comme la courbe limite des diagrammes de fréquences des valeurs observées classées en des intervalles de plus en plus étroits. La surface d'une bande verticale mesure la probabilité que la variable aléatoire soit comprise entre a et b.



La **fonction de répartition** d'une variable aléatoire continue donne la probabilité que cette variable aléatoire prenne des valeurs inférieures à x, elle est alors donnée par :

$$F(x) = \int_{-\infty}^{x} f(u)du$$

L'espérance et la variance se définissent à l'aide d'intégrales (à la place des sommes), mais conservent la signification et les propriétés précédentes. L'espérance est alors donnée par :

$$E(X) = \int_{D(x)} f(x)dx$$

où $D(x)$ est le domaine des valeurs prises par X.

En général, on considère comme des variables aléatoires continues les variables aléatoires discrètes qui prennent un très grand nombre de valeurs très proches (le revenu, la taille d'une grande population, les ventes d'une entreprise, etc). Ce sont donc des variables aléatoires continues que l'on retiendra dans la suite du cours.

Il y a une famille de lois de variables aléatoires continues très importante en statistique et en économétrie, il s'agit de la **loi normale** (ou loi de Gauss). Elle intervient notamment dans la théorie de l'estimation et des tests. On l'utilisera beaucoup dans la suite de ce cours.

Les variables aléatoires, dont la loi est une loi normale, prennent leurs valeurs dans l'ensemble des nombres réels (x appartient à l'ensemble $]-\infty;+\infty[$), et une variable qui suit une loi normale a bien sûr une probabilité égale à 1 de prendre ses valeurs entre $-\infty$ et $+\infty$ ($\int_{-\infty}^{+\infty} f(x)dx = 1$ avec $f(x)$ définie ci-dessous).

Une variable aléatoire qui suit une loi normale est principalement caractérisée par deux paramètres, l'espérance et la variance.

La fonction de densité de probabilité d'une variable aléatoire X qui suit une loi normale d'espérance $E(X) = \mu$ et de variance $Var(X) = \sigma^2$, notée $X \sim \mathcal{N}(\mu, \sigma^2)$,

est bien connue et est donnée par :

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

avec x les valeurs prises par cette variable aléatoire X.

La courbe de densité d'une telle variable est une courbe en cloche (dite courbe de Gauss) symétrique autour de μ et plus ou moins évasée selon la valeur de σ.

Toute variable aléatoire X qui suit une loi $\mathcal{N}(\mu, \sigma^2)$ peut se ramener à une variable aléatoire qui suit la plus simple des lois normales, la **loi normale centrée réduite**, c'est-à-dire dont l'espérance est nulle et la variance vaut 1. Notons Z cette variable aléatoire qui suit une loi normale centrée réduite. On note $Z \sim \mathcal{N}(0, 1)$ et on a :

$$Z = \frac{X - \mu}{\sigma}$$

On vérifie en effet que $E(Z) = \frac{E(X)-\mu}{\sigma} = \frac{\mu-\mu}{\sigma} = 0$ et que $Var(Z) = \frac{Var(X)}{\sigma^2} = \frac{\sigma^2}{\sigma^2} = 1$.

La loi normale centrée réduite est définie par une courbe de densité symétrique autour de 0.

Différentes tables (que l'on peut trouver dans tous les ouvrages de statistiques ou d'économétrie) permettent de calculer la probabilité qu'a une variable aléatoire qui suit une loi normale centrée réduite de prendre des valeurs dans un intervalle donné.

Une valeur que l'on peut retenir est le chiffre de 1.96 pour une probabilité de 95% : une variable aléatoire normale centrée réduite a une probabilité de 0.95 (95% de chance) de prendre des valeurs dans l'intervalle $[-1.96; 1.96]$. On parlera d'intervalle de confiance à 95% (ou à 5% d'erreur) pour la variable $\mathcal{N}(0, 1)$:

$$Pr(-1.96 < Z < 1.96) = 0.95$$

De cette valeur, on en déduit un intervalle de confiance à 95% pour une variable X qui suit une loi $\mathcal{N}(\mu, \sigma^2)$:

$$Pr(\mu - 1.96\sigma < X < \mu + 1.96\sigma) = 0.95$$

puisque $X = \sigma Z + \mu$.

De manière générale, écrire $Pr(-1.96 < Z < 1.96) = 0.95$ est équivalent à écrire $Pr(|Z| > 1.96) = 1 - 0.95 = 0.05$. Pour une valeur quelconque du niveau de confiance $1 - \alpha$, ou autrement dit du risque d'erreur α, on note t_α la valeur telle que $Pr(|Z| > t_\alpha) = \alpha$.

Exemple : on suppose que la durée de vie en heures d'un certain type d'appareil électronique suit une loi normale $\mathcal{N}(10000, 2000^2)$. Un appareil donné a alors une probabilité de 0.95 de durer entre $10000 - 1.96 \times 2000$ heures et $10000 + 1.96 \times 2000$ heures, soit entre 6080 et 13920.

Une propriété importante, qui explique l'importance des lois normales, est que la somme d'un grand nombre de variables aléatoires indépendantes, et dont aucune

n'est d'échelle prépondérante, suit approximativement une loi normale (c'est l'énoncé littéraire du théorème central limite ou de la loi des grands nombres).

Ceci explique pourquoi beaucoup de variables qui peuvent être considérées comme la résultante additive d'un grand nombre de déterminants dont aucun n'est dominant, montrent des distributions normales (par exemple la taille des françaises).

Malgré l'importance des lois normales, on gardera à l'esprit qu'il existe bien d'autres lois de distribution, pas forcément bien connues d'ailleurs. Ainsi, toute variable ne suit pas forcément une loi normale. Ainsi, quand on fait l'hypothèse qu'une variable aléatoire suit une loi normale, il faudra vérifier si cela est acceptable ou pas par l'examen des observations de l'échantillon recueilli.

Dans la suite du cours, nous utiliserons d'autres lois qui découlent de la loi normale et qui sont aussi très utilisées en économétrie : les lois de Student, les lois de Fisher, les lois du chi-deux. Nous les définirons plus loin.

2 Covariance entre deux variables aléatoires

On définit la covariance entre deux variables aléatoires X et Y comme l'espérance du produit des deux variables aléatoires quand les deux sont définies en déviation à leur moyenne :

$$Cov(X,Y) = E[(X - E(X))(Y - E(Y))] = E(XY) - E(X)E(Y)$$

La covariance est une mesure de la relation linéaire entre les deux variables. Si les deux variables sont "en même temps" inférieures à leur moyenne respective et "en même temps" supérieures à leur moyenne respective, alors leur covariance est positive. Si, quand l'une est en dessous de sa moyenne, l'autre est au dessus de sa propre moyenne, alors la covariance est négative.

Par exemple, les variables aléatoires taille et poids des gens ont sans doute une covariance positive, alors que la rapidité aux 100 mètres et l'âge des athlètes ont sans doute une covariance négative.

La valeur de la covariance dépend des unités dans lesquelles X et Y sont mesurées. On préfère alors le coefficient de corrélation, défini en dehors de toute mesure :

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Il varie entre -1 et +1. S'il est positif, c'est que les deux variables varient dans le même sens, alors que s'il est négatif, c'est qu'elles varient en sens opposé. S'il est proche de 1 ou de -1, c'est que la relation entre les variables est forte, alors que s'il est proche de 0, c'est que les variations des variables ne sont pas (ou peu) liées.

Ce coefficient de corrélation nous indique à quel point la relation entre les deux variables est forte (mais attention, on ne sait rien sur le sens de la causalité). Nous le retrouverons dans l'étude de la régression linéaire.

Propriétés:
Si X et Y sont deux variables aléatoires, alors :

$$E(X+Y) = E(X) + E(Y) \quad Var(X+Y) = Var(X) + Var(Y) + 2Cov(X,Y)$$

On dit que deux variables aléatoires X et Y sont indépendantes si le fait d'avoir une certaine valeur pour X n'est pas relié à la valeur de Y et vice versa. Dans ce cas, on a :

$$E(XY) = E(X)E(Y) \text{ et donc } Cov(X,Y) = 0$$

Attention, deux variables peuvent avoir une covariance nulle et ne pas être indépendantes! par exemple si la relation qui les lie est plutôt non linéaire.

3 Estimation

L'estimation est l'un des buts de la statistique et de l'économétrie. L'idée est de relier une valeur **non aléatoire mais inconnue** (la vraie valeur qui nous intéresse) à une valeur **aléatoire mais observable** (l'estimation de la valeur inconnue), en mesurant autant que possible la précision de l'estimation, c'est-à-dire la confiance à accorder à cette valeur.

Supposons, par exemple, que l'on s'intéresse à l'espérance inconnue d'une variable aléatoire (par exemple la durée moyenne d'un certain équipement) ou alors, la moyenne d'une variable sur une population très grande (la taille moyenne des terriens). Elles prennent une certaine valeur (non aléatoire) mais inconnue (ou alors connue, mais très coûteuse à obtenir dans le cas d'une population de très grande taille). En revanche, si nous disposons d'un échantillon, c'est-à-dire d'observations du phénomène étudié (soit plusieurs équipements, soit un sondage sur une partie de la population), on va pouvoir estimer l'espérance théorique à partir des informations issues de l'échantillon dont on dispose, c'est-à-dire trouver une valeur pour approcher l'espérance inconnue, mais cela sur la base d'un échantillon particulier.

Pour illustrer les notions d'échantillon et d'estimation, prenons tout d'abord le cas d'une variable aléatoire discrète. Supposons par exemple que l'on s'intéresse à la probabilité d'apparition de Pile quand on lance une pièce de monnaie, ou la probabilité d'apparition d'une certaine valeur quand on lance un dé. Bien sûr, on sait que la probabilité théorique d'obtenir Pile pour une pièce équilibrée est 0.5 et que la probabilité théorique d'obtenir 1 quand on lance un dé est 1/6. Mais supposons que l'on ne connaisse pas ces probabilités théoriques, ou alors que l'on souhaite vérifier que la Pièce ou le dé considéré ne sont pas pipés, et que l'on cherche à estimer ces probabilités d'apparition. C'est ce qui se passe en général pour les variables aléatoires que l'on étudie : on ne connaît pas leur loi, et on ne peut qu'obtenir des estimations des paramètres de la loi.

Pour estimer la probabilité d'obtenir Pile, on va lancer un grand nombre de fois la pièce de monnaie et on va observer les **réalisations** de la variable aléatoire. Ces observations ont une valeur déterminée, donnée, et il n'y a plus aucun caractère d'incertitude. Cependant, ce sont les résultats d'une expérience aléatoire. A partir de ces observations, qui constituent un échantillon, on va pouvoir calculer

la fréquence d'apparition de Pile lors de nos n lancers (nombre de fois où on a obtenu Pile divisé par nombre de lancers), on parle de **fréquence empirique**. La fréquence empirique a de "bonnes propriétés" dans le sens où elle s'approche de la probabilité théorique (p_i) quand n devient grand. En effet, en augmentant le nombre de répétitions, on s'attend à obtenir des fréquences (empiriques) de plus en plus proches des probabilités (théoriques).

Supposons que l'on lance 10 fois une pièce de monnaie, et que l'on obtienne $\{P; F; F; P; F; P; P; F; F; F\}$. La fréquence d'apparition de Pile de 0.4. Si on refait l'expérience de lancer 10 fois la pièce, on risque d'obtenir d'autres valeurs pour la fréquence. C'est pour cela que l'on parle de fréquence empirique, alors que la probabilité d'obtenir Pile reste toujours égale à 0.5, d'où le nom de probabilité théorique. En augmentant le nombre de répétitions, on s'attend à obtenir des valeurs (pour la fréquence empirique) très proche de 0.5 (probabilité théorique).

Supposons maintenant que l'on s'intéresse au lancé d'un dé et à la valeur moyenne obtenue. Pour un échantillon de taille n (valeurs des n observations quand on a répété n fois l'expérience), on peut calculer la moyenne des valeurs, appelée **moyenne empirique** et généralement notée $\bar{x}$:

$$\bar{x} = \sum_{i=1}^{n} f_i x_i = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Par exemple, supposons qu'on lance un dé 20 fois et que l'on obtienne

$$\{5; 4; 6; 2; 3; 2; 1; 1; 6; 3; 3; 6; 2; 6; 4; 6; 3; 6; 3; 4\}$$

alors $\bar{x} = 3.8$, alors que $E(X)$, la moyenne théorique, reste égale à 3.5. De la même manière que les fréquences empiriques s'approchent des probabilités théoriques quand n devient très grand, on s'attend à ce que la moyenne empirique s'approche de la moyenne théorique (espérance) quand n devient très grand.

D'une manière générale, supposons que l'on s'intéresse à la meilleure estimation possible de l'espérance (inconnue) d'une variable aléatoire. Supposons que l'on dispose d'un échantillon de taille n de cette variable aléatoire. On cherche à déterminer une règle, qui nous donne une valeur pour chaque échantillon possible. On parle **d'estimateur**. La valeur obtenue sur un échantillon particulier sera appelée une **estimation** (on dit aussi une estimation ponctuelle). L'estimateur est une règle alors que l'estimation donne une valeur. L'estimateur est donc une variable aléatoire puisqu'il dépend de l'échantillon. Pour un échantillon donné, on va obtenir une valeur, mais on obtiendrait sans doute une autre valeur si on disposait d'un autre échantillon.

On demande à l'estimateur qu'il ait de "bonnes propriétés", notamment qu'il soit **non biaisé** (ou sans biais), c'est-à-dire qu'en moyenne, il donne la vraie valeur du paramètre que l'on cherche. On veut de plus qu'il soit **convergent**, c'est-à-dire que sa variance tende vers 0 quand on augmente le nombre d'observations de l'échantillon. Ainsi, on s'assure d'obtenir une estimation proche de la vraie valeur du paramètre que l'on cherche à estimer.

Supposons que l'on s'intéresse à une variable aléatoire X qui suit une certaine loi (inconnue) d'espérance μ et de variance σ^2, μ et σ^2 étant inconnus.

Un échantillon de taille n est une série de variables aléatoires $X_1, X_2, \ldots, X_n$ indépendantes, suivant la même loi que X.

Un "bon" estimateur de μ, l'espérance inconnue de X, est la moyenne empirique, définie par :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Il est très important de noter que $\bar{X}$ est une variable aléatoire, dont la valeur dépend des valeurs prises par $X_1, X_2, \ldots, X_n$, même si la moyenne de la population μ reste inchangée. Ceci car $\bar{X}$ dépend des observations que l'on va observer et donc de l'échantillon considéré.

En effet, il n'y a pas de raison que, si l'on tire aléatoirement 1 échantillon de 1000 personnes dans la population totale et qu'on leur demande leur taille, la moyenne ($\bar{X}$) soit égale à la vraie valeur (μ).

Si on recommence 100 fois (soit 100 échantillons de 1000 personnes), il n'y a pas de raison qu'on trouve 100 fois la même moyenne de la taille. On trouvera 100 valeurs différentes, mais on espère que la moyenne de ces 100 valeurs soit très proche de la vraie taille moyenne de la population (estimateur sans biais).

On conçoit bien aussi que si on augmente le nombre de personnes dans l'échantillon (de 1000 on passe à 100000), on va se rapprocher de la vraie valeur et on va avoir un nombre plus précis (estimateur convergent).

On peut montrer que $\bar{X}$ est un "bon" estimateur, c'est-à-dire qu'il est sans biais (en espérance, on obtient la vraie valeur inconnue du paramètre μ) :

$$E(\bar{X}) = E\left(\frac{\sum_i X_i}{n}\right) = \frac{\sum_i \mu}{n} = \mu$$

et convergent (l'estimation est plus précise quand le nombre de répétitions augmente) :

$$Var(\bar{X}) = \frac{\sum_i \sigma^2}{n^2} = \frac{\sigma^2}{n} \rightarrow^{n \rightarrow \infty} 0$$

Supposons que, maintenant, on s'intéresse à la différence de taille des individus, donc à la variance de la taille. Un choix raisonnable pour estimer la variance serait :

$$\frac{1}{n} \sum_i (X_i - \bar{X})^2$$

Le problème est que cet estimateur est biaisé si la moyenne est inconnue et que l'on doit l'estimer. Un estimateur sans biais de la variance est :

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

L'idée est que notre échantillon contient n observations, mais pour estimer la variance, on a besoin d'estimer au préalable la moyenne. Ceci impose une contrainte entre les n observations, qui laisse $n-1$ observations non contraintes avec lesquelles on peut estimer la variance. On dit qu'on a perdu un degré de liberté.

Si on s'intérrese maintenant à la relation entre la taille et le poids, on va alors chercher une estimation de la covariance entre ces deux variables. Pour calculer la covariance empirique, on retiendra :

$$\widehat{cov}(X,Y) = \frac{1}{n-1}\sum_i (X_i - \bar{X})(Y_i - \bar{Y})$$

Avec ces estimateurs empiriques, on peut définir le coefficient de corrélation empirique, noté r_{XY} :

$$r_{XY} = \frac{\widehat{cov}(X,Y)}{S_X S_Y}$$

où S_X et S_Y sont les estimateurs de la variance de X et de Y.

On a vu comment obtenir des estimations de la moyenne, de la variance et de la covariance pour un échantillon donné. Mais, quelle est la confiance que l'on peut accorder à ces valeurs obtenues (quand on cherche la vraie valeur d'un paramètre)?

Pour le savoir, il faut construire des intervalles de confiance. Pour cela, il faut connaître la loi de l'estimateur (par exemple de $\bar{X}$). Concernant $\bar{X}$, le théorème central limite nous donne un résultat très utile.

Le théorème central limite :

Si la variable aléatoire X est de moyenne μ et de variance σ^2, alors la distribution empirique de $\bar{X}$ devient approximativement normale de moyenne μ et de variance σ^2/n quand n augmente. On note :

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

On connaît alors la loi de la moyenne empirique[2]. On remarque que l'estimation de l'espérance théorique est d'autant plus précise ($var(\bar{X})$ est d'autant plus faible) que le nombre d'observations n de l'échantillon est élevé.

Puisque $\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, alors $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$. D'après ce que l'on a vu précédemment, on peut déduire un intervalle de confiance pour cette variable centrée réduite : elle a 95% de chance d'être comprise entre -1.96 et 1.96, ou $(1-\alpha)$% d'être comprise entre $-t_\alpha$ et t_α.

Ainsi, un intervalle de confiance pour la vraie valeur μ inconnue à un seuil d'erreur de α% est donné par :

$$[\bar{X} - t_\alpha \times \sigma/\sqrt{n}; \bar{X} + t_\alpha \times \sigma/\sqrt{n}]$$

En général, quand μ est inconnue, alors σ est aussi inconnue! On montre qu'il est acceptable, pour un échantillon de taille suffisante, de remplacer la valeur inconnue

[2]La loi de S^2 est présentée dans le paragraphe sur la loi du chi-deux.

de l'écart-type σ par son estimation S, calculée sur l'échantillon, et on peut alors calculer l'intervalle de confiance du paramètre μ inconnu, ou de manière équivalente, la précision de l'estimation effectuée sur l'échantillon. Cependant, dans ce cas, $\frac{\bar{X}-\mu}{S/\sqrt{n}}$ suit, non pas une loi normale centrée réduite, mais une loi de Student (que l'on présente ci-après). Disons juste que pour n suffisamment grand, les valeurs t_α correspondant à une loi normale et à une loi de Student sont les mêmes.

Par exemple, 15000 personnes ont passé un concours, 600 premières copies prises au hasard sont déjà corrigées et on a trouvé une moyenne de 11.3 et un écart-type de 2.1. On veut estimer la moyenne générale avec un risque de 5% d'erreur (ou à 95% de confiance). On obtient l'intervalle de confiance $[11.3 - 1.96 \times 2.1/\sqrt{600}; 11.3 + 1.96 \times 2.1/\sqrt{600}]$, soit $[11.13; 11.47]$.

Attention: on retiendra que l'estimation par intervalle de confiance donne une réponse en terme de précision (par l'intervalle proposé), mais aussi de risque (par la valeur du seuil d'erreur, contraire du seuil de confiance). Ainsi, un statisticien qui fait par exemple des estimations au risque de 10% dans des conditions correctes doit s'attendre à obtenir une fourchette erronée, c'est-à-dire ne contenant pas la vraie valeur, environ une fois sur 10.

4 Quelques lois dérivées de la loi normale

4.1 Loi du chi-deux

La somme du carré de n variables aléatoires indépendantes distribuées selon des lois normales centrées réduites (moyenne 0 et variance 1) est distribuée selon une loi du **chi-deux** à n degrés de liberté, notée χ^2_n.

La distribution du chi-deux est toujours positive, et est disymétrique.

Si on calcule la variance empirique S^2 de n observations tirées d'une distribution normale de variance σ^2, alors $(n-1)S^2/\sigma^2$ sera distribué selon un chi-deux à $n-1$ degrés de liberté.

4.2 Loi de Student

Si X est distribuée selon une loi normale centrée réduite, et que Z est distribuée selon un chi-deux à n degrés de liberté, et si X et Z sont indépendantes, alors $X/\sqrt{Z/n}$ est distribuée selon une **loi de student** à n degrés de liberté.

La loi de student ressemble à la loi normale en étend plus aplatie (elle devient normale pour des échantillons de grande taille).

Si X est normale, $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ est aussi normale de moyenne nulle et de variance égale à 1. Mais si σ^2 est inconnue, on doit la remplacer par la variance empirique s^2. Puisque $(n-1)s^2/\sigma^2$ suit un chi-deux et que $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ est $\mathcal{N}(0,1)$, alors

$$\frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)s^2/\sigma^2}}\sqrt{n-1} = \frac{(\bar{X}-\mu)}{s/\sqrt{n}}$$

suit une distribution de student à $n-1$ degrés de liberté.

Cela permet de tester si la moyenne d'une variable aléatoire est égale à une valeur donnée, même quand la variance est inconnue (on reverra la loi de student lors de l'étude des tests d'une seule contrainte dans le modèle de régression).

4.3 Loi de Fisher

Si X et Z sont indépendantes et distribuées selon des lois du chi-deux à n_1 et n_2 degrés de liberté, alors $\frac{X/n_1}{Z/n_2}$ est distribuée selon une **loi de Fisher** à n_1 et n_2 degrés de liberté.

La distribution de Fisher est centrée vers la gauche et toujours positive.

Cette distribution permet de tester des hypothèses jointes (impliquant plusieurs contraintes, comme on le verra plus tard dans le modèle de régression multiple), par exemple l'égalité de plusieurs paramètres à une certaine valeur, ou alors l'égalité des variances de deux échantillons (que l'on étudiera plus tard).

Par exemple, pour tester $\sigma_X^2 = \sigma_Y^2$ (variance respective de X et de Y), on peut calculer la statistique s_X^2/s_Y^2. Si X et Y sont indépendantes, alors $(n_1-1)s_X^2/\sigma_X^2$ est distribuée selon un chi-deux à $n_1 - 1$ degrés de liberté et $(n_2-1)s_Y^2/\sigma_Y^2$ est distribuée selon un chi-deux à $n_2 - 1$ degrés de liberté; alors

$$\frac{(n_1-1)s_X^2/\sigma_X^2}{n_1-1} \Big/ \frac{(n_2-1)s_Y^2/\sigma_Y^2}{n_2-1}$$

suit un Fisher à $n_1 - 1$ et $n_2 - 1$ degrés de liberté.

On reprendra cela plus tard, dans les tests de validation du modèle de régression estimé.

Chapitre 2 :
Le modèle de régression simple

L'analyse de la régression est l'outil le plus utilisé en économétrie. L'idée est de décrire et d'évaluer la relation entre une variable (dite variable à expliquer ou variable dépendante) souvent notée y, et une (ou plusieurs) variable(s), dite(s) variable(s) explicative(s) ou indépendante(s), à partir d'un échantillon de taille n, valeurs de ces variables pour n observations (individus ou pays ou dates).

Nous allons tout d'abord supposer qu'il n'y a qu'une seule variable explicative, notée x. On cherche alors à étudier la relation entre 2 variables, y et x, ou plus précisément l'influence de x sur y. On parle de régression simple. On étendra ensuite les résultats au cas de la régression multiple, c'est-à-dire lorsqu'il y a plusieurs variables explicatives (chapitre 3).

Pour étudier le lien entre les variables y et x, on doit tout d'abord se donner une relation (ou forme fonctionnelle) entre les deux. On retient une relation linéaire[3] :

$$y = a + bx$$

Puisque la relation qu'on étudie est une simplification de la réalité, que l'on a forcément oublié des facteurs explicatifs de y, on introduit un terme aléatoire (une perturbation, appelé encore résidu) noté u à cette relation déterministe. Ainsi le modèle que l'on étudie est le suivant :

$$y = a + bx + u$$

avec y la variable à expliquer, x la variable explicative, u le terme d'erreur, qui est supposé être une variable aléatoire (inobservable, mais dont on suppose que l'on connaît la loi).

a et b sont les paramètres inconnus, appelés coefficients de régression, que l'on cherche à déterminer (à estimer).

On suppose que l'on dispose d'un échantillon de n observations pour y et x, soit $\{(y_1, x_1), \ldots, (y_n, x_n)\}$.

Ainsi, pour chaque observation i, on peut écrire :

$$y_i = a + bx_i + u_i$$

L'objectif est de déterminer des valeurs pour les paramètres a et b à partir des n observations sur les variables x et y. Graphiquement, il s'agit de trouver les paramètres de la droite de régression, qui passe au milieu du nuage de points dessiné dans le plan (x, y).

On remarque que les paramètres a et b sont les mêmes pour toutes les observations, autrement dit que l'influence de x sur y est la même pour toutes les observations.

[3] Hypothèse peu restrictive, voir discussion plus loin.

Pour préciser la démarche adoptée en économétrie, prenons un exemple simple (très répandu dans les ouvrages d'économétrie), l'étude d'une fonction de consommation. La démarche de l'économètre se décompose comme suit :

a) Prendre pour point de départ une théorie économique, qui donne la relation à étudier :

   On va supposer que la consommation des agents dépend de leur revenu disponible (*Cf.* théorie keynésienne).

b) Formaliser le modèle et recueillir les observations (constituer l'échantillon) :

   On considère un modèle linéaire du type $C = C_0 + bR$ avec C la consommation, R le revenu (toutes deux des variables), C_0 et b des paramètres à déterminer, représentant la consommation incompressible et la propension marginale à consommer.

   On choisit le type de données que l'on va utiliser pour estimer C_0 et b : par exemple, on dispose des données pour la consommation et le revenu pour un grand nombre de ménages à un instant donné C_i, R_i $i = 1, \ldots, n$.

c) Examiner dans quelle mesure la théorie est validée ou non par les résultats obtenus, et si elle n'est pas rejetée, évaluer les paramètres du modèle :

   Est-ce que la relation entre C et R est acceptable? Est-elle linéaire? Est-ce que R influence significativement C ou pas, c'est-à-dire est-ce que $b = 0$? (on fera des tests). Quelles sont les valeurs acceptables pour b et C_0 et la confiance à accorder à ces valeurs?

L'économètre va alors recueillir des données : par exemple les valeurs de la consommation et du revenu de n ménages. Chaque ménage i ($i = 1, \ldots, n$) étant ainsi caractérisé par une valeur pour sa consommation C_i et pour son revenu R_i, on peut représenter les ménages dans le plan (R_i, C_i). On obtient un nuage de points (plus ou moins aligné selon la linéarité de la relation pour les ménages étudiés).

S'il existait une relation certaine entre consommation et revenu des ménages, et que cette relation était précisément la même pour tout le monde, on aurait pour chaque individu :

$$C_i = C_0 + bR_i$$

Dans ce cas, toutes les observations appartiendraient à la même droite (dans le plan (R_i, C_i). Il suffirait alors de connaître les observations pour 2 ménages seulement pour trouver les valeurs des paramètres C_0 et b. Ce cadre de figure ne se rencontre jamais car la réalité est plus complexe. En effet, aucun ménage ou presque ne vérifie exactement la fonction de consommation keynésienne : Certains ménages sont plus dépensiers (une plus grande préférence pour le présent peut les pousser à consommer une proportion plus importante de leur revenu que la moyenne des ménages); D'autres ménages sont très exposés au risque de chômage par exemple (ils cherchent à consommer moins pour économiser, pour se constituer une épargne de précaution); On peut penser aussi à quelqu'un qui détiendrait un portefeuille de titres et qui anticiperait une hausse des cours de la Bourse, ou un déménagement prévu, un mariage à fêter, la préparation d'un long voyage, l'existence de revenus exceptionnels non anticipés, etc.

En fait, il existe une infinité de facteurs explicatifs de la consommation qu'il est impossible d'intégrer dans le modèle. Le modèle est une simplification de la réalité. Ainsi, la fonction de consommation considérée est affectée d'une incertitude : chaque ménage est un cas particulier dont le comportement de consommation s'écarte du modèle théorique.

Pour gérer cette incertitude, on utilise une approche probabiliste en introduisant une variable aléatoire appelée perturbation aléatoire : elle est appelée ainsi car elle perturbe une relation stable (qui est donnée par la théorie économique de manière complètement déterministe). Nous formulerons des hypothèses sur cette variable aléatoire et plus précisément sur son espérance, sa variance, sa loi (hypothèses qu'il faudra vérifier a posteriori).

Le modèle économétrique que l'on considérera est alors le suivant :

$$C_i = C_0 + bR_i + u_i$$

On observe C_i et R_i pour chaque ménage mais on ne connaît pas C_0 et b (les paramètres à estimer). Les réalisations des perturbations u_i sont inobservées. Elles résument notre incertitude : elles incorporent l'ensemble des facteurs explicatifs de la consommation qui n'ont pas été pris en compte dans le modèle.

L'essentiel est de construire une approximation acceptable de la relation économique étudiée. Nous verrons dans quels cas l'approximation constituée par le modèle est acceptable. Notons que ce qui nous intéresse, c'est de mesurer correctement l'influence des variables figurant dans le modèle sur le phénomène étudié.

Par exemple, notre modèle est réducteur dans le sens où il n'intègre pas explicitement une influence du risque de chômage sur la consommation. Cette approximation sera jugée comme acceptable si elle n'introduit pas d'erreur dans l'évaluation du paramètre b.

Bien entendu, l'économètre peut s'intéresser à d'autres modèles, par exemple, l'estimation d'une fonction de production Cobb Douglass, où la production Y (variable endogène) dépend des facteurs de production, le capital K et le travail L, ainsi que le temps t :

$$Y = AL^{\alpha}K^{1-\alpha}B^t$$

On remarque que ce modèle n'est pas linéaire tel que, mais on peut le rendre linéaire (dans les variables) si on prend le logarithme de cette équation. En effet, on obtient :

$$y = a + \alpha k + (1-\alpha)l + tb$$

où on note en minuscule le logarithme des variables ($k = \ln K, l = \ln L$) ou des paramètres ($a = \ln A, b = \ln B$). Le modèle économétrique à estimer est dit modèle de régression multiple, car il comporte plusieurs variables explicatives (capital, emploi, temps) au phénomène étudié (production de l'entreprise). Si nous disposons d'observations dans le temps pour les variables, le modèle est donné par :

$$y_t = a + \alpha k_t + (1-\alpha)l_t + tb + u_t$$

L'économètre peut aussi s'intéresser à l'estimation d'une équation de salaire, dans laquelle on cherche à mesurer l'effet d'une année d'expérience supplémentaire

sur le salaire perçu, ou l'effet de l'âge sur le salaire. Le modèle s'écrit par exemple :

$$w = a + bExp + cEtude + dAge + eAge^2 + fFemme$$

avec w le logarithme du salaire perçu par les individus, Exp le nombre d'années d'expérience, $Etude$ le nombre d'années d'études, Age l'âge de la personne (Age^2 étant introduit pour rendre compte de l'effet non linéaire de l'âge sur le salaire : le salaire augmente avec l'âge mais augmente de moins en moins vite) et enfin $Femme$ étant une variable qui vaut 1 si la personne est une femme et 0 sinon. Supposons que l'on dispose de données individuelles, le modèle économétrique (de régression multiple) à estimer sera alors :

$$w_i = a + bExp_i + cEtude_i + dAge_i + eAge_i^2 + fFemme_i + u_i$$

Quelles données retenir?

Nous pouvons disposer en général de plusieurs types de données :

- Les données en coupe instantanée, ou coupe transversale, correspondent à l'observation à un moment donné de différents individus (ménages, entreprises, secteurs, pays, etc).
- Les séries chronologiques ou séries temporelles correspondent à des observations de variables (souvent des agrégats de variables) dans le temps (à intervalle régulier).
- Les données individuelles-temporelles, ou données de panel, qui combinent l'aspect individuel et l'aspect temporel.

Le modèle de régression étudié dans ce cours peut être utilisé pour étudier une variable en coupe instantanée ou une série temporelle (les données de panel nécessitent quelques traitements distincts).

Les variables peuvent être mesurées :
→ au niveau individuel (plutôt pour étudier des comportements microéconomiques, mais toutes les données ne sont pas disponibles comme cela, par exemple la durée du travail dans chaque entreprise).
→ au niveau agrégé (soit nationale, sectorielle ou régionale). Ce sont souvent des séries temporelles. C'est souvent le niveau pertinent pour estimer des modèles macroéconomiques permettant d'effectuer des simulations de politiques économiques. Cependant, l'agrégation engendre beaucoup de pertes d'information : les estimations sont moins précises et on s'expose à des biais d'agrégation dès que les comportements des agents microéconomiques ne sont pas homogènes.

Les variables peuvent être :
→ quantitatives (consommation, revenu, etc)
→ qualitatives (être homme ou femme, être ou non diplômé de l'enseignement supérieur, habiter tel département).

Dans ce cours, nous étudions le modèle de régression linéaire, qui explique le phénomène y (toutes variables quantitatives que l'on cherche à caractériser : la consommation des ménages, les ventes d'une entreprise, le salaire des gens, etc) par un modèle linéaire en fonction de K variables explicatives (quantitatives ou qualitatives) $x_1, x_2, \ldots, x_K$:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \quad i = 1, \ldots, n$$

On dit que y est la **variable à expliquer** et que les x_j, $j = 1, \ldots, K$ sont les K **variables explicatives**. Ces variables sont observées sur $i = 1, \ldots, n$ observations. ε est une perturbation aléatoire.

Le modèle est qualifié de linéaire car y est une fonction linéaire des paramètres $(\beta_0, \beta_1, \ldots, \beta_K)$. L'hypothèse de linéarité nécessite que le terme d'erreur soit introduit de manière additive et que la forme fonctionnelle soit linéaire dans les paramètres. Cette hypothèse n'est pas si restrictive puisqu'un grand nombre de formes fonctionnelles sont linéaires ou quasi-linéaires (linéaire après transformation), par exemple $Y = AL^\alpha K^\beta$ est linéaire après transformation logarithmique : $y = a + \alpha l + \beta k$ avec $y = \ln Y, l = \ln L, k = \ln K, a = \ln A$.

Autres exemples de modèles quasi-linéaires : $y = Ax^\alpha e^\varepsilon$; $y = \alpha + \beta \cos(x) + \varepsilon$; $y = \alpha + \beta/x + \varepsilon$; $y = \alpha + \beta \ln x + \varepsilon$.

La transformation la plus utilisée reste la transformation logarithmique. Celle-ci permet, outre le fait de linéariser certains modèles, d'interpréter les paramètres du modèle comme des élasticités, puisque, si le modèle explique $\ln y$ en fonction de $\ln x_k$, alors le coefficient associé à $\ln x_k$ est l'élasticité de y par rapport à x_k :

$$\beta_k = \frac{d \ln y}{d \ln x_k} = \frac{dy/y}{dx_k/x_k} = \left(\frac{\partial y}{\partial x_k}\right)\left(\frac{x_k}{y}\right)$$

alors que l'élasticité dans le modèle linéaire sans logarithme, où y est expliqué par x_k, n'est pas constante :

$$\left(\frac{\partial y}{\partial x_k}\right)\left(\frac{x_k}{y}\right) = \frac{\beta_k x_k}{y}$$

L'intérêt d'un modèle de régression, relativement au calcul des simples corrélations entre variables, est qu'il permet d'étudier la proposition chère aux économistes : "toute chose égale par ailleurs". En effet, une corrélation entre deux variables peut être élevée, parce que ces deux variables sont toutes deux influencées par une troisième. Le coefficient de corrélation ne nous permet pas de mesurer la relation entre ces deux variables uniquement, indépendamment de l'influence de la troisième variable. En revanche, le modèle de régression permet de déterminer l'effet d'une variable sur une autre, les autres variables explicatives étant supposées inchangées (coefficient de régression dans une régression comportant plusieurs variables explicatives). Par exemple, un modèle de régression expliquant le salaire des individus en fonction par exemple du niveau d'étude et de l'âge, va nous permettre d'étudier l'influence du niveau d'étude sur le salaire des gens, à âge donné. On sait en effet que l'âge a un effet sur le salaire et sur le niveau d'étude. Si on souhaite comparer le salaire de deux individus ayant le même âge pour des niveaux d'étude différents, il

se peut que, dans l'échantillon qu'on observe, il n'y ait pas deux personnes de même âge et de niveau d'étude différent. Ainsi, la simple observation de nos données ne nous permet pas de répondre à cette question. Le modèle de régression multiple nous permettra de répondre à cette question en estimant tout simplement le coefficient du niveau d'étude dans la régression du salaire sur le niveau d'étude et sur l'âge.

Dans ce premier chapitre, nous nous restreignons au modèle de régression simple, dans lequel il n'y a qu'une seule variable explicative au phénomène y à expliquer. Ainsi, pour chaque observation i, on peut écrire :

$$y_i = a + bx_i + u_i$$

Nous allons tout d'abord énoncer les hypothèses que l'on adopte dans le modèle de régression, puis nous présenterons les méthodes de régression (la méthode du Maximum de vraisemblance ainsi que la méthode des MCO). Nous présenterons alors l'analyse de la variance du modèle puis les tests d'hypothèses.

1 Les hypothèses du modèle

Afin d'estimer les paramètres a et b, on doit se donner des hypothèses sur le terme d'erreur, ainsi que sur le modèle. De ces hypothèses découlera une méthode d'estimation.

Les hypothèses que l'on pose sont les suivantes :

1. $E(u_i) = 0 \ \forall i = 1, \ldots, n$: les erreurs sont de moyenne nulle, ou autrement dit, on ne se trompe pas en moyenne
2. $var(u_i) = \sigma^2 \ \forall i = 1, \ldots, n$: la variance est la même pour tout i (hypothèse d'homoscédasticité)
3. $cov(u_i, u_j) = 0 \ i \neq j$: indépendance des erreurs, autrement dit, le fait de faire une erreur pour une observation i n'implique rien sur l'erreur de l'observation j
4. x_i est une variable certaine (non aléatoire), elle est parfaitement connue[4]
5. $u_i \sim \mathcal{N}(0, \sigma^2)$: hypothèse de normalité des erreurs

Les hypothèses 2 et 3 ne sont pas forcément vérifiées, tout dépend de la nature de la variable y et surtout de la nature des observations (données individuelles ou temporelles). Une fois les paramètre a et b estimés, les résidus u_i seront estimés et on vérifiera a posteriori si les hypothèses sur les erreurs sont vérifiées ou non. Si elles ne le sont pas, on adaptera la méthode d'estimation (*Cf.* chapitre 4).

[4]Très souvent, il est difficile de supposer que la variable x est certaine, autrement dit qu'elle n'est pas aléatoire, alors que la variable y est aléatoire. Cependant, on montre que les résultats obtenus sous l'hypothèse que x est une variable certaine sont maintenus en supposant que x est aléatoire mais que $cov(u_i, x_i) = 0$, c'est-à-dire que la distribution des u ne dépend pas des valeurs de x.

L'hypothèse de normalité est nécessaire pour mener une inférence, c'est–à–dire des tests sur les paramètres estimés. On devra vérifier, sur la base des résidus estimés, que cette hypothèse est acceptable.

Conséquences des hypothèses :
Puisque les erreurs sont de moyenne nulle et que x_i est une variable non aléatoire ($E(x_i) = x_i$), alors $E(y_i) = a+bx_i$. Ainsi, le modèle théorique est vérifié en moyenne, même si pour chaque individu, il y a une erreur.

Puisque les erreurs sont de même variance, alors $var(y_i) = var(u_i) = \sigma^2 \quad \forall i = 1, \ldots, n$. Ainsi, les données sont telles que la variance est constante (on ne peut pas avoir des observations de variance très différente, par exemple des petites et des grandes entreprises dans le même échantillon, il faudra dans ce cas adapter la méthode d'estimation. Voir chapitre 4).

Puisque les erreurs ne sont pas corrélées, $cov(y_i, y_j) = cov(u_i, u_j) = 0$ pour $i \neq j$. Ainsi, si les observations sont temporelles, cela signifie que la valeur prise par y à une date ne dépend pas de la valeur prise par y à la date d'avant par exemple. Ce qui est très peu probable avec des données temporelles, et il faudra donc adapter la méthode d'estimation. Voir chapitre 4.

On peut déduire de ces différentes hypothèse que chaque y_i est supposé suivre une loi normale de moyenne $a + bx_i$ et de variance σ^2.

A partir de ce modèle et des hypothèses formulées, on peut développer la méthode d'estimation. Nous présenterons deux méthodes d'estimation : la méthode du maximum de vraisemblance et la méthode des MCO (qui donnent exactement les mêmes résultats dans ce cadre).

2 Méthodes d'estimation

Le but de l'économètre est de déterminer la valeur des paramètres inconnus (a et b ici) le mieux possible. Il s'agit de trouver l'estimateur de ces paramètres qui soit sans biais et le plus précis possible.

2.1 La méthode des MCO

Le problème est de déterminer les paramètres estimés ($\hat{a}$ et $\hat{b}$) de telle sorte que l'ajustement, $\hat{y}_i = \hat{a} + \hat{b}x_i$ soit aussi proche que possible de l'observation y_i, ou autrement dit, que l'erreur (estimée) $\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{a} - \hat{b}x_i$ soit aussi proche que possible de 0 et cela pour chaque i. La mesure de la proximité que l'on retient constitue le critère d'ajustement. On retient le critère des moindres carrés ordinaires, c'est–à–dire qu'on retient les valeurs $\hat{a}$ et $\hat{b}$ qui minimisent la somme des carrés des résidus :

$$(\hat{a}, \hat{b}) = \arg\min_{a,b} \sum_{i=1}^{n} u_i^2 = \arg\min_{a,b} \sum (y_i - a - bx_i)^2$$

Les conditions du premier ordre, appelées équations normales, sont les suivantes :

$$\begin{aligned} \sum_{i=1}^{n}(y_i - a - bx_i) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - a - bx_i) &= 0 \end{aligned}$$

Ainsi, pour $\hat{a}$ et $\hat{b}$ solutions de ce système, la première équation est $\sum_i \hat{u}_i = 0$, elle signifie que les paramètres doivent être tels que les résidus estimés sont de moyenne nulle. La seconde condition s'écrit quant à elle $\sum_i x_i\hat{u}_i = 0$, elle implique que x_i et les résidus estimés sont orthogonaux ($cov(x_i, \hat{u}_i) = 0$).

Après quelques réarrangements, on obtient :

$$\begin{aligned} \hat{a} &= \bar{y} - \hat{b}\bar{x} \\ \hat{b} &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{cov(y_i, x_i)}{var(x_i)} \end{aligned}$$

avec $\bar{x} = \frac{\sum_i x_i}{n}$ et $\bar{y} = \frac{\sum_i y_i}{n}$.

$\hat{a}$ est tel que la droite passe par le point moyen du nuage $(\bar{x}, \bar{y})$ et $\hat{b}$ est donné par le rapport de la covariance entre x et y et la variance de x.

2.2 La méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est une méthode d'estimation très générale applicable dans tous les modèles de régression. On montre que cette méthode donne toujours des estimateurs sans biais et efficaces (c'est-à-dire dont la variance est minimale). Dans le cas du modèle linéaire, cette méthode donne les mêmes estimateurs que la méthode des MCO.

On a vu que les hypothèses faites sur les erreurs u_i impliquent que les y_i suivent une loi Normale de moyenne $a + bx_i$ et de variance σ^2, et qu'ils sont indépendants entre eux ($cov(y_i, y_j) = 0\ i \neq j$).

Ainsi, sachant que $y_i \sim \mathcal{N}(a + bx_i, \sigma^2)$, alors la densité de probabilité de y_i est :

$$f(y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - a - bx_i)^2\right]$$

Sachant que les y_i sont indépendants, la densité jointe des observations $(y_1, \ldots, y_n)$ est donnée par le produit des densités individuelles :

$$f(y_1, \ldots, y_n) = \Pi_{i=1}^{n} f(y_i) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right]$$

Cette fonction (fonction des paramètres a, b, σ^2) est appelée la fonction de vraisemblance et est souvent notée $L(a, b, \sigma^2)$.

La méthode d'estimation du maximum de vraisemblance suggère que l'on choisisse les paramètres a, b, σ^2 qui maximisent la fonction de vraisemblance $L(a, b, \sigma^2)$. On préfère maximiser le logarithme de cette fonction, noté $\log L(a, b, \sigma^2)$, pour des

raisons pratiques, les résultats étant les mêmes puisque les deux fonctions atteignent leur maximum au même point.

La log-vraisemblance est donnée par :

$$\log L(a, b, \sigma^2) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = c - \frac{n}{2}\log\sigma^2 - \frac{\sum u_i^2}{2\sigma^2}$$

où $c = -\frac{n}{2}\log 2\pi$ ne dépend pas de a, b, σ^2.

Afin de maximiser cette fonction, nous écrivons les conditions du premier ordre :

$$\begin{aligned}
\frac{\partial \log L}{\partial a} &= -\frac{1}{2\sigma^2}\frac{\partial \sum u_i^2}{\partial a} = 0 \\
\frac{\partial \log L}{\partial b} &= -\frac{1}{2\sigma^2}\frac{\partial \sum u_i^2}{\partial b} = 0 \\
\frac{\partial \log L}{\partial \sigma^2} &= -\frac{n}{2\sigma^2} + \frac{\sum u_i^2}{2\sigma^4} = 0
\end{aligned}$$

Ainsi, on remarque que les deux premières conditions sont exactement équivalentes aux conditions qui permettent de minimiser la somme des carrés des erreurs. Ainsi, les valeurs de a et b qui maximisent la vraisemblance sont exactement les mêmes que celles qui satisfont les MCO.

Si on remplace a et b par leur estimateur $\hat{a}$ et $\hat{b}$, alors on obtient $\hat{u}_i$ et la dernière condition nous donne :

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

Une fois que l'on a les estimations des paramètres, on peut calculer la log-vraisemblance estimée :

$$\max \log L = c - \frac{n}{2}\log\frac{SCR}{n} - \frac{n}{2}$$

où $SCR = \sum_i \hat{u}_i$, la somme des carrés des résidus estimés. Cette expression sera très utile pour faire des tests sur les paramètres (tests du ratio de vraisemblance, du multiplicateur de Lagrande, de Wald).

2.3 Quelques remarques sur les estimateurs

On rappelle que les estimateurs MCO de a et b sont donnés par :

$$\begin{aligned}
\hat{a} &= \bar{y} - \hat{b}\bar{x} \\
\hat{b} &= \frac{\sum(y_i - \bar{y})(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} = \frac{\sum y_i x_i - n\bar{y}\bar{x}}{\sum x_i^2 - n\bar{x}^2} = \frac{cov(y_i, x_i)}{var(x_i)}
\end{aligned}$$

• L'expression des estimateurs montre que si on multiplie toutes les observations y_i et les x_i par un même coefficient, la valeur de $\hat{a}$ et $\hat{b}$ n'est pas modifiée. Mais de façon générale, les valeurs estimées dépendent des unités de mesure pour les variables. En conséquence, un coefficient b par exemple élevé ne signifie pas nécessairement que la variable explicative x a une forte influence sur y. Dans la pratique, il convient donc d'interpréter prudemment les résultats obtenus, en examinant non seulement les valeurs des coefficients, mais aussi le sens économique des

variables du modèle et leur unité de mesure (ainsi que la précision de l'estimation comme on le verra ensuite).

• On peut modifier l'écriture des équations normales afin de faire intervenir les deux variables explicatives, x et la constante :

$$\begin{aligned}\sum_{i=1}^{n} 1(y_i - a - bx_i) &= 0 \\ \sum_{i=1}^{n} x_i(y_i - a - bx_i) &= 0\end{aligned}$$

Ainsi, les résidus estimés sont de moyenne nulle ($\sum_i \hat{u}_i = 0$) car la constante est dans les variables explicatives. Dans ce cas, puisque l'on a $y_i = \hat{y}_i + \hat{u}_i$, alors $\bar{y} = \bar{\hat{y}}$, la moyenne de la série ajustée est égale à la valeur de la moyenne des observations.

Ainsi, le principe de la méthode des MCO consiste à décomposer y_i en deux éléments, et plus particulièrement à trouver $\hat{y}_i$ dans l'espace des variables explicatives (ici, l'espace engendré par la constante et x_i), telle que la distance entre y_i et $\hat{y}_i$ soit la plus petite possible. Autrement dit, puisque $y_i = \hat{y}_i + \hat{u}_i$, la méthode des MCO consiste à minimiser l'erreur faite $\hat{u}_i$ quand on choisit $\hat{y}_i$ pour y_i. Or, en définissant la distance entre y_i et $\hat{y}_i$ par la norme euclidienne, on sait (d'après le théorème du plus court chemin) que cette distance minimale est obtenue en définissant $\hat{y}_i$ comme la projection orthogonale de y_i sur l'espace engendré par les variables explicatives.

• On peut remarquer, d'après la définition de $\hat{b}$, que l'on obtiendrait les mêmes estimateurs en projetant y_i sur x_i et une constante ou en projetant $y_i - \bar{y}$ sur $x_i - \bar{x}$. Il s'agit du théorème de Frisch-Waugh. Ce théorème nous dit que le coefficient estimé pour b dans la régression $y_i = a + bx_i + u_i$ est le coefficient de la régression du résidu de y_i sur une constante sur le résidu de x_i sur une constante. Or régresser une variable sur une constante donne comme coefficient estimé la moyenne de cette variable. Ainsi, il est équivalent de mettre une constante dans une régression ou de ne pas en mettre mais de travailler sur des variables centrées.

• Puisque les estimateurs $\hat{a}$ et $\hat{b}$ sont des fonctions linéaires des observations y_i, et que y_i est supposé suivre une loi normale, alors les estimateurs $\hat{a}$ et $\hat{b}$ suivent une loi normale.

On peut montrer très facilement qu'ils sont sans biais, c'est–à–dire que leur espérance est égale à la vraie valeur du paramètre qu'ils estiment ($E(\hat{a}) = a$ et $E(\hat{b}) = b$).

Leur variance est donnée par :

$$var(\hat{b}) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \qquad var(\hat{a}) = \sigma^2 \frac{\sum_i x_i^2}{n \sum_i (x_i - \bar{x})^2} \qquad cov(\hat{a}, \hat{b}) = \frac{-\bar{x}\sigma^2}{\sum_i (x_i - \bar{x})^2}$$

Ainsi

$$\hat{a} \sim \mathcal{N}(a, var(\hat{a})) \Longrightarrow \frac{\hat{a} - a}{\sqrt{var(\hat{a})}} \sim \mathcal{N}(0, 1)$$

$$\hat{b} \sim \mathcal{N}(b, var(\hat{b})) \Longrightarrow \frac{\hat{b} - b}{\sqrt{var(\hat{b})}} \sim \mathcal{N}(0, 1)$$

On remarque que les variances des coefficients estimés varient directement avec $var(u_i) = \sigma^2$. Ainsi, on pourra obtenir des estimateurs d'autant plus efficaces que cette variance est faible. De plus, les variances des estimateurs varient inversement avec $var(x_i)$. Ainsi, plus les x_i sont dispersés et plus on peut avoir des estimateurs précis. En effet, si les variables x_i varient très peu, on aura du mal à obtenir de bons estimateurs de la droite de régression puisque toutes les observations seront concentrées.

On a vu que la variance des estimateurs dépend de σ^2 qui est inconnue. Il faut alors trouver un estimateur de σ^2. On montre que :

$$\frac{1}{n-2} \sum_i \hat{u}_i^2$$

est un estimateur sans biais de la variance σ^2. On a $n-2$ et non pas n pour définir la variance estimée des résidus car pour estimer les résidus $\hat{u}_i$, on a dû estimer 2 paramètres, à savoir $\hat{a}$ et $\hat{b}$.

Ainsi, la variance estimée par maximum de vraisemblance devient valide quand le nombre d'observations n est élevé.

Puisque σ^2 doit être estimée, alors les paramètres estimés suivent une loi de student et non plus une loi normale (voir chapitre 1) :

$$\frac{\hat{a} - a}{\sqrt{v\hat{a}r(\hat{a})}} \sim \mathcal{S}t_{n-2} \qquad \frac{\hat{b} - b}{\sqrt{v\hat{a}r(\hat{b})}} \sim \mathcal{S}t_{n-2}$$

On se servira de ce résultat pour faire des tests sur les paramètres a et b.

3 Décomposition de la variance et qualité de la régression

L'idée est de savoir quelle est la part de la variation de y qui est expliquée par les variations de x. Pour cela, décomposons la variance de y_i. Puisque $y_i = \hat{y}_i + \hat{u}_i$, en retranchant $\bar{y}$ (qui est égal à $\bar{\hat{y}}$) des 2 cotés, on obtient:

$$y_i - \bar{y} = \hat{y}_i - \bar{\hat{y}} + \hat{u}_i$$

Or, comme $\hat{y}_i$ et $\hat{u}_i$ sont orthogonaux, alors la variation totale à expliquer, ou Somme des Carrés Totale $SCT = \sum (y_i - \bar{y})^2$, peut se décomposer en Somme des Carrés Expliquée (par le modèle, ou plus précisément par la variable x) $SCE = \sum (\hat{y}_i - \bar{y})^2$ et en Somme des Carrés des Résidus (partie que le modèle n'explique pas) $SCR = \sum \hat{u}_i^2$. On a alors l'équation d'analyse de la variance suivante :

$$SCT = SCE + SCR$$

Remarque : tout ceci repose sur le fait que les résidus estimés sont centrés, et donc qu'il y a une constante dans le modèle et que la méthode d'estimation est les MCO.

On définit alors le coefficient de détermination, qui mesure la part de la variance expliquée par le modèle dans la variance totale :

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

Il est compris entre 0 et 1. Plus il est proche de 1 et plus la régression permet d'expliquer une grande partie de la variance totale de la variable à expliquer.

Remarque : on peut montrer que ce coefficient de détermination R^2 est égal au coefficient de corrélation entre y et x, $r_{xy} = cov(y,x)/\sigma_x\sigma_y$, dans le cadre du modèle de régression simple (une seule variable explicative) :

$$R^2 = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} = \frac{\hat{b}^2\sum_i(x_i - \bar{x})^2}{\sum_i(y_i - \bar{y})^2} = \frac{Cov(x_i,y_i)^2}{var(x_i)var(y_i)} = r_{xy}^2$$

Attention, le jugement sur la valeur de R^2 est très subjectif. Bien que ce coefficient soit très facile à comprendre, il faut se garder d'y attacher trop d'importance car il est loin de fournir un critère suffisant pour juger de la qualité d'une régression.

• En effet, la valeur de ce critère est aisément manipulable, par exemple elle dépend de la forme sous laquelle on a introduit les variables (en log ou en taux de croissance). On peut donc facilement l'améliorer ou le détériorer en modifiant la forme fonctionnelle dans laquelle la variable y est spécifiée (niveau, log, ratio, taux de croissance).

Exemple de limite de R^2 : Si au lieu d'estimer $y_i = a + bx_i + u_i$, on estime $z_i = \alpha + \beta x_i + w_i$ avec $z_i = y_i - x_i$ et $\beta = b - 1$, on obtiendra un R^2 supérieur avec le second modèle si $b < 1/2$!

• Si le modèle ne comporte pas de terme constant, l'équation d'analyse de la variance n'est plus vérifiée en général.

• Enfin, le coefficient de détermination augmente mécaniquement quand on ajoute une variable explicative, même si celle-ci n'a pas beaucoup de rapport avec y.[5]

[5]Afin de montrer cela, nous considérons un modèle de régression multiple, *Cf.* chapitre 3. Supposons que l'on ait $y = X\beta + u_1$ et $\hat{y}_1$ son ajustement. On ajoute une variable explicative et on a $y = X\beta + \delta w + u_2$ et on note $\hat{y}_2$ son ajustement.

Ainsi $\hat{u}_1 = y - \hat{y}_1 = y - \hat{y}_2 + \hat{y}_2 - \hat{y}_1 = \hat{u}_2 + \hat{y}_2 - \hat{y}_1$. Or $\hat{u}_2$ est orthogonal à l'espace des X (car il est orthogonal à l'espace des (X,w)), donc il est orthogonal à $\hat{y}_2 - \hat{y}_1$. On peut donc écrire $||\hat{u}_1||^2 = ||\hat{u}_2||^2 + ||\hat{y}_2 - \hat{y}_1||^2$, ce qui implique $||\hat{u}_1||^2 > ||\hat{u}_2||^2$ et donc le R^2 du premier modèle est plus petit que le R^2 du second modèle, de manière purement mécanique, indépendamment de la pertinence de la variable w (dès que son coefficient est non nul).

Ainsi, on préfère un coefficient de détermination ajusté par le nombre de variables explicatives, noté $\bar{R}^2$. *Cf.* chapitre suivant.

4 Les tests d'hypothèse

Dans le cadre de ce chapitre, nous présentons l'idée des tests d'hypothèses, ainsi que l'application au test de student. Dans le chapitre suivant de régression multiple, nous présenterons les autres tests.

4.1 Généralités sur les tests

On va chercher à tester la validité d'hypothèses concernant les (vrais) paramètres du modèle.

On se donne une **hypothèse nulle**, notée H_0 que l'on teste, et une **hypothèse alternative**, notée H_1.

La procédure est basée sur la construction d'une **statistique**, calculée sur l'échantillon aléatoire afin de décider, avec un niveau de confiance raisonnable, si on peut supposer que les données de l'échantillon suivent l'hypothèse nulle (c'est-à-dire si on peut supposer que l'hypothèse nulle est acceptable).

La statistique retenue dépend de l'hypothèse que l'on teste (une statistique de Student quand on ne teste qu'une contrainte, une statistique de Fisher quand on teste plusieurs contraintes).

On se donne ensuite une **règle de décision** pour savoir si l'hypothèse nulle doit être acceptée ou rejetée. Plus précisément, on se donne une zone de rejet de l'hypothèse nulle.

Par exemple, si l'on veut tester si un paramètre est égal à une certaine valeur, la règle de décision sera qu'on rejette H_0 si le paramètre estimé est trop loin de la valeur testée. Il reste à définir la notion de "trop loin".

Puisque l'échantillon est aléatoire, et que la statistique est calculée sur la base de l'échantillon, alors la statistique de test est elle aussi aléatoire. Ainsi, la même procédure de test peut conduire à des conclusions différentes sur des échantillons différents.

En fait, il y a deux manières pour que la procédure de test fasse une erreur :

- On parle d'**erreur de première espèce** quand la procédure conduit à rejeter H_0 quand H_0 est vraie
- On parle d'**erreur de seconde espèce** quand la procédure conduit à ne pas rejeter H_0 alors que H_0 est fausse

Il y a une probabilité non nulle pour que l'estimation du paramètre soit assez éloignée de la valeur testée, même si l'hypothèse nulle est vraie. Cette situation conduit à une erreur de première espèce. La probabilité d'une erreur de première espèce est la **taille du test**, notée α. Elle est aussi appelée seuil d'erreur (ou niveau de significativité, *significance level*).

Ce seuil est choisi par l'économètre, il peut être changé en modifiant la règle de décision. Ainsi, on peut réduire l'erreur de première espèce en rendant la région de rejet de H_0 très petite, mais on sera alors conduit à ne jamais rejeter H_0, même si elle est fausse, c'est-à-dire à augmenter l'erreur de seconde espèce! Il y a un arbitrage entre les deux, et ce qu'on veut, c'est que les deux soient suffisamment petites.

L'erreur de seconde espèce dépend bien sûr de l'hypothèse alternative, et donc de la valeur du paramètre. Ainsi, ce que l'on veut, c'est que sachant le seuil d'erreur

α qu'on se donne et la procédure de test adoptée, le risque de seconde espèce ne soit pas trop élevé.

Ainsi, on retient généralement $\alpha = 1\% = 0.01, 5\% = 0.05, 10\% = 0.1$.

On appelle **puissance** du test la probabilité qu'on rejette l'hypothèse nulle sachant qu'elle est fausse. On dit qu'un test est plus puissant s'il a une plus grande puissance que tout test pour une même taille.

4.2 Tests de student

Ce test est utilisé pour tester une seule contrainte sur les paramètres, par exemple la nullité d'un paramètre (permettant de tester la significativité d'une variable) ou l'égalité d'un paramètre à une certaine valeur donnée, ou une fonction de paramètres.

Ainsi, si on teste l'égalité d'un paramètre, par exemple b, à une valeur donnée b_0, les hypothèses nulle et alternative sont données par :

$$\begin{aligned} H_0 &: b = b_0 \\ H_1 &: b \neq b_0 \end{aligned}$$

Sachant l'estimation obtenue pour le paramètre b, soit $\hat{b}$, et sachant d'autre part que cet estimateur suit une loi de Student, on a l'intervalle de confiance suivant pour le vrai paramètre b :

$$\hat{b} - St_{n-2}\hat{\sigma}_{\hat{b}} < b < \hat{b} + St_{n-2}\hat{\sigma}_{\hat{b}}$$

où St_{n-2} est la valeur critique lue dans une table de loi de Student à $n-2$ degrés de liberté au seuil de α, c'est–à–dire la valeur telle qu'une variable aléatoire de Student à $n-2$ degrés de liberté ait $\alpha\%$ d'être supérieure à cette valeur. Si $\alpha = 5\%$, on a 95% de chance d'obtenir une valeur pour b comprise dans cet intervalle de confiance.

L'idée du test est donc de rejeter H_0 si b_0 excède la limite supérieure ou est inférieur à la limite inférieure. Autrement dit, on rejette H_0 si :

$$\left| \frac{\hat{b} - b_0}{\hat{\sigma}_{\hat{b}}} \right| > St_{n-2}$$

Ainsi, pour tester H_0, on utilise la statistique de student, définie par :

$$t = \frac{\hat{b} - b_0}{\hat{\sigma}_{\hat{b}}}$$

Cette statistique suit, sous H_0, une loi de Student à $n-2$ degrés de liberté.

La région de rejet de H_0 au seuil de α est alors :

$$R_\alpha = \{|t| > St_{n-2}\}$$

Si la valeur calculée pour la statistique t est dans la région de rejet, alors on dit qu'au seuil de α, on peut rejeter H_0 (avec un risque d'erreur de α). Si la statistique calculée n'est pas dans la région de rejet, alors on dit qu'on ne peut pas rejeter H_0 avec le risque d'erreur de α.

Les logiciels permettent de conclure sur le rejet ou non de l'hypothèse nulle sans avoir à regarder les valeurs critiques dans une table statistique. En effet, ils donnent la valeur de la statistique de test ainsi que la probabilité du test, appelée *p-value*. La **p-value** est le niveau d'erreur exact associé à un résultat de test particulier. Elle mesure la vraisemblance d'une erreur de première espèce, soit la probabilité de rejeter à tord l'hypothèse nulle. Plus la p-value est élevée, plus il est probable que l'on se trompe en rejetant l'hypothèse nulle, et il est alors préférable de ne pas rejeter H_0. Plus la p-value est faible, plus on est rassuré pour la rejeter car moins on a de chance de faire une erreur en la rejetant.

Une p-value égale à 0.07 indique que l'on peu rejeter l'hypothèse nulle à un seuil d'erreur de 10%, mais qu'on ne peut pas rejeter l'hypothèse nulle à un seuil d'erreur de 5%. Cela signifie que 7% de la distribution de la variable qui suit une loi de Student est en dehors de l'intervalle de plus ou moins 1.96 écart type.

Test d'absence de significativité d'une variable : Pour tester la nullité d'un coefficient, c'est-à-dire la non significativité de la variable x par exemple, on utilise la statistique de student avec $b_0 = 0$.

Ainsi, on définit

$$\begin{aligned} H_0 &: b = 0 \\ H_1 &: b \neq 0 \end{aligned}$$

La statistique de test est la statistique de student suivante :

$$t = \frac{\hat{b}}{\hat{\sigma}_{\hat{b}}}$$

Cette statistique suit, sous H_0, une loi de Student à $n-2$ degrés de liberté.

La région de rejet de H_0 au seuil de α est alors :

$$R_\alpha = \{|t| > St_{n-2}\}$$

Si la statistique de test calculée est dans la zone de rejet de H_0, ou autrement dit si la p-value est inférieure à $\alpha\%$, alors on conclut que la variable x est significative au seuil de $\alpha\%$. Sinon, c'est-à-dire si la p-value est supérieure à $\alpha\%$, alors on ne peut pas dire que la variable x est significative au seuil d'erreur de $\alpha\%$.

Si la variable x n'est pas significative, alors le modèle n'est pas pertinent. Si elle est significative, on peut interpréter son signe et sa valeur. Cependant, il faut vérifier, sur la base des résidus estimés, la validité du modèle, à savoir si les hypothèses faites sur le modèle, sont acceptables ou pas. C'est ce que nous verrons dans le chapitre suivant, après avoir généralisé les résultats du modèle de régression simple au cas de plusieurs variables explicatives.

Chapitre 3 :
Le modèle de régression multiple

On étudie dans ce chapitre le modèle de régression multiple, consistant à expliquer les variations de la variable à expliquer y à l'aide de K variables explicatives $x_1, \ldots, x_K$. Il s'agit de déterminer l'influence des variables explicatives sur y et de vérifier si les variables explicatives sont réellement "explicatives". Nous étudierons enfin les tests permettant de vérifier la validité des hypothèses effectuées.

Le modèle de régression multiple s'écrit :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \quad i = 1, \ldots, n$$

ou encore

$$y = e\beta_0 + x_1\beta_1 + \ldots + x_K\beta_K + \varepsilon \Longleftrightarrow y = X\beta + \varepsilon$$

avec

$$y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \quad e = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \quad x_j = \begin{pmatrix} x_{j1} \\ \vdots \\ x_{j1} \end{pmatrix}_{j=1,\ldots,K} \quad X = [e\ x_1 \ldots x_K] \quad \beta = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_K \end{pmatrix}$$

y est un vecteur à n éléments, X est une matrice à n lignes et $K+1$ colonnes, β est un vecteur à $K+1$ éléments, et ε est un vecteur à n éléments.

Si on ne parle que d'une observation, on notera :

$$y_i = X_i'\beta + \varepsilon_i$$

où $X_i' = [1\ x_{1i} \ldots x_{Ki}]$ le vecteur ligne comprenant l'ensemble des variables explicatives pour l'observation i.

1 Hypothèses sur le modèle de régression multiple

Avant d'étudier la méthode d'estimation des paramètres inconnus β, nous présentons les hypothèses sur le modèle, nécessaires pour que la méthode d'estimation par MCO soit applicable.

- Linéarité du modèle : le modèle est linéaire dans les paramètres β à estimer.
- X est une matrice $(n, K+1)$ de rang $K+1$, *i.e.* de plein rang colonne, c'est-à-dire que les colonnes de X sont linéairement indépendantes et qu'il y a au moins K observations. Il s'agit d'une condition d'identification.

  Si l'on avait $rang(X) < K+1$, cela signifierait qu'il existe au moins une variable explicative qui peut s'écrire comme une combinaison linéaire d'une ou des autres variables explicatives : cette variable explicative serait donc superflue, elle n'apporterait rien à l'explication de y déjà fournie par les autres variables explicatives.

Cette hypothèse signifie que le modèle est correctement écrit : il n'y a pas de redondance dans la liste des variables explicatives. Si ce n'est pas le cas, on parle de colinéarité (entre deux variables) ou de multicolinéarité (entre plus de deux variables), nous reverrons cela à la fin du chapitre.

- $\lim_{n\to\infty} \frac{X'X}{n} = V_X$ où V_X est une matrice définie positive, donc régulière.

  Ainsi, la matrice de variance-covariance empirique des variables explicatives converge vers une matrice finie et définie positive. Lorsque le modèle comprend une constante, cette hypothèse signifie que les moyennes, les variances et les covariances des variables explicatives tendent vers des limites finies quand $n \to \infty$. L'idée important formalisée ici est que les variables explicatives conservent toujours une certaine variance quand $n \to \infty$.

- $E(\varepsilon_i) = 0$ (comme dans le modèle de régression simple)

- $var(\varepsilon_i) = \sigma^2$ (comme dans le modèle de régression simple)

- $cov(\varepsilon_i, \varepsilon_j) = 0$ pour $i \neq j$ (comme dans le modèle de régression simple)

  Ces deux dernières hypothèses s'écrivent, sous forme matricielle :

  $$E(\varepsilon\varepsilon') = \sigma^2 I_n$$

  où I_n est la matrice identité d'ordre n (constituée de 1 sur la diagonale et de 0 sinon). En effet, toutes les variances sont égales à σ^2 (éléments de la diagonale de V), et les covariances sont nulles (les éléments en dehors de la diagonale).

- ε_i suit une loi Normale $\mathcal{N}(0, \sigma^2)$.

- $cov(\varepsilon_i, X_i') = 0$[6]

Ainsi, on en déduit que $E(y) = X\beta$ et $V(y) = V(\varepsilon) = \sigma^2 I$.

[6]Il est habituel de supposer que les régresseurs sont non stochastiques, c'est-à-dire que l'analyste choisit la valeur des régresseurs puis observe y; les régresseurs sont exogènes et la variable explicative est endogène. Par exemple, on cherche à expliquer le rendement d'une parcelle de terre à partir de l'engrais et de la pluie; ou alors l'effet d'une politique budgétaire ou monétaire sur le PIB.

Cette hypothèse est surtout plus pratique, elle permet d'utiliser les résultats standards de statistiques. Les variables x_j sont des constantes connues.

Une autre possibilité est de supposer que les observations sur x_j sont fixées dans des échantillons répétés, de telle manière qu'on travaille conditionnellement à l'échantilon qu'on a observé. Ainsi, on suppose que la régression et les hypothèses s'appliquent à l'échantillon qu'on a observé. On modélise y conditionnellement aux réalisations X_i' observées dans l'échantillon.

Mais on a aussi des modèles où les régresseurs ont le même statut que la variable à expliquer : dans la fonction de consommation keynésienne, pourquoi aurait-on des hypothèses différentes sur le revenu et sur la consommation?

Dans ce cas, on fait les hypothèses suivantes et la plupart des résultats que l'on énoncera plus tard sont maintenus :

• la distribution de chaque variable explicative est indépendante des vrais paramètres du modèle.

• la distribution des variables explicatives est distribuée indépendamment des erreurs du modèle.

Cette dernière hypothèse implique que X_i' et ε_i sont indépendants. Cette hypothèse est fondamentale : elle conduit à des estimateurs sans biais. Cela signifie que l'approximation constituée par le modèle est acceptable si les perturbations sont indépendantes des déterminants de y_i qui ont été retenus dans la liste de variables explicatives.

2 L'estimation par MCO

Le but de l'économètre est de déterminer la valeur du vecteur de paramètres β le mieux possible sur la base des observations $i = 1, \ldots, n$ dont il dispose. Il s'agit donc de construire un estimateur $\hat{\beta}$ du vecteur β qui soit sans biais et convergent dans le modèle $y = X\beta + \varepsilon$.

On choisit $\hat{\beta}$ de façon à ce que l'ajustement $\hat{y}_i = X_i'\hat{\beta}$ soit aussi proche que possible des observations y_i, c'est–à–dire qui minimise la somme des carrés des résidus (critère des Moindres Carrés Ordinaires) :

$$\min_{\beta} \sum \varepsilon_i^2 = \sum (y_i - X_i'\beta)^2 = \varepsilon'\varepsilon = (y - X\beta)'(y - X\beta)$$

La condition nécessaire s'écrit :

$$-2X'y + 2X'X\beta = 0$$

La solution vérifie alors les **équations normales** :

$$X'X\hat{\beta} = X'y$$

Il s'agit d'un système de $K+1$ équations à $K+1$ inconnues (les composantes de β). Ce système admet une solution unique si les $K+1$ équations sont indépendantes, c'est–à–dire si $X'X$ est régulière. Cette condition est vérifiée dès que $rg(X) = K+1$ car, dans ce cas, $rg(X'X) = K+1$.

La solution (unique) est alors donnée par :

$$\hat{\beta} = (X'X)^{-1}X'y$$

Propriétés de $\hat{\beta}$:

$\hat{\beta}$ est un estimateur sans biais :

$$E(\hat{\beta}) = E[(X'X)^{-1}X'(X\beta + \varepsilon)] = \beta + (X'X)^{-1}E(X'\varepsilon) = \beta$$

On peut calculer la variance de l'estimateur $\hat{\beta}$, afin de connaître la précision de notre estimation :

$$V(\hat{\beta}) = (X'X)^{-1}X'V(y)X(X'X)^{-1} = \sigma^2(X'X)^{-1}$$

On montre que l'estimateur sans biais de la variance des erreurs est donné par :

$$\hat{\sigma}^2 = \frac{1}{n-K-1}\sum_i \hat{\varepsilon}_i^2 = \frac{SCR}{n-K-1}$$

où $\hat{\varepsilon}_i = y_i - X_i'\hat{\beta}$.

Enfin, puisque $\hat{\beta} = (X'X)^{-1}X'y$ s'écrit en fonction de y, et que y est supposé suivre une loi normale, alors $\hat{\beta}$ suit aussi une loi normale d'espérance $E(\hat{\beta}) = \beta$ et de variance $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$. Puisque l'on doit estimer σ^2, alors $\hat{\beta}$ suit, non pas une loi normale, mais une loi de student à $n-K-1$ degrés de liberté.

Matrice de projection

Bien entendu, on retrouve l'idée que la méthode d'estimation par MCO partitionne y en 2 parties orthogonales. On peut réinterpréter l'estimation effectuée comme la détermination de la projection orthogonale de y sur l'espace engendré par les X, et la série ajustée $\hat{y}$ est alors orthogonale au résidu $\hat{\varepsilon}$.

Cette interprétation permet de retrouver directement la définition de l'estimateur des MCO. En effet, puisque $\hat{\varepsilon}$ est orthogonal à l'ensemble des X, on a :

$$X'\hat{\varepsilon} = 0 \Leftrightarrow X'(y - X\hat{\beta}) = 0 \Leftrightarrow X'y = X'X\hat{\beta}$$

A partir de la définition de l'estimateur MCO, on peut définir les matrices de projection :

$$\hat{\varepsilon} = y - X\hat{\beta} = y - X(X'X)^{-1}X'y = [I - X(X'X)^{-1}X']y = M_X y$$

La matrice M_X est fondamentale dans la théorie de la régression. Elle est symétrique ($M'_X = M_X$), idempotente ($M_X^2 = M_X$). Appliquée à y, elle donne les résidus de la régression de y sur X.

Ainsi, $M_X X = 0$: quand X est régressé sur X, le fit (ou ajustement) est parfait et le résidu est nul.

On a

$$\hat{y} = y - \hat{\varepsilon} = [I - M_X]y = P_X y$$

La matrice P_X telle que $M_X = I - P_X$ ou $P_X = X(X'X)^{-1}X'$ est aussi symétrique et idempotente. C'est une matrice de projection. Appliquée à y, elle donne la série ajustée. On a alors $P_X X = X$ et $P_X M_X = M_X P_X = 0$.

La notion de corrélation partielle

En utilisant les matrices de projection, on peut étudier les régressions partitionnées, afin de voir l'effet de l'ajout ou de l'oubli d'une variable dans la régression.

Ceci nous permet d'étendre la notion de corrélation simple afin d'étudier le lien entre la variable à expliquer et UNE variable explicative prise séparément : dans les variations de y, qu'est-ce qui est du à la variation d'une variable explicative, l'autre étant maintenue constante. Il s'agit du coefficient de corrélation partielle : c'est la corrélation entre y et X_1 une fois qu'on a retiré l'effet des autres variables à la fois sur y et sur X_1.

Supposons que le modèle s'écrive :

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon = [X_1 \;\; X_2] \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} + \varepsilon$$

X_1 et X_2 pouvant être des matrices à plusieurs colonnes (contenant plusieurs variables). Les équations normales donnent :

$$\begin{aligned} X_1'X_1\beta_1 + X_1'X_2\beta_2 &= X_1'y \\ X_2'X_1\beta_1 + X_2'X_2\beta_2 &= X_2'y \end{aligned}$$

soit

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'(y - X_2\beta_2)$$

• Si $X_1'X_2 = 0$, alors $\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'X_1$, résultat de la régression de y sur X_1. Ainsi, si les deux variables sont orthogonales, le coefficient associé à X_1 est le même que celui de la régression de y uniquement sur X_1.

• Sinon, le coefficient estimé est celui de la régression de $y - X_2\beta_2$ sur X_1. Ainsi, il correspond à l'effet de y sur X_1 une fois qu'on a retiré l'effet de X_2 sur y.

Voyons ce que donne $\hat{\beta}_2$:

$$X_2'X_1(X_1'X_1)^{-1}X_1'(y - X_2\beta_2) + X_2'X_2\beta_2 = X_2'y$$

soit

$$(X_2'X_2 - X_2'X_1(X_1'X_1)^{-1}X_1'X_2)\beta_2 = X_2'y - X_2'X_1(X_1'X_1)^{-1}X_1'y$$

Ainsi,

$$\hat{\beta}_2 = (X_2'M_1X_2)^{-1}X_2'M_1y$$

Il s'agit du coefficient de la régression de M_1y sur M_1X_2, qui sont respectivement les résidus de la régression de y (et X_2) sur X_1.

C'est donc le coefficient de la régression de y sur X_2 une fois qu'on a retiré l'effet de X_1 sur chaque variable de la régression, c'est donc l'effet net de X_2 sur y.

Ceci est le **théorème de Frisch-Waugh** : le coefficient de X_2 dans la régression de y sur X_1 et X_2 est aussi le coefficient de la régression des résidus de y sur X_1 sur les résidus de X_2 sur X_1.

Conséquences du théorème de Frisch-Waugh:

Supposons que le vrai modèle soit tel que y est expliqué par X_1 et X_2. Même si on ne s'intéresse qu'à l'influence des variables X_1 sur y, le théorème de Frisch-Waugh implique qu'il faut quand même prendre en compte la présence des X_2 dans la liste des variables explicatives de y. Sinon, l'estimateur sera biaisé, sauf si X_1 et X_2 sont orthogonales.

Ceci permet de comprendre les bases du modèle de régression multiple. On a dit que quand on avait $y = X\beta + \varepsilon$, les autres déterminants de y que les X sont dans la perturbation. Mais afin de bien connaître l'influence des X sur y, il faut que les perturbations soient orthogonales au X.

Ceci nous permet aussi de retrouver le concept de toute chose égale par ailleurs.

Application : Etude de l'introduction d'une variable indicatrice

Supposons que nous disposions d'observations temporelles pour estimer le modèle de régression multiple. Supposons que sur notre échantillon, nous avons une observation particulière $\tilde{t}$, qui peut ne pas être représentative du phénomène que l'on cherche à modéliser. Pour prendre en compte cette éventualité, on intègre dans le modèle une variable indicatrice (ou variable muette), $d = 1$ si $t = \tilde{t}$ et 0 sinon.

Sous forme matricielle, le modèle s'écrit :

$$y = X\beta + d\alpha + \varepsilon$$

Estimer ce modèle revient à estimer l'influence de X sur y en éliminant l'observation $t = \tilde{t}$. Ainsi, introduire la variable d revient à considérer l'observation $\tilde{t}$ comme à

part. En effet, quand on régresse X sur d, on obtient X partout (puisque quand on régresse sur 0, l'ajustement est 0 et donc le résidu est la variable elle même) sauf pour $t = \tilde{t}$ où on obtient 0 (puisque quand on régresse sur 1, on obtient la même valeur et le résidu est nul!). C'est la même chose avec y. Ainsi, quand on calcule l'estimateur de β, on voit intervenir $\sum_{t \neq \tilde{t}} X_t X_t'$ et $\sum_{t \neq \tilde{t}} X_t y_t$.

$\hat{\beta}$ est donc l'estimateur des MCO calculé en excluant l'observation $\tilde{t}$. Quant à l'estimation de α, on trouve que $\hat{\alpha} = y_{\tilde{t}} - X_{\tilde{t}}' \hat{\beta}$, la différence entre ce qui a été observé en $\tilde{t}$ et ce qui est prédit à partir de l'estimation sans inclure l'observation $\tilde{t}$. Afin de tester si l'observation $t = \tilde{t}$ est aberrante, on étudiera la nullité de α.

3 Equation d'analyse de la variance et qualité de l'ajustement

L'idée est de savoir si le modèle (les variables explicatives présentes dans le modèle) permet de bien expliquer les variations de la variable endogène y.

De la même manière que dans le modèle de régression simple, on a

$$y = \hat{y} + \hat{\varepsilon}$$

et comme on l'a vu précédemment, on a $X'\hat{\varepsilon} = 0$, ce qui implique que $\sum_i y_i = \sum_i \hat{y}_i$ donc $\bar{y} = \bar{\hat{y}}$.

On retrouve alors l'équation d'analyse de la variance :

$$SCT = SCE + SCR$$

et bien entendu, on en déduit le coefficient de détermination :

$$R^2 = \frac{SCE}{SCT}$$

4 Les tests d'hypothèse

4.1 Test de student

La mise en place d'un test de student, pour tester la nullité d'un coefficient de régression (par exemple $H_0 \ : \ \beta_1 = 0$) ou alors pour tester une fonction des paramètres du modèle (par exemple $H_0 \ : \ \alpha + \beta = 1$, c'est-à-dire hypothèse de rendements d'échelle constants, dans le modèle $y = \alpha l + \beta k + c + u$), se fait comme dans le modèle de régression simple, à l'exception bien sûr que la statistique de student, par exemple, pour le test de $H_0 \ : \ \beta_1 = 0$, donnée par :

$$t = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}}$$

suit sous H_0, une loi de Student à $n - K - 1$ degrés de liberté (et non $n - 2$ comme dans le modèle de régression simple) puisqu'on a estimé $K + 1$ degrés de liberté.

4.2 Test de Fisher

La statistique de Fisher permet de tester plusieurs contraintes dans un modèle général. Elle consiste à comparer deux modèles emboîtés, c'est-à-dire que le modèle sous H_0 est obtenu en imposant des contraintes sur les paramètres dans le modèle sous H_1. Ainsi, on teste le modèle (Ω_0) sous l'hypothèse nulle H_0 contre le modèle (Ω_1) sous H_1.

La statistique de Fisher est alors :

$$F = \frac{(SCR_0 - SCR_1)/(K_1 - K_0)}{SCR_1/(n - K_1 - 1)}$$

où SCR_0 et SCR_1 sont respectivement la somme des carrés des résidus estimée sous H_0 et sous H_1. $K_1 + 1$ et $K_0 + 1$ sont respectivement le nombre de paramètres estimés sous H_0 et sous H_1. Ainsi, $K_1 - K_0$ est le nombre de contraintes testées.

La statistique suit sous H_0 une loi de Fisher à $K_1 - K_0$ et $n - K_1 - 1$ degrés de liberté. La région de rejet de H_0 est donnée par :

$$R_\alpha = \{F > Fisher^{\alpha}_{K_1-K_0,n-K_1-1}\}$$

où $Fisher^{\alpha}_{K_1-K_0,n-K_1-1}$ est la valeur lue dans une table de Fisher à $K_1 - K_0$ et $n - K_1 - 1$ degrés de liberté au seuil α.

Application a) Test de significativité globale de la régression

Il s'agit de tester la nullité de tous les paramètres sauf la constante dans le modèle :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i \quad (\Omega)$$

Ainsi, on teste :

$$H_0 : \ \beta_1 = \beta_2 = \ldots = \beta_K = 0 \qquad \text{soit} \ \ y_i = \beta_0 + \varepsilon_i \quad (\Omega_0)$$

contre

$$H_1 \ : \text{ il existe au moins une inégalité : on a le modèle général } (\Omega)$$

L'estimation du modèle sous H_0 conduit à $\hat{\beta}_0 = \bar{y}$ et donc $SCR_0 = \sum_i (y_i - \bar{y})^2 = SCT$. La somme des carrés des résidus sous l'hypothèse alternative (SCR_1) est tout simplement SCR (de (Ω)).

La statistique de Fisher est alors définie par :

$$F = \frac{(SCT - SCR)/K}{SCR/(n - K - 1)} = \frac{R^2/K}{(1 - R^2)/(n - K - 1)}$$

Ce test revient à regarder si le R^2 est suffisamment proche de 1, et donc si la régression est globalement significative, c'est-à-dire si l'ensemble des variables explicatives prises globalement est significatif.

Attention, le fait de rejeter H_0 ne signifie pas que toutes les variables sont significatives, mais que globalement la régression est bonne et donc qu'il y a des variables

significatives dans l'ensemble. Une fois ce test effectué, il est nécessaire d'aller étudier la significativité de chaque variable prise séparément par un test de student.

Remarque : on peut tester la significativité d'une seule variable par un test de Fisher. Dans ce cas, $F = t^2$ et cela revient donc à faire un test de student.

Application b) Test de constance des paramètres ou de l'absence de rupture

On veut étudier si les paramètres sont différents avant et après une certaine date, ou alors s'ils diffèrent pour deux groupes de population distincts. Il s'agit d'un test de changement de structure.

Pour cela, on doit au préalable s'assurer que la variance de l'erreur est constante, c'est-à-dire que la variance du premier sous-échantillon (notée σ_1^2) est la même que la variance du second sous-échantillon (notée σ_2^2), sinon on ne serait plus sous les hypothèses permettant d'appliquer les MCO (on parle de test d'égalité des variances, on verra cela dans le chapitre suivant).

Sous l'hypothèse que $\sigma_1^2 = \sigma_2^2$, on peut faire le **test de Fisher d'égalité des paramètres** entre les deux sous-échantillons. Il s'agit de tester :

$$H_0 \ : \begin{cases} \beta_0^{(1)} = \beta_0^{(2)} \\ \beta_1^{(1)} = \beta_1^{(2)} \\ \vdots \\ \beta_K^{(1)} = \beta_K^{(2)} \end{cases}$$

dans le modèle :

$$y_i = \begin{cases} \beta_0^{(1)} + \beta_1^{(1)} x_{1i} + \ldots + \beta_K^{(1)} x_{Ki} + \varepsilon_i & i \in \text{ population (1)} \\ \beta_0^{(2)} + \beta_1^{(2)} x_{1i} + \ldots + \beta_K^{(2)} x_{Ki} + \varepsilon_i & i \in \text{ population (2)} \end{cases}$$

La statistique de Fisher est alors :

$$F = \frac{(SCR_0 - SCR_1)/(K_1 - K_0)}{SCR_1/(n - K_1 - 1)}$$

Or, sous H_0, le modèle est :

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_K x_{Ki} + \varepsilon_i$$

Donc SCR_0 est la somme des carrés des résidus dans ce modèle et $K_0 = K + 1$.

Pour le modèle sous H_1, on a $SCR_1 = \sum_{i \in (1)} \hat{\varepsilon}_i^2 + \sum_{i \in (2)} \hat{\varepsilon}_i^2$. On peut donc estimer la régression sur chaque sous-population et calculer chaque SCR, leur somme donnant SCR_1. $K_1 = (K + 1) + (K + 1)$.

Ainsi,

$$F = \frac{(SCR_0 - SCR_1)/K}{SCR_1/(n - 2K - 2)}$$

5 Variables indicatrices

Une variable indicatrice, dite aussi variable muette, est une variable qui vaut 0 ou 1. Elle peut être introduite dans le modèle de régression multiple pour tester la présence de certains effets (changement temporel, homogénéité du comportement entre différents groupes, etc) que l'on ne peut pas séparer simplement. Pour étudier ce type d'effets, il suffit d'introduire la variable indicatrice valant 1 pour l'effet que l'on souhaite isoler, et 0 sinon, dans la régression étudiée. La méthode par MCO reste valide et les tests décrits précédemment aussi.

Afin de comprendre l'idée, prenons un exemple simple, consistant à comparer la moyenne de deux sous-populations. La comparaison de la moyenne d'une variable selon 2 groupes peut être formulée comme cela :

$$\begin{array}{ll} y_i = \mu + \varepsilon_i & \text{si } i \in (1) \\ y_i = \mu + \delta + \varepsilon_i & \text{si } i \in (2) \end{array}$$

Le test d'égalité des moyennes entre les deux sous-populations consiste alors à tester $\delta = 0$ par un test de student, mais pour cela, il faut avoir réécrit ces 2 modèles dans une seule régression. On introduit une variable indicatrice d :

$$d_i = \begin{cases} 0 & \text{si } i \in (1) \\ 1 & \text{si } i \in (2) \end{cases}$$

et on récrit une seule régression :

$$y_i = \mu + \delta d_i + \varepsilon_i \quad i = 1, \ldots, n$$

Ainsi, δ mesure la différence de moyenne (du salaire ou de la taille par exemple) entre les 2 groupes (homme-femme, noir-blanc, etc). Pour savoir si la différence est significative, il suffit de tester $H_0 \ : \ \delta = 0$ par un test de Student.

On aurait aussi pu écrire :

$$y_i = \mu_1 h_i + \mu_2 d_i + \varepsilon_i$$

avec $h_i = 1$ si $i \in (1)$ et 0 sinon et $h_i = 1 - d_i$. C'est bien sûr la même chose, avec $\mu_1 = \mu$ et $\mu_2 = \mu + \delta$. Tester l'égalité des moyennes revient alors à tester : $H_0 \ : \ \mu_1 = \mu_2$ par un test de student :

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\hat{V}(\hat{\mu}_1) + \hat{V}(\hat{\mu}_2) - 2\hat{cov}(\hat{\mu}_1, \hat{\mu}_2)}}$$

Cependant, on préférera la première écriture à celle-ci puisque cette dernière nous empêche de conserver la constante dans le modèle, et nous empêche aussi d'introduire d'autres variables indicatrices, puisque l'écriture suivante :

$$y_i = \mu + \mu_1 h_i + \mu_2 d_i + \varepsilon_i$$

poserait des problèmes de multicolinéarité puisque $h_i + d_i$ est égal au vecteur constant.

Prendre en compte un changement temporel :

Introduire la variable $D_t = 1$ si $t < 1973$ et 0 sinon dans un modèle où les données sont temporelles permet d'étudier l'effet du choc pétrolier sur le phénomène étudié. Cela permet donc de prend explicitement en compte des changements temporels.

Supposons que l'on s'intéresse à la relation entre les dépenses de consommation C et le revenu R au cours d'une période donnée (de 1 à T), et on cherche à savoir s'il y a eu un changement dans le comportement de consommation après la date t_1 (par exemple après 1973). Le modèle intégrant le changement temporel s'écrit :

$$\begin{aligned} C_t &= \beta_0 + \beta_1 R_t + \varepsilon_t \qquad t = 1, \ldots, t_1 \\ C_t &= \alpha_0 + \alpha_1 R_t + \varepsilon_t \quad t = t_1 + 1, \ldots, T \end{aligned}$$

Soit on estime le modèle sur chaque sous-période et on construit un test de Fisher, où $H_0 \ : \ \beta_0 = \alpha_0, \beta_1 = \alpha_1$ sous l'hypothèse que la variance de ε_t reste constante entre les deux sous-périodes (section précédente), soit on estime le modèle sur l'ensemble de la période en introduisant la variable indicatrice $D_t = 1$ si $t \leq t_1$ et 0 sinon. Le modèle estimé est alors :

$$C_t = \beta_0 + \phi_0 D_t + \beta_1 R_t + \phi_1 (D_t R_t) + \varepsilon_t$$

avec $\alpha_0 = \beta_0 + \phi_0$ et $\alpha_1 = \beta_1 + \phi_1$. Pour tester l'hypothèse d'absence de changement temporel en t_1, on doit tester $H_0 \ : \ \phi_0 = \phi_1 = 0$ par un test de Fisher. Dans ce modèle, ϕ_0 mesure directement la différence dans la moyenne de la consommation entre les deux sous-périodes et ϕ_1 mesure directement la différence dans la propension à consommer entre les deux sous-périodes.

On pourrait aussi introduire la variable $D_t = 1$ si $t = 1968$ et 0 sinon , qui permet alors de mesurer l'effet de l'année 1968 pour voir si le comportement est différent cette année là.

Prendre en compte des effets individuels :

Introduire la variable $D_i = 1$ si l'individu i est un homme et 0 s'il est une femme dans un modèle de régression où les données sont individuelles permet d'étudier la présence de discrimination de genre, par exemple sur le salaire. Le modèle est alors donné par :

$$w_i = \beta_0 + X_i' \beta + \delta D_i + \varepsilon_i$$

où X_i sont les variables explicatives du salaire, par exemple l'age, le niveau d'expérience ou le niveau d'étude. Le coefficient δ mesure alors l'accroissement de salaire moyen pour les hommes relativement aux femmes. On peut tester sa significativité par un test du Student.

On aurait aussi pu chercher si les effets des variables explicatives X_i sont les mêmes pour les hommes et pour les femmes (si une année d'expérience professionelle en plus accroît le salaire du même montant pour les hommes et pour les femmes). Il faut alors estimer le modèle suivant :

$$w_i = \beta_0 + X_i' \beta + (X_i' D_i) \alpha + \delta D_i + \varepsilon_i$$

α mesure la différence d'effet des variables X_i sur w_i quand on est un homme plutôt que quand on est une femme.

Prendre en compte des effets saisonniers :

Introduire les variables $T_{jt} = 1$ si t correspond au $j$ème trimestre et 0 sinon pour $j = 1, \ldots, 4$ dans un modèle où les observations sont trimestrielles permet d'étudier la présence d'effets saisonniers.

Supposons que l'on dispose de données trimestrielles, qui semblent comporter un effet saisonnier. Afin de tester cet effet, et de l'enlever s'il est présent, on introduit les 4 variables saisonnières $T_{1t}, T_{2t}, T_{3t}, T_{4t}$. Le modèle est alors :

$$y_t = \delta_0 + \delta_1 T_{1,t} + \delta_2 T_{2,t} + \delta_3 T_{3,t} + \delta_4 T_{4,t} + \varepsilon_t \Longrightarrow y = X\delta + \varepsilon$$

Cependant, la matrice X de régresseurs ne sera plus de plein rang colonne puisque la somme des 4 dernières colonnes est égale à la constante. En effet, on a :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 \\ . & . & . & . & . \\ 1 & . & . & . & . \end{pmatrix}$$

Comme on préfère en général garder la constante dans le modèle, il faut réécrire différemment le modèle. On peut par exemple centrer l'effet saisonnier, c'est-à-dire imposer une contrainte de telle sorte que l'espérance de la variable soit la même avec l'effet saisonnier que sans. Cela signifie que :

$$E[\delta_0 + \delta_1 T_{1t} + \delta_2 T_{2t} + \delta_3 T_{3t} + \delta_4 T_{4t}] = \delta_0$$

ou autrement dit, que :

$$\sum_{j=1}^{4} \delta_j = 0 \Longrightarrow \delta_4 = -(\delta_1 + \delta_2 + \delta_3)$$

Le modèle se réécrit alors :

$$y_t = \delta_0 + \delta_1(T_{1,t} - T_{4,t}) + \delta_2(T_{2,t} - T_{4,t}) + \delta_3(T_{3,t} - T_{4,t}) + \varepsilon_t \Longrightarrow y = X\delta + \varepsilon$$

où la nouvelle matrice des régresseurs est donnée par :

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & -1 & -1 & -1 \\ 1 & 1 & 0 & 0 \\ . & . & . & . \\ 1 & . & . & . \end{pmatrix}$$

Tester l'absence d'un effet saisonnier revient à tester :

$$H_0 \ : \delta_1 = \delta_2 = \delta_3 = 0$$

dans le modèle ci-dessus, par un test de Fisher.

Si l'hypothèse nulle est rejetée, alors il y a un effet trimestriel significatif et la variable y désaisonnalisée (ou CVS pour Corrigée des Variations Saisonnières) est donnée par :

$$y_t^{CVS} = y_t - \left[\hat{\delta}_1(T_{1,t} - T_{4,t}) + \hat{\delta}_2(T_{2,t} - T_{4,t}) + \hat{\delta}_3(T_{3,t} - T_{4,t})\right]$$

Généralisation au delà d'une variable dichotomique :

Une variable dichotomique est une variable qui peut prendre uniquement deux valeurs. Supposons maintenant que l'on dispose d'une variable qui peut prendre plus de deux valeurs. Dans ce cas, il faut alors introduire plus de 1 variable indicatrice.

Supposons, par exemple, que l'on étudie l'effet des diplômes sur le salaire. On dispose alors d'une variable diplôme, qui nous dit que la personne a, soit 0 diplôme, soit un diplôme niveau BAC, soit un diplôme supérieur.

Si on code cette variable :

$$d_i = \begin{cases} 1 & \text{si l'individu } i \text{ est sans diplôme} \\ 2 & \text{si l'individu } i \text{ a un diplôme niveau BAC} \\ 3 & \text{si l'individu } i \text{ a un diplôme supérieur} \end{cases}$$

et qu'on introduise cette variable dans la régression, alors le coefficient associé sera sans doute positif, signifiant que plus un individu a de diplômes, plus son salaire est élevé, mais il est insatisfaisant de se dire que l'augmentation de salaire pour chaque diplôme est la même. En effet, le coefficient qui sera estimé correspond à l'augmentation de salaire quand on passe de sans diplôme à BAC, ou quand on passe de BAC à diplôme supérieur.

Une manière plus satisfaisante serait d'écrire :

$$w_i = \beta age_i + \delta_1 SD_i + \delta_2 BAC_i + \delta_3 SUP_i + \varepsilon_i$$

où SD_i (respectivement BAC_i et SUP_i) vaut 1 si l'individu i est sans diplôme (respectivement a un BAC, et diplôme supérieur) et 0 sinon. Ainsi, δ_1 est le coefficient associé à sans diplôme, δ_2 à BAC et δ_3 à diplôme supérieur. Cependant, on note qu'on ne peut plus conserver la constante dans le modèle puisque la somme des trois variables indicatrices, SD_i, BAC_i, SUP_i est égale au vecteur constant (colinéarité). Ainsi, on préférera retirer une des variables (on dit que l'on met en "référence" la modalité correspondante) et n'introduire que les deux autres, ce qui nous permet de conserver la constante. Le modèle s'écrit alors, si on met "sans diplôme" en référence :

$$w_i = \beta_0 + \beta age_i + \delta_2 BAC_i + \delta_3 SUP_i + \varepsilon_i$$

Les valeurs δ_i sont interprétables : δ_2 mesure l'accroissement de salaire quand on a le BAC relativement à ne pas avoir de diplôme, et δ_3 mesure l'accroissement de salaire quand on a un diplôme supérieur relativement à ne pas avoir de diplôme.

6 D'autres tests

Comme dans le chapitre précédent, pour le modèle de régression simple, on peut facilement montrer que l'estimateur par MCO du vecteur de paramètre β correspond exactement à l'estimateur par Maximum de Vraisemblance, dès que le nombre d'observations est assez grand.

Trois procédures de tests, asymptotiquement équivalents, sont très souvent utilisés quand le nombre d'observations est assez grand (c'est-à-dire quand l'estimateur du maximum de vraisemblance est efficace) : le test de Wald, le test du ratio de vraisemblance et le test du multiplicateur de Lagrange (noté LM).

Ces tests reposent sur une estimation des paramètres réalisée par la méthode du maximum de vraisemblance. La log-vraisemblance est définie par :

$$L(y_1, \ldots, y_n; \beta, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{n}{2}\log(2\pi) - \frac{\sum_i (y_i - X_i'\beta)^2}{2\sigma^2}$$

La méthode d'estimation du maximum de vraisemblance (MV) consiste à choisir $\hat{\beta}$ et $\hat{\sigma}^2$ de façon à maximiser la log-vraisemblance. Il s'agit de trouver les valeurs de β et σ^2 qui annulent les dérivées premières de la log-vraisemblance par rapport à chaque paramètre.

Supposons que l'on veuille tester une contrainte sur les paramètres β, notée H_0 : $c(\beta) = 0$. Les trois tests permettant de tester cette hypothèse sont les suivants :

- **Ratio de vraisemblance**: si H_0 est valide, alors imposer cette contrainte ne doit pas conduire à une importante diminution de la log-vraisemblance. Le test est donc basé sur la différence de log-vraisemblance entre le modèle non contraint L_1 et le modèle contraint L_0 :

$$LR = -2(L_0 - L_1)$$

- **Test Wald**: si H_0 est valide, alors $c(\hat{\beta})$ doit être proche de 0 puisque l'estimateur du MV est convergent.

$$W = [c(\hat{\beta})]'[\ Var(c(\hat{\beta}))]^{-1}[c(\hat{\beta})]$$

ou si on récrit la contrainte testée sous la forme $H_0 : R\beta - q = 0$

$$W = [R\hat{\beta} - q]'[R\ Var(\hat{\beta})R']^{-1}[R\hat{\beta} - q]$$

Si on ne teste qu'une seule contrainte, W revient au carré du student.

- **Test LM**: si H_0 est valide, alors l'estimateur contraint doit être proche du β qui maximise la vraisemblance. Ainsi, la pente de la log-vraisemblance pour l'estimateur de β contraint sous H_0, noté $\hat{\beta}_{H_0}$, doit être proche de 0.

$$LM = \left(\frac{\partial lnL}{\partial \beta}\Big|_{\beta=\hat{\beta}_{H_0}}\right)' \left[I(\beta)\Big|_{\beta=\hat{\beta}_{H_0}}\right]^{-1} \left(\frac{\partial lnL}{\partial \beta}\Big|_{\beta=\hat{\beta}_{H_0}}\right)$$

On peut estimer la matrice d'information de Fisher $I(\beta)$ par le produit des gradients.

Si on teste la nullité de tous les coefficients sauf la constante, on montre que le test LM peut se réécrire comme $LM = nR^2$ où R^2 est le coefficient de détermination de la régression générale.

Chacune de ces statistiques suit un chi-deux sous H_0 avec comme degré de liberté le nombre de contraintes testées.

7 Quelques problèmes rencontrés dans la régression : tests des hypothèses du modèle

L'estimation de β, ainsi que les tests présentés précédemment, ont été définis sachant les hypothèses effectuées pour le modèle de régression. Nous devons maintenant vérifier que ces hypothèses sont acceptables pour que les résultats obtenus soient valides. Sinon, les résultats ne peuvent être interprétés.

La première hypothèse concernant le modèle de régression était la linéarité. On peut vérifier s'il n'aurait pas été pertinent de mettre des variables explicatives sous une forme différente. Pour cela, on peut observer la représentation graphique des résidus estimés en fonction de chaque variable explicative afin de voir si une relation quelconque n'apparaît pas. Par exemple, on peut observer une relation quadratique entre les résidus et une variable explicative, indiquant qu'on a omis la variable explicative au carré dans la régression. Il suffit alors de tester la significativité de cette variable au carré dans la régression.

Concernant les hypothèses sur la matrice des régresseurs, il s'agit de vérifier l'absence de multicolinéarité entre les régresseurs. Ceci fait l'objet de la première sous-section ci-dessous.

Nous étudierons ensuite le problème de la spécification du modèle, à savoir les risques que posent l'omission de variables ou la présence de variables supplémentaires dans la régression. Ceci nous conduira ensuite à étudier le choix du nombre de variables explicatives à introduire et la sélection de modèles parmi plusieurs.

Concernant les hypothèses sur les erreurs, il s'agit de mettre en œuvre des tests de normalité, d'homoscédasticité et d'absence d'autocorrélation.

7.1 Le problème de multicolinéarité

Si deux variables explicatives sont parfaitement corrélées, alors la matrice de régresseurs X n'est pas de plein rang colonne et $X'X$ ne sera pas inversible : on ne pourra pas calculer l'estimateur des MCO. De plus, le modèle n'est pas identifiable, on ne pourra pas estimer l'effet propre à chaque variable.

En revanche, si la corrélation n'est pas parfaite mais très élevée, on parle de quasi-colinéarité, on pourra calculer l'estimateur des MCO. Cependant, la variance de l'estimateur sera très élevée et l'interprétation des coefficients estimés sera difficile. En effet, un coefficient de régression mesure, ce qui, dans la variation de y, est du au changement de la variable explicative associée, toute chose égale par ailleurs. Or si cette variable est fortement corrélée avec une autre, alors un changement dans la variable va conduire à un changement dans l'autre variable! Face à un tel problème,

appelé aussi problème d'identification, la notion de paramètres n'a pas de sens. Ils seront biaisés et non efficaces.

S'il y a un problème entre plusieurs variables, on parle de multicolinéarité, cela veut dire que plusieurs vecteurs de paramètres différents donnent la même espérance de y.

Comment détecter la présence de multicolinéarité? En présence de régresseurs fortement corrélés, on peut observer les symptômes suivants :

- de petits changements dans les données produisent de grands changements dans les paramètres estimés;
- les coefficients ont de grands écart-types estimés et de faibles niveaux de significativité tout en étant globalement significatifs (F élevé et R^2 élevé);
- quand on a 2 (ou plus de 2) variables explicatives qui sont corrélées, elles ne paraissent pas significatives quand elles sont toutes dans la régression, mais si on enlève une variable non significative, l'autre (ou les autres) le deviennent quelque soit la variable qu'on retire;
- les coefficients de corrélation partielles (entre une variable explicative et les autres) sont très élevés;
- les coefficients peuvent avoir de mauvais signe ou des valeurs impossibles!

Dans la pratique, il est impossible d'avoir des régresseurs parfaitement orthogonaux : $R_k^2 \neq 0$ où R_k^2 est le R^2 de la régression de la $k$ième variable sur toutes les autres. Or la variance du coefficient associé à la $k$ième variable augmente avec R_k^2. A l'extrême, quand $R_k^2 = 1$, alors la variance devient infinie car elle s'écrit :

$$\frac{\sigma^2}{(1 - R_k^2) \sum_i (x_{ik} - \bar{x}_k)^2}$$

Ainsi, on sera amener à accepter la non-significativité de cette variable à tord.

Comment détecter la multicolinéarité? Quand est-ce qu'on doit considérer que la variance d'un coefficient est trop élevée et que cela peut poser problème?

Certains logiciels donnent un critère, appelé VIF (*Variance Inflation Factor*) : $\frac{1}{1-R_k^2}$ pour chaque coefficient de la régression comme statistique de diagnostic. Il mesure l'augmentation de la variance du paramètre estimé due à la présence de colinéarité. S'il est élevé, cela pose problème, mais il n'existe pas de règle pour savoir s'il est trop élevé pour affecter les valeurs estimées.

On peut aussi calculer les coefficients de corrélation partiel. Certains estiment que quand le R_k^2 est supérieur au R^2 global, la multicolinéarité est forte (Klein 1962).

Un autre indicateur repose sur les valeurs propres de $X'X$. Le logiciel calcule un critère (*Condition index*), qui est égal à la racine carré du ratio de la plus grande valeur propre à chaque valeur propre. Si ce critère dépasse 100, il est fort probable qu'il y ait au des problèmes lors de l'estimation.

Comment faire en présence de multicolinéarité?

Il n'y a pas de solution miracle! Quand cela est possible, il faut reformuler le modèle, une possibilité est bien sûr d'enlever une des variables qui pose problème, mais attention, cela peut causer des biais si cette variable est explicative.

7.2 Spécification du modèle – Choix de variables explicatives

Tout ce qu'on a fait précédemment repose sur l'hypothèse que le bon modèle est connu et qu'il est donné par : $y = X\beta + \varepsilon$. Mais il se peut qu'on ait oublié des variables explicatives ou qu'il y ait des variables non pertinentes.

a) Omission de variables pertinentes

Supposons que le vrai modèle soit :

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon$$

mais que l'économètre a oublié X_2 et qu'il régresse seulement y sur X_1, il obtient :

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 + (X_1'X_1)^{-1}X_1'\varepsilon$$

En prenant l'espérance, on voit que l'estimateur est biaisé (sauf si $X_1'X_2 = 0$ ou $\beta_2 = 0$).

La variance de l'estimateur est :

$$var(\hat{\beta}_1) = \sigma^2(X_1'X_1)^{-1}$$

S'il avait fait la bonne régression en incluant X_2, la variance de β_1 aurait été celle de l'élément en haut à gauche de la matrice $\sigma^2(X'X)^{-1}$, soit :

$$var(\hat{\beta}_{1,2}) = \sigma^2(X_1'M_2X_1)^{-1} = \sigma^2[X_1'X_1 - X_1'X_2(X_2'X_2)^{-1}X_2'X_1]^{-1}$$

Pour comparer les deux variances, il est plus pratique de prendre leur inverse :

$$var(\hat{\beta}_1)^{-1} - var(\hat{\beta}_{1,2})^{-1} = (1/\sigma^2)X_1'X_2(X_2'X_2)^{-1}X_2'X_1$$

qui est non négative. Ainsi, même si $\hat{\beta}_1$ est biaisé, il a une variance plus faible (puisque la variance de l'inverse est plus grande).

On peut remarquer que l'estimateur de σ^2 sera biaisé aussi car :

$$\hat{\varepsilon}_1 = M_1y = M_1(X_1\beta_1 + X_2\beta_2 + \varepsilon) = M_1X_2\beta_2 + M_1\varepsilon$$

Donc

$$E[\hat{\varepsilon}_1'\hat{\varepsilon}_1] = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2 tr(M_1) = \beta_2'X_2'M_1X_2\beta_2 + \sigma^2(n - K_1)$$

Cet estimateur reste biaisé même si $X_1'X_2 = 0$ (régresseur orthogonal qui permet de ne pas avoir un coefficient $\hat{\beta}_1$ biaisé).

Ainsi, oublier une variable explicative pertinente conduit à des estimateurs biaisés pour les autres paramètres, et à des variances estimées de ces paramètres erronées. Faut-il alors préférer mettre beaucoup de variables explicatives quitte à ce qu'elles ne soient pas pertinentes?

b) Variables supplémentaires

Supposons que le vrai modèle ne comporte que X_1 et que l'économètre régresse y sur X_1 et sur X_2. Dans ce cas, il est facile de montrer que l'estimateur sera non biaisé, on aura en effet :

$$E[\beta] = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ 0 \end{bmatrix}$$

Ainsi, l'estimateur de σ^2 ne sera pas biaisé non plus.

Alors, où est le problème? en fait, on utilise de l'information inutile, ce qui réduit la précision. Ainsi, si X_2 est fortement corrélée avec X_1, cela va conduire à augmenter de manière inutile la variance de l'estimateur, et donc à accepter l'hypothèse selon laquelle les variables ne sont pas pertinentes (et donc à oublier des variables).

Il importe donc de disposer de critères afin de choisir le "meilleur modèle", c'est–à–dire le nombre de variables explicatives à intégrer dans le modèle et l'ensemble pertinent de variables à retenir.

c) Sélection de modèles

Supposons que l'on dispose d'un (large) ensemble $x_1, x_2, \ldots, x_K$ de variables explicatives. L'idée est de savoir quel sous-ensemble retenir parmi cet ensemble de variables explicatives afin que le modèle ainsi retenu soit le mieux spécifié possible, c'est–à–dire qu'il n'y ait pas de variables omises (faut-il encore qu'on ait considéré initialement un ensemble suffisamment large de variables) et pas de variables supplémentaires.

Nous allons présenter des critères statistiques permettant de guider l'économètre dans le choix de modèle. Il s'agit tout d'abord de critères permettant de choisir le modèle qui conduit au meilleur ajustement tout en étant parcimonieux (critères d'information et méthode *stepwise*) puis de critères permettant de choisir le modèle qui fournit les meilleures performances en prévision.

Critères d'ajustement

Nous avons déjà présenté le R^2 comme indicateur d'ajustement, mais celui-ci a l'inconvénient de ne pas pouvoir diminuer quand on introduit des variables explicatives. En effet, plus on met de variables explicatives et plus il augmente par un effet purement mécanique. A la limite, quand le nombre de variables explicatives est égal au nombre d'observations, on obtient un R^2 égal à 1. La variance de y est expliquée à 100%, quelque soit la pertinence économique des variables explicatives utilisées, pourvu que les hypothèses du modèle soient respectées (notamment indépendance linéaire des vecteurs des observations des variables explicatives). Ainsi, un choix de variables reposant sur le R^2 conduira toujours à inclure toutes les variables explicatives proposées, y compris des variables non pertinentes.

On considère alors le coefficient de détermination corrigé (par le nombre de variables explicatives dans la régression) afin de tenir compte du critère de parcimonie et de corriger l'effet mécanique du R^2. Ce R^2 ajusté est défini par :

$$\bar{R}^2 = 1 - \frac{n-1}{n-K}(1-R^2) = 1 - \frac{n-1}{n-K}\frac{SCR}{SCT}$$

On remarque que le $\bar{R}^2$ peut diminuer quand on ajoute une variable si la SCR du modèle ne diminue pas suffisamment.

Ainsi, on peut calculer le $\bar{R}^2$ pour tous les modèles possibles obtenus à partir de toutes les combinaisons possibles de $1, 2, 3, \ldots, K$ variables parmi l'ensemble $x_1, x_2, \ldots, x_K$. On retient l'ensemble de variables explicatives qui conduit à maximiser ce critère.

Soulignons cependant que ce critère, pour pertinent qu'il soit, ne suffit pas à guider le modélisateur dans son choix de variables explicatives. Il présente, à l'instar du R^2, l'inconvénient d'être aisément manipulable par une transformation du vecteur y. De plus, il a le défaut de ne pas pouvoir s'interpréter à partir de l'équation d'analyse de la variance. Il peut en particulier être négatif.

Que ce soit le R^2 ou le $\bar{R}^2$, ces coefficients doivent toujours être utilisés avec précaution et jamais comme critère unique pour juger de la qualité d'une régression.

Pour comparer 2 modèles qui n'ont pas les mêmes variables explicatives, on ne peut utiliser le R^2 que si c'est la même variable y que l'on cherche à expliquer (même nombre d'observations et même forme retenue pour y) et que si les 2 modèles ont le même nombre de variables explicatives.

Critères d'information

Ces critères sont utilisés généralement pour choisir le nombre de variables explicatives à introduire dans le modèle. Les plus connus sont le critère d'Akaïke (AIC) et celui de Schwarz (BIC). Pour un modèle avec K variables explicatives, ils sont donnés par :

$$AIC = s_y^2(1 - R^2)e^{2(K+1)/n}$$

et

$$BIC = s_y^2(1 - R^2)n^{(K+1)/n}$$

On cherche le nombre K qui minimisent ces critères. En effet, ce critères diminuent quand le R^2 augmente, mais ils augmentent quand le nombre de régresseurs augmente. Ainsi, ils permettent de pénaliser la variance expliquée du nombre de degré de liberté, et donc de trouver un modèle parcimonieux.

On utilise en général le log de ces critères :

$$AIC = \log(SCR/n) + \frac{2(K+1)}{n}$$

et

$$BIC = \log SCR/n + \frac{(K+1)\log n}{n}$$

Ces 2 critères ont chacun leurs avantages et aucun n'est vraiment meilleur. Le critère de Schwarz, qui pénalise fortement pour la perte de degré de liberté, conduira plus souvent à un modèle très parcimonieux.

Méthodes automatiques de choix de variables :

Dans le cadre de l'estimation MCO d'un modèle de régression multiple, les logiciels d'économétrie proposent généralement des procédures d'aide au choix d'un modèle, c'est–à–dire au choix d'un ensemble de variables explicatives. Trois méthodes peuvent être utilisées :

- La méthode *forward* : elle consiste à introduire comme première variable explicative la variable la plus corrélée à la variable à expliquer. Si cette variable est significative (au seuil d'erreur qu'on peut donner), elle recherche parmi l'ensemble des variables explicatives restantes celle dont le coefficient de corrélation partielle est le plus élevé, c'est–à–dire la corrélation entre y et cette variable, conditionnellement à la présence de la variable déjà introduite. Si cette nouvelle variable n'est pas significative, la procédure propose de retenir le modèle obtenu avant l'introduction de cette dernière variable. Si elle est significative, la procédure continue d'introduire des variables jusqu'à ce qu'il n'y ait plus de variable significative à introduire.

- La méthode *backward* : elle consiste à introduire toutes les variables explicatives et à regarder leur significativité. Si certaines variables ne sont pas significatives, elle retire la moins significative. Une fois cette variable retirée, elle regarde la significativité des restantes et retire, parmi les non significatives, la moins significative. La procédure propose alors à retenir le modèle où toutes les variables restantes sont significatives.

- La méthode *stepwise* : elle commence comme la méthode forward, puis quand elle ajoute une variable, elle teste si les variables précédemment introduites sont toujours significatives. Sinon, elle retire la moins significative et tente ensuite d'en introduire une autre. Elle combine donc les deux méthodes précédentes.

Critères basés sur les performances en prévision des modèles

Une fois les paramètres estimés, on peut les utiliser pour construire une prévision pour la valeur future (inconnue) de y, notée y_0. Cette valeur (inconnue) est telle que :

$$y_0 = X_0'\beta + \varepsilon_0$$

et on prédira :

$$\hat{y}_0 = X_0'\hat{\beta}$$

Ainsi, l'erreur de prévision (estimée) est donnée par :

$$\hat{\varepsilon}_0 = y_0 - \hat{y}_0 = X_0'(\beta - \hat{\beta}) + \varepsilon_0$$

Sa variance est donc donnée par :

$$var(\hat{\varepsilon}_0) = \sigma^2 + X_0'[\sigma^2(X'X)^{-1}]X_0'$$

Ainsi, la largeur de l'intervalle de prévision dépend de la distance de X_0' avec la moyenne des données. C'est très intuitif : plus le point prévu est loin de la moyenne de notre expérience, plus le degré d'incertitude est grand.

L'intervalle de prévision pour y_0 est

$$\hat{y}_0 \pm t_{\alpha/2}\widehat{\sigma}(\hat{\varepsilon}_0)$$

Une fois la prévision effectuée, on va chercher à mesurer les performances du modèle de régression en termes de prévisions. Pour cela, on utilise des prévisions

faites pour des observations que l'on observe réellement, afin de pouvoir juger de la qualité de la prévision effectuée (de pouvoir calculer les erreurs de prévision). On estime le modèle sur un certain nombre de points et on prévoit pour les observations suivantes. On répète cette procédure en rajoutant une observation à l'estimation. On compare ensuite les valeurs prévues aux vraies valeurs observées.

A partir des prévisions effectuées, on calcule des critères permettant de mesurer la qualité des prévisions effectuées (ces critères sont à minimiser). Les plus connus sont, d'une part, la racine carré de la moyenne des erreurs de prévision au carré (*Root Mean Square Error*) :

$$RMSE = \sqrt{(1/k)\sum_i (y_i - \hat{y}_i)^2}$$

et, d'autre part, la moyenne des erreurs de prévision en valeur absolue (*Mean Absolute Error*) :

$$MAE = (1/k)\sum_i |y_i - \hat{y}_i|$$

où k est le nombre de prévisions construites.

On choisit alors l'ensemble de variables explicatives qui conduit à minimiser les critères RMSE et MAE. Ces critères ne pondèrent pas de la même manière les erreurs de prévision. Le RMSE pondère davantage les erreurs au carré, alors que le MAE n'y donne pas plus de poids.

7.3 Tests sur les erreurs

Nous avons supposé, pour les erreurs du modèle $y = X\beta + \varepsilon$, que :

$$\begin{array}{ll} E(\varepsilon_i) = 0 & \\ var(\varepsilon_i) = \sigma^2 & \forall i \\ cov(\varepsilon_i, \varepsilon_j) = 0 & i \neq j \\ \varepsilon_i \sim \mathcal{N}(0, \sigma^2) & \end{array}$$

Sur la base des erreurs estimées, il faut vérifier si ces hypothèses peuvent être acceptées. Nous présentons les tests de l'hypothèse de variance constante, appelée homoscédasticité, et d'absence d'autocorrélation dans le chapitre suivant.

Concernant l'hypothèse que chaque erreur doit être de moyenne nulle, nous avons vu que si la constante est présente dans le modèle, alors les erreurs estimées sont par définition de moyenne nulle.

En revanche, si la constante n'a pas été mise dans le modèle, il faut tester que les erreurs estimées sont de moyenne nulle. Pour cela, on fait un test de student en utilisant la moyenne empirique des erreurs estimées :

$$\hat{m} = \frac{\sum_i \hat{\varepsilon}_i}{n}$$

Le test de $H_0 \ : \ m = 0$ contre $H_1 \ : \ m \neq 0$ est basé sur la statistique de student :

$$t = \frac{\hat{m}}{\sqrt{\sigma/n}}$$

Elle suit une loi de Student sous H_0.

Concernant l'hypothèse de normalité, il s'agit de tester que les erreurs estimées $\hat{\varepsilon}_i$ suivent une loi normale, c'est–à–dire ne présentent pas d'asymétrie (*Skewness*) ni d'applatissement (*kurtosis*).

Le coefficient de Skewness est donné par :

$$\beta_1^{1/2} = \frac{\mu_3}{\mu_2^{3/2}}$$

et le coefficient de kurtosis est donné par :

$$\beta_2 = \frac{\mu_4}{\mu_2^2}$$

où

$$\mu_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$$

est le moment centré d'ordre k de la variable x.

Si la distribution est normale et le nombre d'observations grand, alors :

$$\beta_1^{1/2} \sim \mathcal{N}\left(0, \sqrt{6/n}\right) \qquad \beta_2 \sim \mathcal{N}\left(3, \sqrt{24/n}\right)$$

On construit alors les statistiques :

$$\nu_1 = \frac{\beta_1^{1/2}}{\sqrt{6/n}} \text{ et } \qquad \nu_2 = \frac{\beta_2 - 3}{\sqrt{24/n}}$$

qui suive chacune une $\mathcal{N}(0,1)$.

Le test de Jarque Bera permet de tester simultanément l'absence d'asymétrie et l'absence d'applatissement. La statistique de test est donnée par :

$$JB = \frac{n}{6}\beta_1 + \frac{n}{24}(\beta_2 - 3)^2$$

Cette statistique suit, sous l'hypothèse nulle de normalité, une loi du χ_2^2.

Si l'hypothèse de normalité est rejetée, cela pose problème pour les tests présentés précédemment. Si le problème est lié à la présence de kurtosis, cela peut provenir d'une variance non constante (on peut prendre le logarithme de la série ou étudier plus précisément l'hétéroscédasticité) ou de la présence d'observations aberrantes. Dans ce dernier cas, il vaut mieux réestimer la régression en introduisant une variable indicatrice pour l'observation aberrante afin de rendre les résidus normaux.

Chapitre 4 :
Modèle de régression généralisé et MCG

1 Présentation du problème

Nous avons fait l'hypothèse que les erreurs du modèle $y = X\beta + \varepsilon$ étaient non corrélées ($cov(\varepsilon_i, \varepsilon_j) = 0 \quad i \neq j$) et de même variance ($var(\varepsilon_i) = \sigma^2 \ \forall i$). Or, la plupart des variables économiques présentent de l'autocorrélation (c'est–à–dire une dépendance temporelle qui implique que $cov(\varepsilon_t, \varepsilon_{t-1}) > 0$) et l'étude d'observations individuelles conduit souvent à de l'hétéroscédasticité (différence de variances entre les groupes). Dans ces cas, l'hypothèse $V(\varepsilon) = \sigma^2 I$ n'est plus vérifiée.

Supposons que notre modèle soit toujours $y = X\beta + \varepsilon$ avec $E(\varepsilon) = 0$ mais :

$$V(\varepsilon) = \sigma^2 \Omega$$

où Ω est une matrice définie positive (différente de la matrice identité), dont la forme dépend du problème rencontré :

- En présence d'hétéroscédasticité, quand les erreurs ne sont pas de même variance : $var(\varepsilon_i) \neq var(\varepsilon_j) \quad i \neq j$, la matrice de variance-covariance s'écrit :

$$\sigma^2\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ & & \vdots & \\ 0 & 0 & \dots & \sigma_n^2 \end{pmatrix}$$

- En présence d'autocorrélation, plutôt présente en séries temporelles pour des variables économiques, quand celles-ci présentent une certaine mémoire, donc quand les réalisations d'une période dépendent de ce qui s'est passé avant. Si le modèle est temporel, par exemple $y_t = X_t'\beta + \varepsilon_t$, on peut avoir $var(\varepsilon_t) = \sigma^2 \ \forall t$ mais $cov(\varepsilon_t, \varepsilon_{t-1}) = \rho_1\sigma^2 \neq 0$ ou de manière générale $cov(\varepsilon_t, \varepsilon_{t-j}) = \rho_j\sigma^2 \neq 0$. On a alors :

$$\sigma^2\Omega = \sigma^2 \begin{pmatrix} 1 & \rho_1 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \dots & \rho_{n-2} \\ & & \vdots & \\ \rho_{n-1} & \rho_{n-2} & \dots & 1 \end{pmatrix}$$

Conséquences sur l'estimateur des MCO:

On montre que l'estimateur de β par MCO, donné par $\hat{\beta} = (X'X)^{-1}X'y$, reste sans biais, mais que la variance donnée par $\sigma^2(X'X)^{-1}$ n'est pas minimale.

La variance doit être définie différemment :

$$var(\hat{\beta}) = \sigma^2(X'X)^{-1}X'\Omega X(X'X)^{-1}$$

L'inférence (les tests) doit être basée sur cette variance et non pas sur celle calculée dans le cadre des MCO, soit $\sigma^2(X'X)^{-1}$.

Même si les MCO ne sont pas biaisés, il y a des cas où ils ne sont pas convergents (la variance de l'estimateur ne tend pas vers 0 quand n tend vers l'infini!). Il faut donc utiliser la bonne variance, mais elle dépend de Ω qui en général n'est pas connue. De plus, il faut déjà s'en rendre compte.

Nous allons alors étudier comment détecter la présence d'hétéroscédasticité ainsi que la présence d'autocorrélation, et nous présenterons les méthodes d'estimation appropriées dans les deux cas.

2 L'hétéroscédasticité

L'hétéroscédasticité correspond au cas où la variance (de y_i et donc de ε_i) n'est pas constante. Ainsi, $var(\varepsilon_i) = \sigma_i^2 \neq \sigma^2$, $i = 1, \ldots, n$.

Généralement, ce problème apparaît lors de l'étude de données individuelles (mais il peut cependant se présenter sur des données temporelles). Il existe plusieurs tests de détection de l'hétéroscédasticité, puisqu'il peut y avoir diverses raisons pour lesquelles la variance n'est pas constante.

2.1 Tests de détection

Il existe plusieurs tests selon le type d'hétéroscédasticité, mais malheureusement, on ne peut pas savoir a priori la forme de l'hétéroscédasticité. L'idée de tous les tests est que les MCO sont convergents même en présence d'hétéroscédasticité. Les résidus estimés par MCO vont donc nous permettre de détecter l'hétéroscédasticité. Ainsi, on commence par estimer le modèle par MCO, on recueille les résidus estimés, et les tests vont portés sur ces résidus estimés. Il s'agit de tester l'absence d'hétéroscédasticité (dite homoscédasticité), c'est-à-dire $H_0 \ : \ \sigma_i^2 = \sigma^2 \ \forall i$, où l'on utilisera $\sum_i \hat{\varepsilon}_i^2$ pour calculer l'estimateur de σ^2.

a) Test général de White

Il s'agit de tester :

$$H_0 : \ \sigma_i^2 = \sigma^2 \ \forall i$$

contre :

$$H_1 : \ \sigma_i^2 \neq \sigma^2$$

Le problème est qu'il faudrait estimer n paramètres : les n variances σ_i^2, $i = 1, \ldots, n$ avec n observations, ce qui est impossible.

White propose un test qui consiste à régresser $\hat{\varepsilon}_i^2$ sur les variables explicatives du modèle ainsi que sur les produits croisés entre ces variables explicatives, et de tester la significativité de ces variables explicatives. Si certaines sont significatives, c'est que la variance n'est pas constante.

La statistique est égale à nR^2 où R^2 est le coefficient de détermination dans cette régression. La statistique est distribuée comme un χ^2_{p-1}, où p est le nombre de régresseurs dans cette régression y compris la constante.

Ce test est très général, on n'a pas besoin de faire une hypothèse sur la forme de l'hétéroscédasticité. C'est à la fois un avantage et un inconvénient, puisque le test va trouver de l'hétéroscédasticité alors que ça peut être un problème de spécification

(comme l'oubli d'une variable au carré dans la régression). De plus, si on rejette l'hypothèse d'homoscédasticité, comme aucune forme particulière n'est testée, le test ne nous permet pas de savoir ce qu'il faut faire!

b) Le test de Goldfeld-Quandt

On suppose que les observations peuvent être divisées en 2 groupes de telle sorte que sous l'hypothèse d'homoscédasticité, les variances des erreurs seraient les mêmes dans les 2 groupes, alors que sous l'hypothèse alternative, elles diffèrent.

Ce test est particulièrement adapté aux observations qui se définissent par groupe (on dispose d'entreprises pour plusieurs secteurs) ou dans les modèles où $\sigma_i^2 = \sigma^2 x_i^2$ pour certaines variables x. En rangeant les observations selon ces x, on peut séparer les observations de forte et de faible variance.

Le test est donc basé sur une division de l'échantillon en 2 groupes de n_1 et n_2 observations. Pour obtenir des estimateurs de la variance statistiquement indépendants, on fait les régressions séparément.

L'hypothèse nulle est alors :

$$H_0 : \ \sigma_1^2 = \sigma_2^2$$

contre

$$H_1 : \ \sigma_1^2 > \sigma_2^2$$

La statistique de test est :

$$F = \frac{\hat{\varepsilon}_1' \hat{\varepsilon}_1 / (n_1 - K - 1)}{\hat{\varepsilon}_2' \hat{\varepsilon}_2 / (n_2 - K - 1)}$$

Sous l'hypothèse nulle, cette statistique suit une Fisher à $n_1 - K - 1$ et $n_2 - K - 1$ degrés de liberté. (Il est préférable de considérer une région de rejet unilatéral mais dans ce cas, il faut bien faire attention de prendre pour σ_1^2 la plus grande variance!)

Pour augmenter la puissance de ce test, Goldfeld et Quandt suggèrent qu'un certain nombre d'observations au milieu de l'échantillon soient exclues. Attention, plus on enlève d'observations, moins il reste de degrés de liberté pour estimer chaque variance, ce qui diminuera la puissance du test. Certains disent qu'il ne faut pas enlever plus du tiers des observations.

c) Le test de Breusch-Pagan

Par rapport au test de White, l'hypothèse nulle est définie de manière plus préciser, c'est-à-dire que l'on identifie une ou plusieurs variables qui pourraient être responsables de l'hétéroscédasticité.

$$H_0 : \ \alpha = 0 \text{ contre } H_1 : \ \sigma_i^2 = \sigma^2 f(\alpha_0 + \alpha' z_i)$$

où z_i est un ensemble de variables explicatives.

La statistique est alors :

$$LM = \frac{1}{2} SCE$$

où SCE est la somme des carrés expliquée de la régression de $\frac{\hat{\varepsilon}_i^2}{\hat{\varepsilon}'\hat{\varepsilon}/n}$ sur z_i.

Sous H_0, LM est distribuée selon un chi-2 avec comme degrés de liberté le nombre de variables dans z_i.

2.2 Correction de l'hétéroscédasticité

Si les tests conduisent à rejeter l'homoscédasticité, l'estimateur des MCO n'est pas efficace, la variance des estimateurs sera trop élevée.

Si le test de Goldfeld-Quandt conduit à rejeter l'homoscédasticité, mieux vaut alors faire l'estimation par groupe. Mais attention, on ne pourra pas faire de test d'absence de rupture puisqu'on n'a pas égalité des variances.

Une autre possibilité est de faire des Moindres Carrés Pondérés. L'idée est de transformer le modèle initial en un modèle où les erreurs sont de même variance. Cela est simple si on connaît la forme de l'hétéroscédasticité.

Supposons par exemple que $var(\varepsilon_i) = \alpha x_{1i}$ avec le modèle $y_i = a + bx_{1i} + cx_{2i} + \varepsilon_i$. Le modèle transformé est obtenu en pondérant les variables par la racine de la variance, soit :

$$\frac{y_i}{\sqrt{x_{1i}}} = a\frac{1}{\sqrt{x_{1i}}} + b\frac{x_{1i}}{\sqrt{x_{1i}}} + c\frac{x_{2i}}{\sqrt{x_{1i}}} + \frac{\varepsilon_i}{\sqrt{x_{1i}}}$$

soit

$$\tilde{y}_i = a_1 + b_1\tilde{x}_{1i} + c_1\tilde{x}_{2i} + \tilde{\varepsilon}_i$$

avec $var(\tilde{\varepsilon}_i) = \alpha$ constante quel que soit i.

Nous pouvons alors appliqué les MCO dans ce modèle puisqu'il vérifie maintenant l'hypothèse d'homoscédasticité.

3 La présence d'autocorrélation

Quand on étudie des données temporelles (variables économiques par exemple), on s'attend à trouver une corrélation non nulle entre l'observation à une date t et les observations précédentes. En effet, ce qui se passe aujourd'hui dépend de ce qui s'est passé dans le passé, par exemple $corr(y_t, y_{t-1}) > 0$. Dans ce cas, on parle d'autocorrélation et les MCO ne sont plus efficaces.

3.1 Test de détection

Il y autocorrélation quand la covariance des erreurs est non nulle. Sur la base du modèle estimé par MCO, on peut calculer la covariance des résidus estimés.

a) Test de Durbin Watson

Ce test est basé sur les résidus estimés de la régression $y_t = X_t'\beta + \varepsilon_t$ par MCO, notés $\hat{\varepsilon}_t$. L'idée est de tester qu'ils ne sont pas autocorrélés, c'est-à-dire que leur covariance est nulle (ce sont des bruits blancs) contre l'hypothèse qu'ils sont autocorrélés à l'ordre 1 (on dit qu'ils suivent un modèle AutoRégressif d'ordre 1, noté AR(1)).

On teste alors :

$$H_0 : \ \rho = 0 \text{ contre } H_1 : \ \hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + \nu_t$$

La statistique est la statique de Durbin-Watson :

$$DW = \frac{\sum_{t=2}^{n}(\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^{n}\hat{\varepsilon}_t^2}$$

Il est facile de montrer qu'elle peut se réécrire comme :

$$DW = 2(1 - \hat{\rho})$$

où $\hat{\rho}$ est le coefficient ρ estimé par MCO dans la régression $\hat{\varepsilon}_t = \rho\hat{\varepsilon}_{t-1} + \nu_t$.

Ainsi, sous l'hypothèse d'absence d'autocorrélation, $\rho = 0$, la statistique de DW est égale à 2. Sinon, elle s'éloigne de 2 : si elle est inférieure à 2 (proche de 0), il y a de l'autocorrélation positive alors que si elle est supérieure à 2 (en étant inférieure à 4), il y a de l'autocorrélation négative.

Il existe des tables afin de savoir si on doit rejeter ou pas l'hypothèse nulle d'absence d'autocorrélation. Ces tables dépendent du nombre de variables explicatives présentes dans le modèle en dehors de la constante et du nombre d'observations.

Ce test ne permet de tester l'absence d'autocorrélation uniquement contre une altenative de type AR(1).

b) Le test de Breusch Godfrey

Il s'agit d'étendre le test précédent à de l'autocorrélation dont la structure est plus complexe qu'un AR(1), soit par exemple la présence d'autocorrélation à un ordre plus élevé que l'ordre 1, on parle d'AR(p).

On teste alors :

$$H_0 : \text{ pas d'autocorrélation contre } H_1 : \ \hat{\varepsilon}_t \sim AR(p)$$

Pour mener ce test, on régresse $\hat{\varepsilon}_t$ sur $x_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-p}$ (en mettant des 0 pour les valeurs manquantes des résidus retardés) et on teste la nullité des retards sur $\hat{\varepsilon}_t$.

La statistique est nR^2, où R^2 est le coefficient de détermination de la régression de $\hat{\varepsilon}_t$ sur $x_t, \hat{\varepsilon}_{t-1}, \ldots, \hat{\varepsilon}_{t-p}$. La statistique suit un chi-2 à p degrés de liberté.

c) Les tests du portemanteau

Il s'agit de tests *fourre-tout* qui permettent de tester la nullité jointe des p premières autocorrélations. Si on accepte l'hypothèse nulle de nullité jointe des p premières autocorrélations des résidus estimés, alors on peut supposer que les résidus sont des bruits blancs, autrement dit, les MCO sont valides. Sinon, ce test ne nous dit rien sur la structure d'autocorrélation.

Le Q de Box Pierce est donné par :

$$Q = n \sum_{j=1}^{p} r_j^2 \sim \chi_p^2$$

où

$$r_j = \frac{\sum_{t=j+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-j}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

Ljung et Box proposent une correction :

$$Q' = n(n+2) \sum_{j} \frac{r_j^2}{n-j} \sim \chi_p^2$$

d) Le h de Durbin en présence d'endogènes retardées

Si parmi les régresseurs, il y a des endogènes retardées y_{t-j}, alors la statistique de DW n'est plus valide et on doit utiliser le h de Durbin :

$$h = \left(1 - \frac{DW}{2}\right)\sqrt{\frac{n}{1 - n\hat{\sigma}^2_{\hat{\beta}(y_{t-1})}}}$$

Cette statistique suit sous H_0 une loi normale centrée réduite.

3.2 Correction de l'autocorrélation

Si on trouve que les résidus $\hat{\varepsilon}_t$ du modèle $y_t = X_t'\beta + \varepsilon_t$ suivent un processus AR(p), c'est que l'on a oublié des variables qui dépendent du passé dans le modèle. Ainsi, on peut ajouter des endogènes retardées $y_{t-1}, \ldots, y_{t-p}$ parmi les variables explicatives, ainsi que les variables explicatives retardées $x_{j,t-1}, \ldots, x_{j,t-p}$. On dit qu'on a blanchit les résidus. En effet, les résidus de cette nouvelle régression ne seront plus caractérisés par la présence d'autocorrélation.

C'est à peu près ce que font les logiciels quand ils utilisent la méthode des Moindres Carrés Généralisés (MCG).

Quand on étudie des variables observées dans le temps, il arrive très souvent qu'elles aient une allure croissante, et parfois (quand elles sont observées à une périodicité intra-annuelle) qu'elles fassent apparaître un effet saisonnier, effet qui revient toutes les saisons de manière assez régulière.

Si on veut étudier la liaison entre plusieurs variables temporelles caractérisées par une allure croissante et de la saisonnalité, il faut retirer l'allure croissante des séries et les effets saisonniers des variables afin de ne pas trouver de corrélation fallacieuse (deux variables qui croissent pour des raisons complètement différentes ou qui ont toutes deux un effet saisonnonnier seront forcément corrélées même si les causes ne sont pas les mêmes!).

Afin d'enlever l'allure croissante des séries, on peut soit régresser les séries sur une tendance linéaire et conserver le résidu (la croissance provient d'une cause exogène) soit prendre le taux de croissance des séries. Le choix entre l'une ou l'autre transformation n'est pas neutre statistiquement et économiquement, mais l'exposé ces conséquences de tel ou tel choix sort largement du cadre de ce cours. On retiendra qu'il faut rendre stationnaire (non croissant ou non décroissant) les séries avant de les étudier.

Afin d'enlever les effets saisonniers, une procédure facile à mettre en oeuvre est de régresser la série sur des variables indicatrices saisonnières, d'estimer les coefficients et de conserver les résidus de cette régression qui correspondront à la série désaisonnalisée.