



Econometrie I et II sur Python

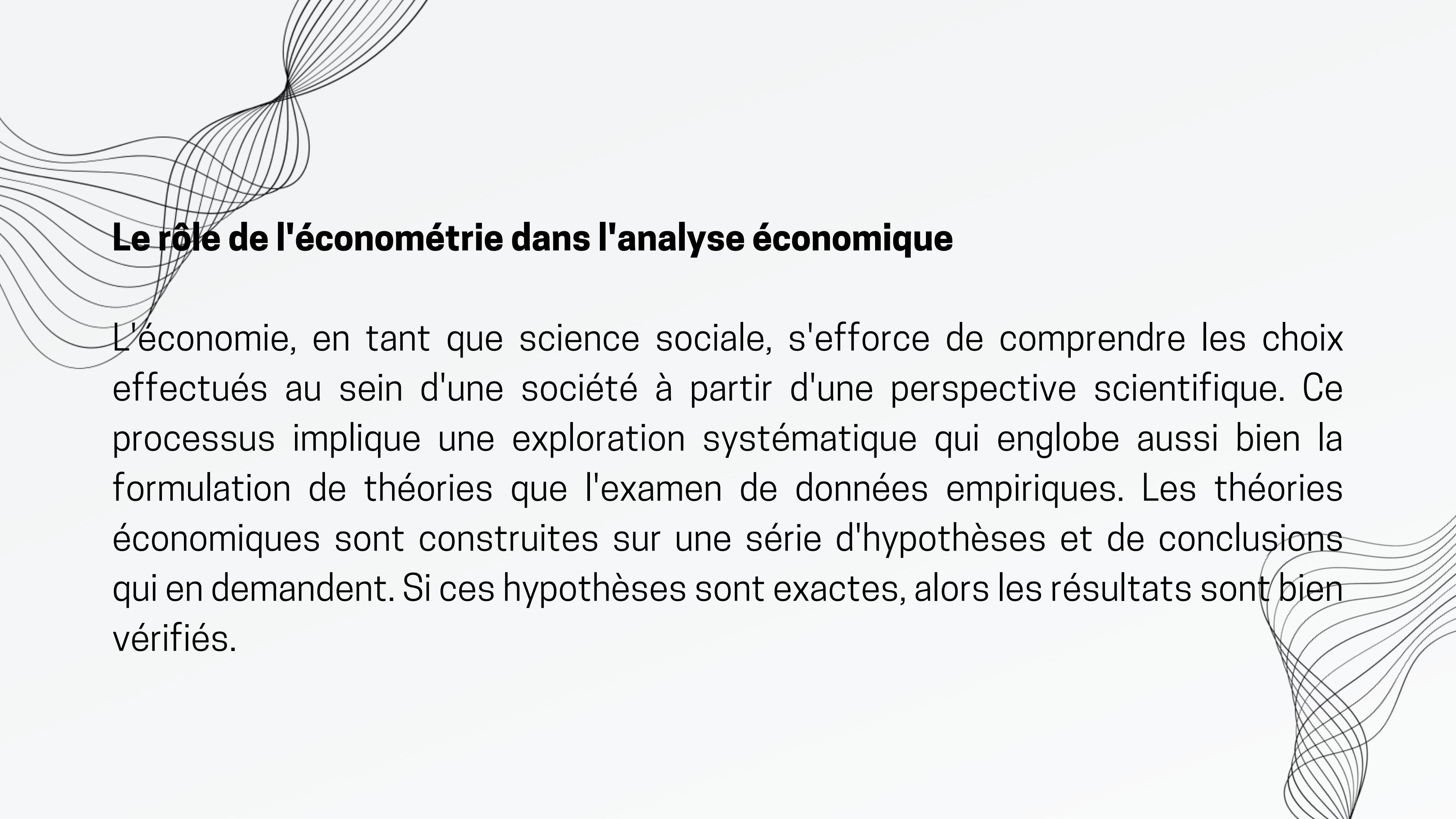
M. Oely Hasina

Datascientist - Developpeur Full stack

Plan du cours

- Partie 1: Introduction à l'économétrie
- Partie 2: Régression linéaire simple
- Partie 3: Régression linéaire multiple
- Partie 4: Tests d'hypothèses et diagnostics de modèle
- Partie 5: Régression non linéaire
- Partie 6: Modèles à variables instrumentales et méthode des moments généralisés (GMM)
- Partie 7: Séries temporelles

Partie 1: Introduction à l'économétrie



Le rôle de l'économétrie dans l'analyse économique

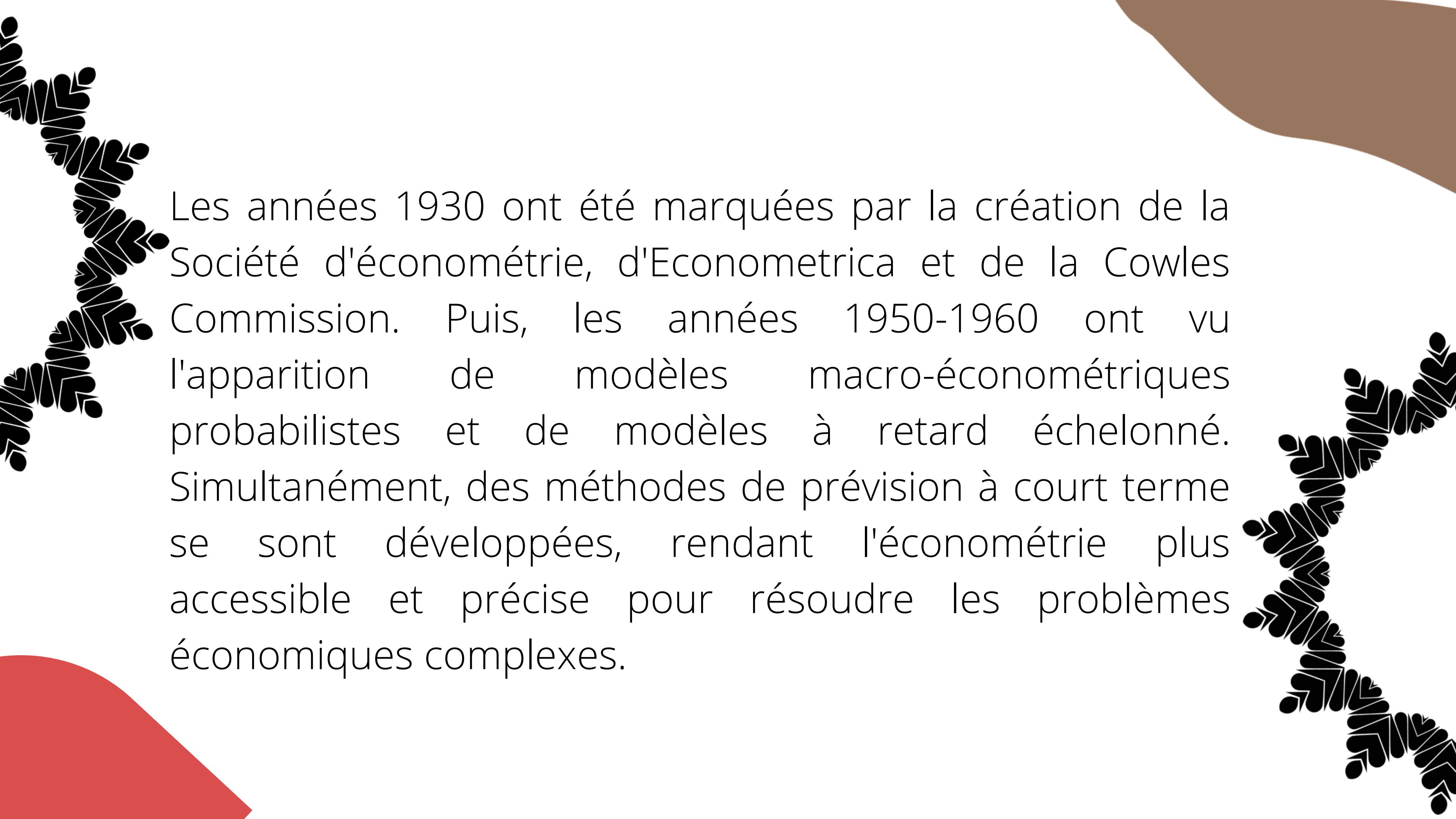
L'économie, en tant que science sociale, s'efforce de comprendre les choix effectués au sein d'une société à partir d'une perspective scientifique. Ce processus implique une exploration systématique qui englobe aussi bien la formulation de théories que l'examen de données empiriques. Les théories économiques sont construites sur une série d'hypothèses et de conclusions qui en demandent. Si ces hypothèses sont exactes, alors les résultats sont bien vérifiés.

Le lien entre la théorie économique et l'économétrie

Comme mentionné précédemment, l'économie est une science sociale dont l'objet d'étude est la société et le comportement de ses constituants, c'est-à-dire les institutions et les ménages. La théorie économique vise à éclaircir les relations entre les variables économiques et utilise ces informations dans un cadre théorique global pour expliquer comment les ressources sont autorisées, comment la production est décidée et comment se fait la répartition au sein d'un système activé dans un contexte de rareté.

Origine:

L'économétrie, née entre les XVIIème et XVIIIème siècles, s'est développée au XIXème siècle avec l'application des mathématiques à l'économie. Au XXème siècle, des chercheurs comme Moore, Tinbergen et Frisch ont consolidé l'économétrie comme une discipline indépendante.



Les années 1930 ont été marquées par la création de la Société d'économétrie, d'Econometrica et de la Cowles Commission. Puis, les années 1950-1960 ont vu l'apparition de modèles macro-économétriques probabilistes et de modèles à retard échelonné. Simultanément, des méthodes de prévision à court terme se sont développées, rendant l'économétrie plus accessible et précise pour résoudre les problèmes économiques complexes.

Définition

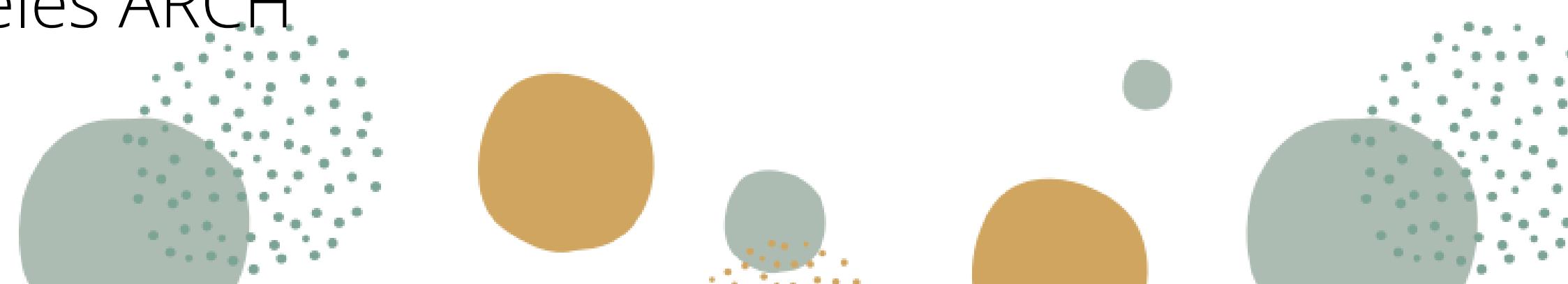
Qu'est-ce que l'économétrie ?

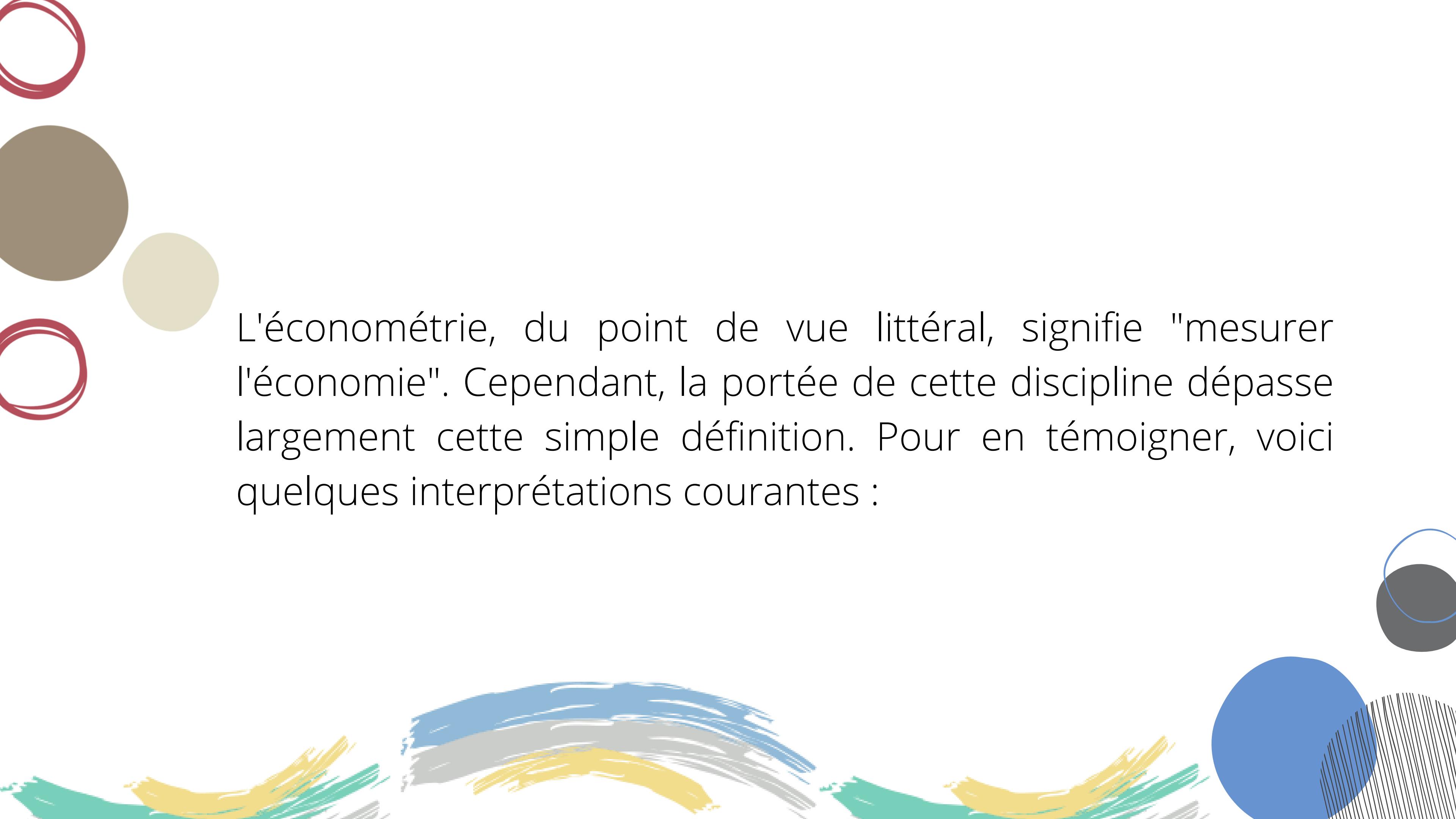
« L'économétrie, c'est l'unification de la théorie économique, des mathématiques et des statistiques. L'économétrie n'est pas assimilable uniquement à la statistique économique ou aux méthodes mathématiques appliquées à l'économie : c'est la conjonction de la théorie économique, de la statistique et des mathématiques » Définition de Ragnar FRISCH, 1933



L'économétrie a été mise à l'honneur par plusieurs prix Nobel

- 1°) 1980 : Lawrence KLEIN (Premiers modèles macro-économétriques, analyse des fluctuations et croissance)
- 2°) 1989 : Trygve HAAVELMO (Introduction de l'approche probabiliste, modèles à équations simultanées)
- 3°) 2000 : James HECKMAN (Théories et méthodes d'analyse des échantillons)
- 4°) 2000 : Daniel McFADDEN (Analyse des choix discrets en économétrie)
- 5°) 2003 : Robert GRANGER & Clive ENGLE (Co-intégration & Volatilité des séries temporelles et des modèles ARCH)





L'économétrie, du point de vue littéral, signifie "mesurer l'économie". Cependant, la portée de cette discipline dépasse largement cette simple définition. Pour en témoigner, voici quelques interprétations courantes :

- **L'économétrie applique les mathématiques et les statistiques aux données économiques** afin de donner une assise empirique aux modèles créés par l'économie mathématique, tout en fournissant des résultats quantifiables.
- **L'économétrie peut être constituée comme l'analyse quantitative des phénomènes économiques actuels**, basée sur le développement parallèle de la théorie et de l'observation, reliée par des méthodes de déduction applicables.
- **L'économétrie peut aussi être définie comme l'utilisation de méthodes statistiques et mathématiques pour l'analyse des données économiques**, dans le but de donner un contenu empirique aux théories économiques, et de les vérifier ou de les infirmer.

Objectifs de l'économétrie:

L'économétrie, en tant que branche de l'économie, vise principalement à atteindre trois objectifs, à savoir l'analyse, l'aide à la décision et la prévision. Chacun de ces objectifs est essentiel pour comprendre et résoudre les problèmes économiques complexes :

Analyse : L'économétrie utilise des données économiques réelles et des statistiques techniques pour tester et vérifier les théories économiques, permettant ainsi de comprendre si ces théories se manifestent dans les faits.

Aide à la décision : En fournissant des estimations numériques des relations économiques, l'économétrie guide les décideurs dans leur processus de prise de décision, en les aidant à baser leurs choix sur des données empiriques plutôt que sur des suppositions.

Prévision : Grâce à l'utilisation de modèles économétriques, l'économétrie permet de prédire les valeurs futures des variables économiques telles que la croissance du PIB, l'inflation, le chômage, etc., ce qui est précieux pour la planification économique et les politiques publiques.

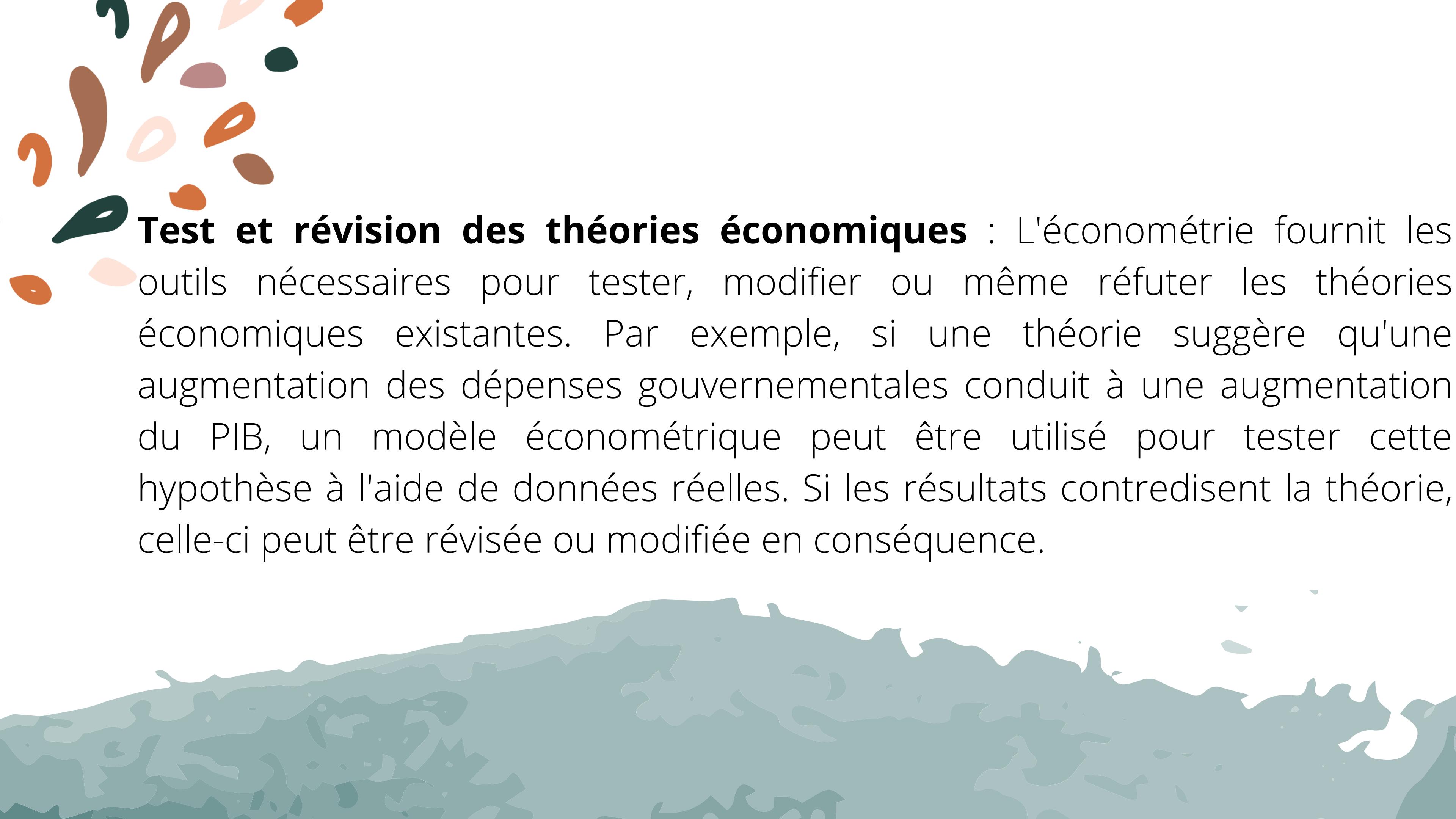
L'utilité de l'analyse économétrique

L'économétrie est un outil fondamental pour l'analyse économique, en particulier lorsque les acteurs économiques sont confrontés à des choix et doivent prendre les meilleures décisions possibles. Son utilité peut être expliquée de manière plus approfondie en se concentrant sur les aspects suivants :



Compréhension des relations économiques : L'économétrie offre un cadre pour comprendre la nature des relations économiques, ce qui est crucial pour prendre des décisions informées. Par exemple, elle peut nous aider à comprendre comment une variation du taux d'intérêt peut influencer l'économie, ou comment une augmentation du salaire minimum peut affecter le taux de chômage.

Quantification des relations économiques : En plus de comprendre le sens d'une relation économique, l'économétrie permet d'en quantifier l'ampleur. Par exemple, elle peut nous aider à déterminer l'impact précis d'une augmentation d'un dollar du salaire minimum sur le taux de chômage. Cette capacité à quantifier les relations économiques est vitale pour prendre des décisions stratégiques stratégiques.



Test et révision des théories économiques : L'économétrie fournit les outils nécessaires pour tester, modifier ou même réfuter les théories économiques existantes. Par exemple, si une théorie suggère qu'une augmentation des dépenses gouvernementales conduit à une augmentation du PIB, un modèle économétrique peut être utilisé pour tester cette hypothèse à l'aide de données réelles. Si les résultats contredisent la théorie, celle-ci peut être révisée ou modifiée en conséquence.

Interprétation des coefficients économiques : Les coefficients économiques donnent une indication de l'importance relative de différentes variables dans une relation économique. L'économétrie fournit des méthodes pour estimer ces coefficients de manière fiable et pour les interpréter correctement. Par exemple, dans un modèle qui décrit la relation entre le revenu et la consommation, le coefficient du revenu nous indique combien la consommation change en moyenne lorsque le revenu change d'une unité.

Les étapes de la méthode économétrique.

L'économétrie suit un processus bien structuré pour atteindre ces objectifs, et ce processus est souvent appelé la méthode économétrique. Elle implique la formulation d'un modèle basé sur une théorie économique, la spécification de ce modèle en une forme testable, l'estimation des paramètres du modèle, et enfin, l'interprétation des résultats pour tirer des conclusions significatives.

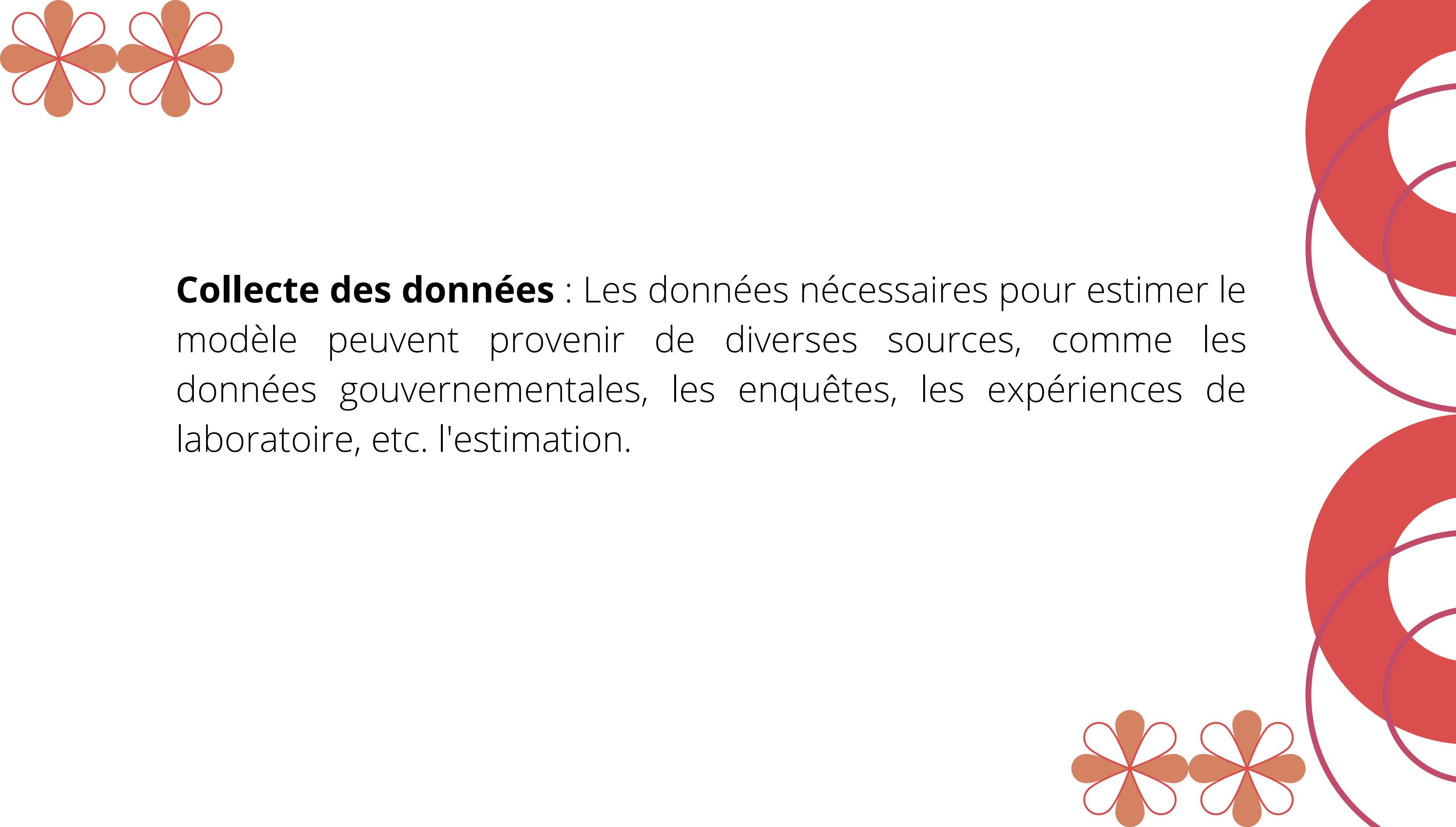
Formulation du modèle économique : lors de cette étape, le but est de formuler un modèle économique qui décrit les relations entre différentes variables économiques. Ce modèle est généralement basé sur une théorie économique existante. Par exemple, on pourrait supposer que le revenu d'un individu dépend de son niveau d'éducation et de son expérience de travail. C'est ici que l'économiste formule ses hypothèses fondées sur la théorie économique, l'intuition et l'expérience passée.



Spécification du modèle économétrique : Une fois le modèle économique formulé, il doit être transformé en un modèle économétrique. Ce modèle est une version quantitative du modèle économique qui peut être testée avec des données réelles. Cela implique de transformer les relations qualitatives du modèle économique en équations mathématiques.



Choix de la méthode d'estimation : Il existe plusieurs méthodes pour estimer les coefficients de votre modèle, comme la méthode des moindres carrés ordinaires (MCO), la méthode du maximum de vraisemblance (MMV), ou encore la méthode des moments généralisés (MMG). Le choix de la méthode dépend du type de données disponibles, des hypothèses faites sur les erreurs et d'autres facteurs.



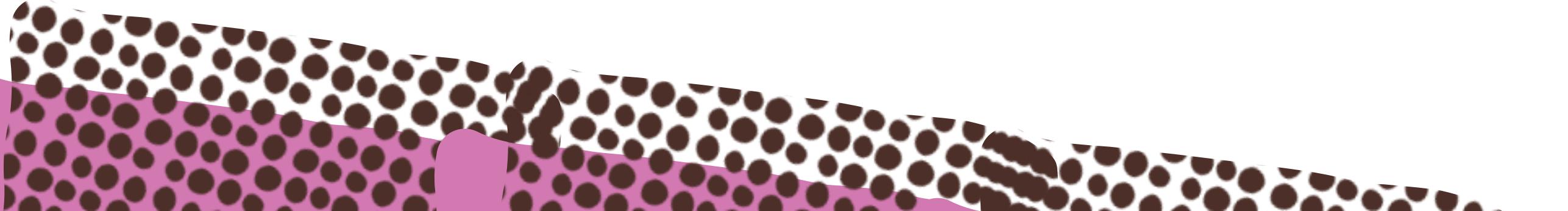
Collecte des données : Les données nécessaires pour estimer le modèle peuvent provenir de diverses sources, comme les données gouvernementales, les enquêtes, les expériences de laboratoire, etc. l'estimation.

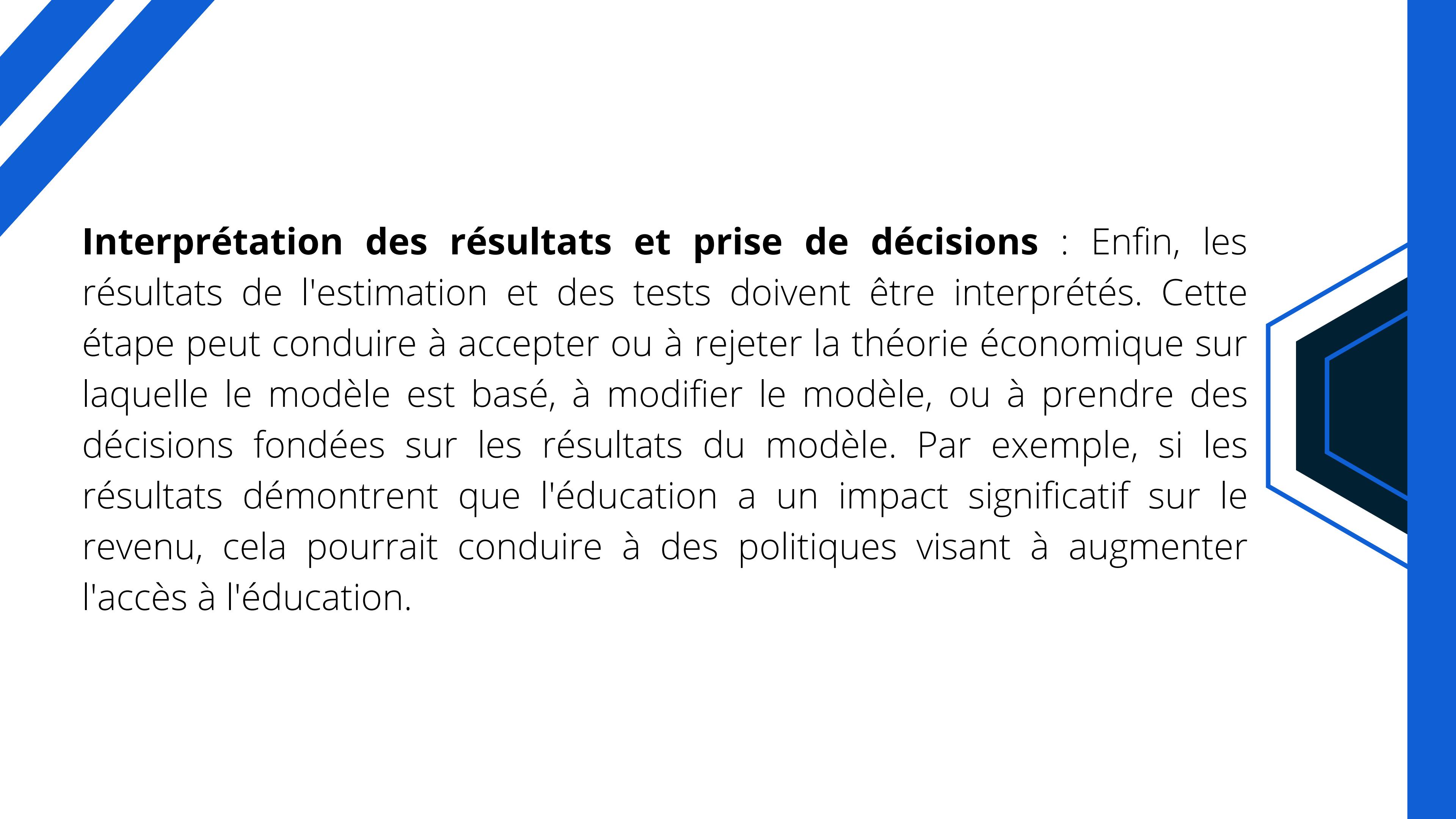


Estimation du modèle : Après avoir collecté les données, le modèle économétrique est affiché. Cela signifie que les coefficients du modèle, qui dégradent la relation entre les variables, sont calculés à partir des données. Cette étape permet d'obtenir une version numérique du modèle qui peut être utilisée pour faire des prédictions ou tester des hypothèses.



Test du modèle : Une fois le modèle produit, il est important de le tester pour s'assurer qu'il est valide. Il existe de nombreux tests que vous pouvez effectuer, comme les tests de spécification pour vérifier que le modèle a la bonne forme, les tests d'hypothèses pour vérifier si les coefficients sont significativement différents de zéro, et les tests d'ajustement pour vérifier si le modèle s'ajuste bien aux données.





Interprétation des résultats et prise de décisions : Enfin, les résultats de l'estimation et des tests doivent être interprétés. Cette étape peut conduire à accepter ou à rejeter la théorie économique sur laquelle le modèle est basé, à modifier le modèle, ou à prendre des décisions fondées sur les résultats du modèle. Par exemple, si les résultats démontrent que l'éducation a un impact significatif sur le revenu, cela pourrait conduire à des politiques visant à augmenter l'accès à l'éducation.

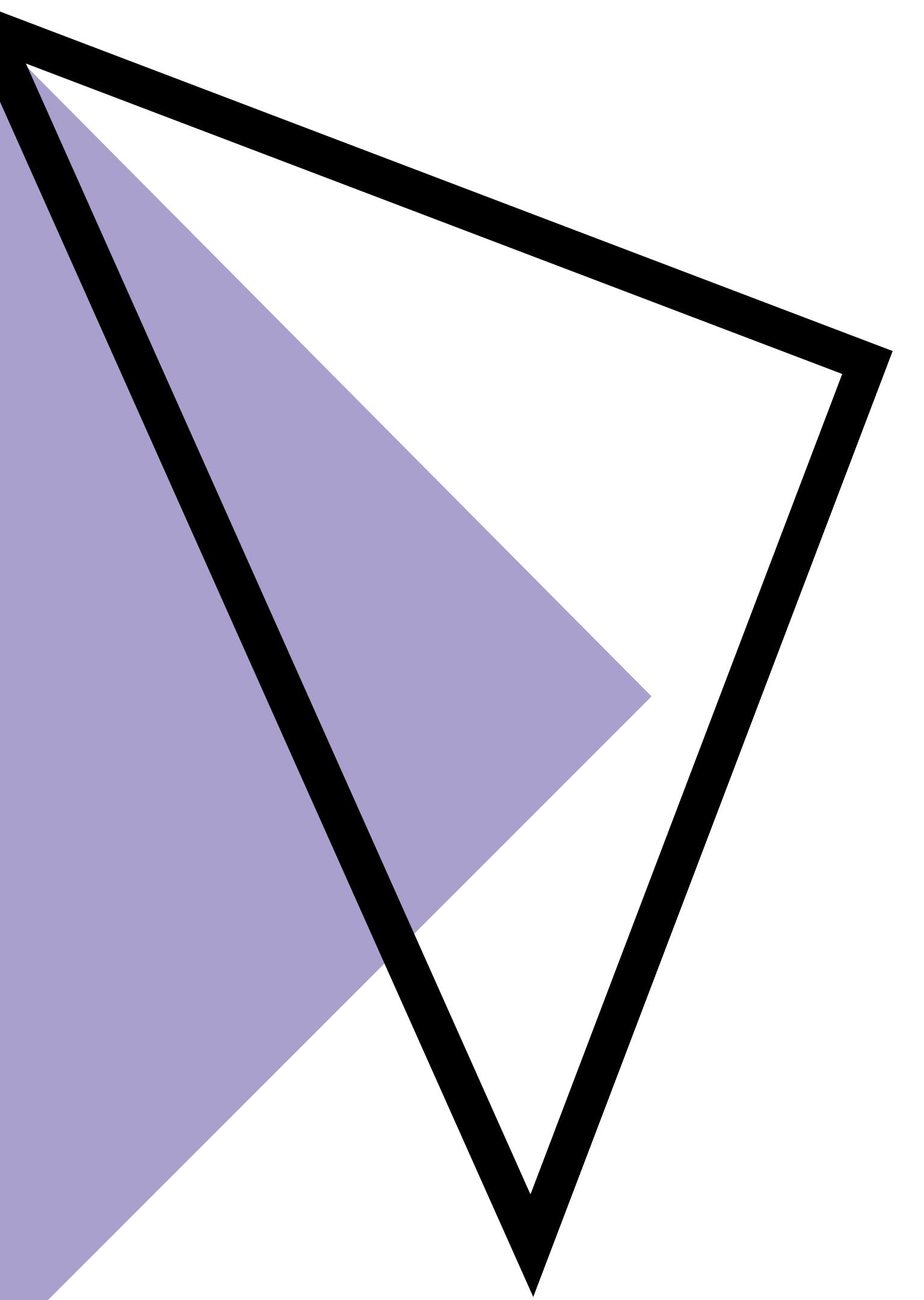


Les données en économétrie

L'économétrie utilise des données variées et profondes pour analyser, prévoir et expliquer les phénomènes économiques. Comprendre et manipuler ces données sont des étapes essentielles pour développer des modèles économiques précis et significatifs.

Types de données utilisées en économie :

- **Séries temporelles** : Ces données sont reportées sur une même entité au fil du temps. Par exemple, le PIB d'un pays collecté annuellement pendant 20 ans est un exemple de série temporelle.
- **Données transversales** : elles représentent les observations recueillies sur plusieurs entités à un instant précis dans le temps. Par exemple, le revenu de différents ménages dans une région donnée collecté au cours d'une année spécifique est un exemple de données transversales.
- **Données de panel** : Aussi connues sous le nom de données longitudinales, elles combinent les séries temporelles et les données transversales. Elles accompagnent plusieurs entités sur une période de temps. Par exemple, le revenu de différents ménages d'une région suivis sur plusieurs années serait une donnée de panel.



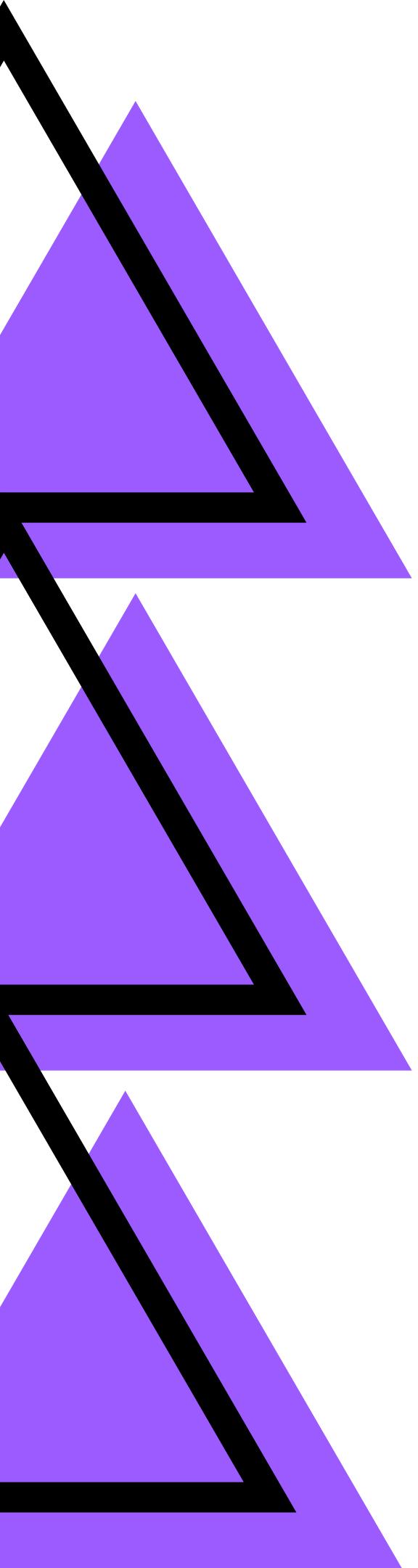
Collecte des données économiques :

La collecte des données est une tâche aussi délicate qu'essentielle en économie. La qualité de notre analyse économétrique dépend fortement de la qualité des données que nous utilisons. Les sources de nos données, leur fiabilité et les contraintes associées à leur collecte sont des facteurs que nous devons prendre en compte lorsque nous entreprenons cette tâche.

Sources :

Parmi les sources les plus courantes, on trouve les organismes publics comme l'Institut National de la Statistique et des Études Économiques (INSEE) en France, le Bureau of Labor Statistics (BLS) aux États-Unis, ou encore l'Office for National Statistics (ONS) au Royaume-Uni. Les bases de données d'entreprises privées, les enquêtes sur les ménages et les consommateurs, et les études académiques sont également des sources précieuses de données économiques.

A decorative element consisting of five horizontal bars at the bottom of the slide. From top to bottom, there are two thick black bars, one thin teal bar, one thin black bar, and one thick teal bar.



Fiabilité :

La fiabilité des données est cruciale pour toute analyse économique. Les chercheurs doivent évaluer la qualité des données, y compris leur exactitude, leur cohérence et leur exhaustivité. Il faut également tenir compte des biais potentiels qui pourraient influencer les résultats.



Contraintes :

La collecte de données économiques peut être confrontée à plusieurs contraintes. Par exemple, les données peuvent ne pas être disponibles pour certaines périodes ou régions. De plus, la collecte de données peut être coûteuse, surtout lorsqu'il s'agit d'enquêtes à grande échelle. Enfin, des problèmes de confidentialité peuvent se poser lorsqu'il s'agit de données sensibles.



Partie 2 : MODELE DE REGRESSION LINEAIRE SIMPLE

Introduction à l'économétrie et aux modèles de régression

"L'économétrie est une discipline qui combine économie, mathématiques et statistiques pour analyser les phénomènes économiques. Les modèles de régression, en particulier le modèle de régression linéaire simple, sont parmi les outils les plus utilisés en économétrie. L'idée est d'expliquer une variable (y) à l'aide d'une autre variable (x)."



La régression linéaire simple, fondamentale en économétrique, examine la relation entre deux variables : une indépendante (x) et une dépendante (y). Cet outil est souvent utilisé pour analyser et prédire les tendances, ou pour établir une relation causale entre ces variables dans un contexte économique.

la régression linéaire simple est utile pour la prévision. Une fois la relation entre les variables établie, elle peut être utilisée pour prédire les effets futurs des variations dans la variable indépendante.

Le modèle de régression linéaire simple - Définition

"Le modèle de régression linéaire simple est défini par l'équation : $y = a + \beta x + u$. Cette équation exprime y (la variable dépendante) comme une fonction linéaire de x (la variable indépendante), plus un terme d'erreur (u)."

L'économétrie s'appuie sur l'estimation des fonctions linéaires, telle que **$y = ax + b$** , pour examiner les théories économiques.

Par exemple, la théorie de Fisher qui établit une relation linéaire entre le taux d'intérêt et l'inflation, peut être représentée par

$r = i - \pi$, où r est le taux d'intérêt réel, i est le taux d'intérêt nominal et π est l'inflation. À l'opposé, des théories comme celle de la croissance basée sur la fonction de production Cobb-Douglas, qui se présente comme **$Y = K^\alpha L^\beta$** , impliquent une fonction multiplicative des variables K et L , et non une fonction linéaire.

Les fonctions multiplicatives peuvent également être évaluées par un modèle linéaire en appliquant une transformation logarithmique à la fonction pour obtenir une fonction linéaire :

$$Y = K \alpha L^\beta$$

$$\ln(Y) = \ln(K \alpha L^\beta)$$

$$\ln(Y) = \alpha \ln(K) + \beta \ln(L)$$

Ainsi, $\ln(Y)$ est une fonction linéaire de $\ln(K)$ et $\ln(L)$.

Les composants du modèle

"Dans notre équation, a est l'ordonnée à l'origine, représentant le niveau de y lorsque x est zéro. β est le coefficient de x , indiquant combien y change lorsque x change d'une unité. u est l'erreur, capturant tous les facteurs affectant y non inclus dans le modèle. Les suppositions clés de notre modèle sont : $E(u) = 0$ (l'espérance de u est nulle), et $Cov(x,u) = 0$ (x et u ne sont pas corrélés)."

Estimation des coefficients :

la méthode des moindres carrés ordinaires (OLS) "Pour estimer les coefficients a et β , nous utilisons la méthode des moindres carrés ordinaires (OLS), qui minimise la somme des carrés des résidus. Un résidu est la différence entre la valeur réelle de y et sa valeur prédictée. Les formules pour les estimations OLS sont : $\hat{\beta} = \text{Cov}(x, y)/\text{Var}(x)$ et $\hat{a} = \bar{y} - \hat{\beta}\bar{x}$."

Interprétation des coefficients

"Interpréter correctement les coefficients est crucial en économétrie. β représente l'effet marginal de x sur y : si x augmente d'une unité, y change de β unités, ceteris paribus (toutes choses étant égales par ailleurs). a est l'intercept, la valeur de y lorsque x est zéro."

Résumé de l'interprétation des coefficients

"Pour illustrer l'interprétation des coefficients, imaginons un modèle où y est le revenu et x est l'âge. Un β de 1000 signifie que chaque année supplémentaire d'âge augmente le revenu de 1000 unités, ceteris paribus.
 a pourrait représenter le revenu de départ à l'âge zéro."



Hypothèses du modèle

"Le modèle de régression linéaire simple repose sur plusieurs hypothèses, comme l'indépendance des erreurs, l'homoscédasticité (variance constante des erreurs), et la normalité des erreurs. Si ces hypothèses sont violées, nos estimations peuvent être incorrectes

Les hypothèses du modèle

H1) $E(u_i) = 0, \forall i \Rightarrow$ en moyenne le terme d'erreur est nul. Autrement dit, les variables spécifiées dans le modèle capturent bien y_i .

H2) x_i est une variable certaine (non stochastique)

H3) $V(x_i) = 1/N \sum_{i=1}^N (x_i - \bar{x})^2 \neq 0 \Rightarrow$ la variance de x est non nulle, i.e., les observations x_i ne prennent pas toutes la même valeur.

H4) $V(u_i) = E(u_i^2) = \sigma^2, \forall i \neq s \Rightarrow$ la variance est la même pour tout les u_i , on dit que les perturbations sont homoscédastiques.

$\text{Cov}(u_i, u_s) = E(u_i, u_s) = 0 \Rightarrow$ la perturbation u_i n'est pas influencée par la perturbation u_s . Dans le cas d'une série temporelle, cela signifie que la perturbation à une période n'est pas influencée par la perturbation à une autre période, i.e., un choc qui s'est produit à une période n'a pas d'influence sur ce qui se passe dans les périodes suivantes.

H5) $u_i \sim N(0, \sigma^2)$ \Rightarrow les erreurs sont indépendentes et identiquement distribuées selon la loi normale. Cette hypothèse de normalité est nécessaire pour réaliser des tests statistiques sur la base des distributions normale, de Student et de Fisher.

H6) $\text{Cov}(x_i, u_i) = 0$ \Rightarrow Cette hypothèse rend compte de l'indépendance entre la partie systématique et la partie aléatoire du modèle



Violations des hypothèses et leurs conséquences

"Il est crucial de comprendre les conséquences des violations des hypothèses. Par exemple, si les erreurs ne sont pas indépendantes (autocorrélation), nos erreurs standard peuvent être sous-estimées, ce qui rend les tests de signification non fiables. Si l'homoscédasticité n'est pas respectée (hétéroscédasticité), l'estimateur des moindres carrés ordinaires n'est plus le meilleur (BLUE), ce qui signifie qu'il existe un autre estimateur plus efficace. Si la normalité des erreurs n'est pas respectée, les tests de signification basés sur la distribution normale ne sont plus valides."



Les différentes méthodes d'estimation

Estimer un paramètre θ consiste à donner une valeur approchée à ce paramètre à partir d'un sondage de la population. Nous allons analyser différents types d'estimation, qui serviront de rappel de notions de statistiques simples.

Supposons que l'on cherche à estimer θ , représentant l'espérance de la population X : $\theta = E(X)$. Dans ce cas, nous pouvons utiliser la moyenne empirique comme estimateur de l'espérance : $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$.

Pour rappel, l'espérance d'une constante est égale à la constante de telle sorte que l'on peut écrire **$E(a) = a$** et **$E(aX) = aE(X)$** . Par ailleurs lorsque X et Y sont indépendantes on peut écrire **$E(XY) = E(X)E(Y)$**

De manière symétrique, si l'on cherche à estimer θ comme étant la variance d'une population \mathbf{X} : $\sigma^2 = V(\mathbf{X}) = E((\mathbf{X} - E(\mathbf{X}))^2)$.

Dans ce cas, nous pouvons utiliser l'estimateur de la variance qui se définit par : $V(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2$.

Pour rappel, $V(\mathbf{x}) = E(x^2) - E(x)^2$.

Preuve :

Si $V(\mathbf{x}) = E(x - E(x))^2$ avec $E(x) = \text{constante}$ (correspond à la moyenne), nous pouvons noter $E(x) = \mu$

$$V(\mathbf{x}) = E(x - E(x))^2$$

$$V(\mathbf{x}) = E(x^2 - 2xE(x) + E(x)^2)$$

$$V(\mathbf{x}) = E(x^2 - 2x\mu + \mu^2)$$

$$V(\mathbf{x}) = E(x^2) - 2\mu^2 + \mu^2$$

$$V(\mathbf{x}) = E(x^2) - E(x)^2$$

Le but d'un estimateur est d'être le plus précis possible pour avoir une erreur d'estimation la plus petite possible. Une mesure de précision est l'erreur quadratique moyenne (MSE) :

$$\mathbf{MSE} = \mathbf{E}(\hat{\mathbf{b}} - \mathbf{b})^2 = 0 \longleftrightarrow \mathbf{MSE} = \mathbf{E}(\hat{\mathbf{b}} - \mathbf{E}(\hat{\mathbf{b}}))^2$$

{z } Variance de l'estimateur + $\mathbf{E}(\mathbf{E}(\hat{\mathbf{b}}) - \mathbf{b})^2$ | {z } Biais de l'estimateur

L'erreur quadratique moyenne (EQM) mesure à la fois le biais et la précision de l'estimation. Un bon estimateur est à la fois peu biaisé et précis. Pour estimer un modèle linéaire, on peut utiliser différentes méthodes :

1. **Méthode des moments** : elle estime les moments de l'échantillon, tels que la moyenne et la variance, sans supposer de distribution spécifique des résidus.
2. **Maximum de vraisemblance** : cette méthode cherche à maximiser la probabilité d'observer les valeurs observées sachant les valeurs estimées, en se basant sur des hypothèses spécifiques sur la distribution des résidus.
3. **Méthode des moindres carrés (MCO)** : elle ajuste une droite au nuage de points en minimisant la somme des carrés des écarts entre les valeurs observées et les valeurs prédictes. Elle est couramment utilisée dans les modèles linéaires.
Ces méthodes permettent d'estimer les paramètres du modèle linéaire en fonction des objectifs d'analyse.

Dans le contexte de l'effet de l'expérience sur les salaires, un modèle non linéaire est utilisé pour capturer la variation des augmentations salariales avec l'expérience.

On estime la fonction **$Wi = a + cexp_i + bexp2i$** , qui prend en compte la théorie économique selon laquelle le rendement marginal de l'expérience diminue avec le temps. La dérivée de cette fonction,

$$\delta Wi / \delta exp = c + 2bexp$$

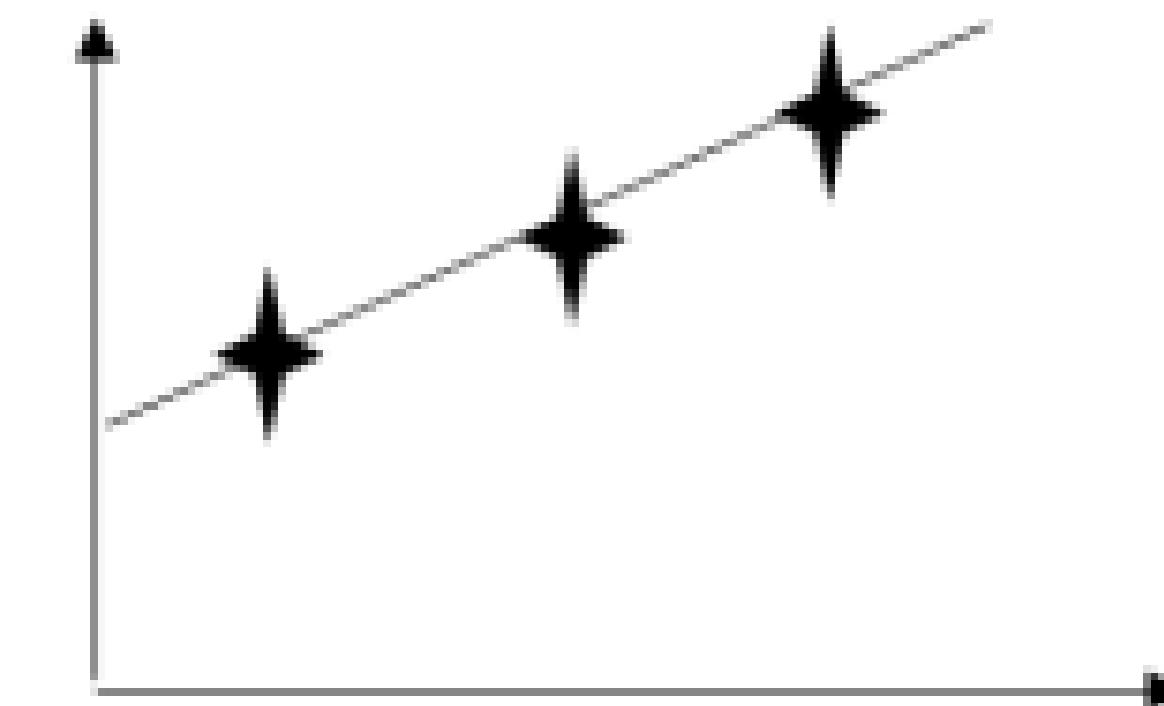
permet d'estimer l'impact d'une année supplémentaire d'expérience sur le salaire, en tenant compte des variations non linéaires.

Cette équation supplémentaire suggère que le gain salarial pour une année d'expérience est de 'c' pour une personne n'ayant aucune expérience. Pour une personne ayant déjà 10 ans d'expérience, le gain est de '**c + 2b10**'.

Le coefficient 'b' détermine si le gain salarial augmente ou diminue avec l'expérience. Le salaire maximal est atteint à un niveau d'expérience 'exp*' où

$$\mathbf{c + 2bexp = 0}, \text{ soit } \mathbf{exp* = -c/2b}.$$

Dans un modèle linéaire parfait, tous les points se trouveraient sur une droite de régression **y = a+bx**.



Dans la réalité, divers facteurs peuvent influencer la variable dépendante 'y', accordant une incertitude. Cela conduit à une déviation de chaque point de l'échantillon par rapport au modèle théorique, ce qui est représenté par le terme d'erreur 'ui'.

Par conséquent, notre équation de base pour estimer un modèle linéaire devient ' **$y_i = a + b x_i + u_i$** ', où 'a' est la constante, 'b' est l'effet marginal, et 'ui' est le terme d'erreur . Si une seule variable indépendante 'x' est utilisée, on parle de modèle linéaire simple ; s'il y a plusieurs variables, c'est un modèle de régression multiple.

Limitations du modèle de régression linéaire simple

"En plus des violations des hypothèses, le modèle de régression linéaire simple a d'autres limites. Il ne peut pas capturer les relations non linéaires ou l'impact de plusieurs variables explicatives. Pour ces situations, nous avons besoin de modèles plus avancés, tels que le modèle de régression linéaire multiple ou les modèles non linéaires."

Modèle de régression multiple :

une extension du modèle simple "Pour inclure plus de variables explicatives, nous pouvons étendre le modèle simple à un modèle de régression multiple. L'équation est $y = a + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + u$. Chaque x_i est une variable explicative différente, et chaque β_i est le changement attendu dans y lorsque x_i change d'une unité, toutes les autres variables restant constantes."

Régression non linéaire :

une autre extension "Si la relation entre y et x n'est pas linéaire, nous pouvons utiliser un modèle de régression non linéaire. Par exemple, **y = $\alpha + \beta_1x + \beta_2x^2 + u$** est un modèle de régression qui capture une relation parabolique entre y et x. β_1 est l'effet de x sur y, et β_2 est l'effet de x^2 sur y. Cette forme de modèle nous permet de capturer des effets qui augmentent ou diminuent avec x."

Le modèle de régression linéaire simple est un outil puissant pour comprendre les relations entre deux variables. Malgré ses limitations et les hypothèses sur lesquelles il repose, il est souvent une première étape utile dans l'analyse des données économiques. À mesure que nous ajoutons plus de variables et explorons des relations non linéaires, nous pouvons développer des modèles qui capturent de manière plus précise les phénomènes économiques complexes.



Partie 3: Régression linéaire multiple



Introduction :

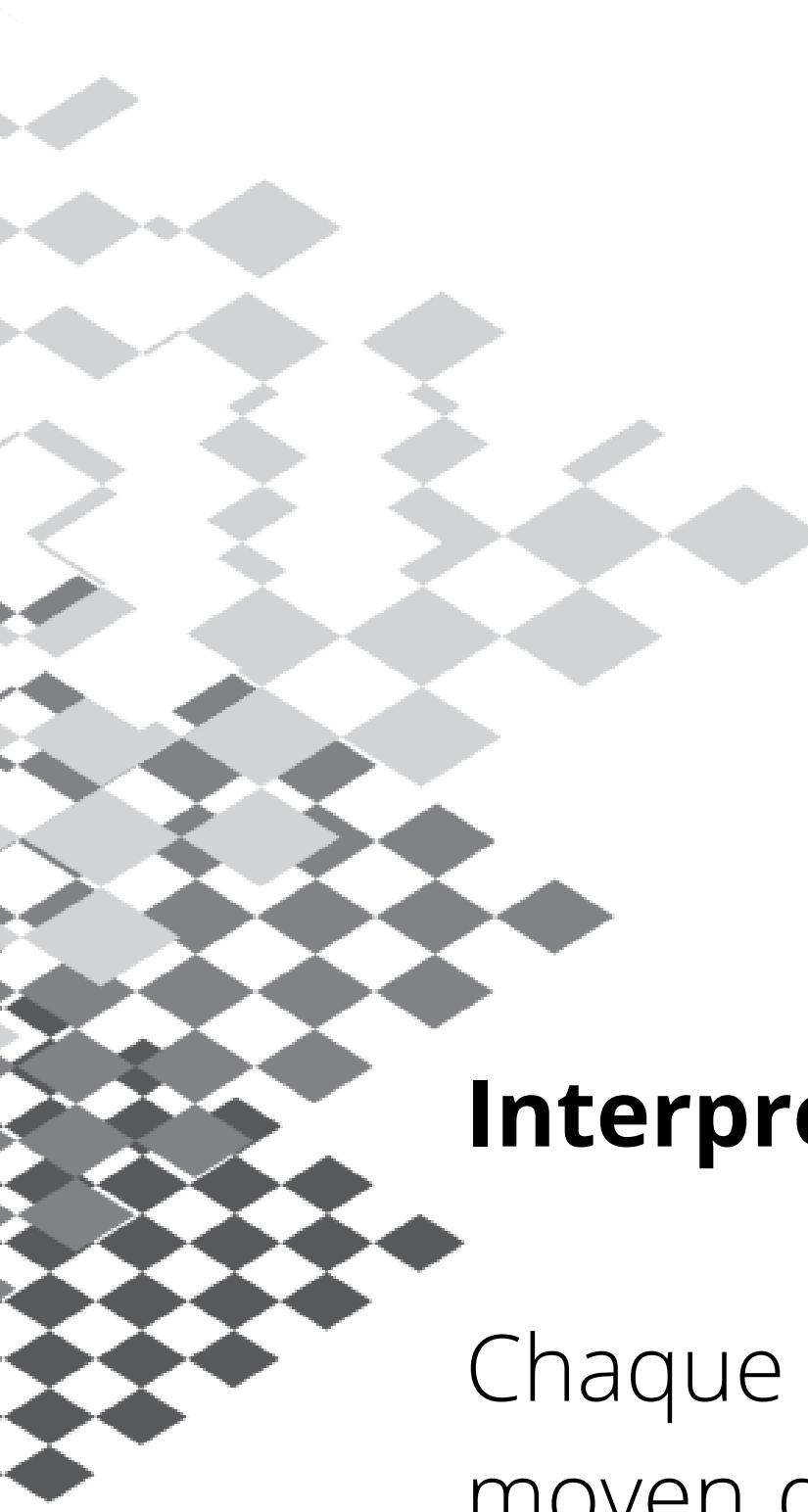
La régression linéaire multiple est une extension du modèle de régression linéaire simple qui comprend plus d'une variable indépendante. Elle permet de modéliser la relation entre une variable dépendante et plusieurs variables indépendantes. Formellement, elle est exprimée comme suit : **$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n * X_n + \epsilon$**

Ici, Y est la variable dépendante, X₁ à X_n sont les variables indépendantes, β₀ à β_n sont les coefficients de régression à estimer, et ε est le terme d'erreur.

Hypothèses de base :

Il y a plusieurs hypothèses clés derrière un modèle de régression linéaire multiple.

- Linéarité : La relation entre les variables indépendantes et la variable dépendante est linéaire.
- Indépendance : Les erreurs sont indépendantes les unes des autres.
- Homoscédasticité : La variance des erreurs est constante pour toutes les valeurs des variables indépendantes.
- Normalité : Les erreurs de suivi d'une distribution normale.



Estimation des coefficients :

Les coefficients β se présentent généralement en utilisant la méthode des moindres carrés ordinaires (OLS), qui minimise la somme des carrés des résidus.

Interprétation des coefficients :

Chaque coefficient β_i (pour $i = 1$ à n) indique l'effet moyen d'une augmentation de 1 unité de la variable indépendante X_i sur la variable dépendante Y , toutes choses étant égales par ailleurs.

Tests de significativité :

Une fois les coefficients révélés, des tests de significativité (comme le test t ou le test F) peuvent être utilisés pour déterminer si les variables indépendantes ont un effet significatif sur la variable dépendante.

1. **Le test t** : Pour chaque coefficient β_i , le test t est utilisé pour tester l'hypothèse nulle que $\beta_i = 0$ (c'est-à-dire que la variable X_i n'a pas d'effet sur Y) contre l'hypothèse alternative que $\beta_i \neq 0$.

1. **Le test F** : Le test F est utilisé pour tester l'hypothèse nulle que tous les coefficients β_i sont égaux à zéro (c'est-à-dire que toutes les variables X sont inutiles pour prédire Y) contre l'hypothèse alternative qu'au moins un β_i est différent de zéro.

1. **Multicollinéarité** : Il s'agit d'un phénomène où deux ou plus variables indépendantes sont fortement corrélées. On peut le détecter en utilisant le Facteur d'Inflation de la Variance (VIF) pour chaque variable. Formule : **VIF(i) = 1/(1 - R^2(i))**.
2. **Hétéroscédasticité** : Elle se produit lorsque la variance de l'erreur varie avec les valeurs des variables indépendantes. Pour la gérer, on peut utiliser des erreurs standard robustes, ou transformer la variable dépendante.

Autocorrélation :

Il s'agit de la corrélation entre les termes d'erreur, qui est un problème commun dans les données de séries temporelles. Le test de Durbin-Watson est généralement utilisé pour détecter l'autocorrélation. Le test produit une statistique qui varie de 0 à 4, où une valeur proche de 2 indique l'absence d'autocorrélation.

Erreurs de spécification du modèle :

Cela se produit lorsque le modèle que nous avons choisi ne reflète pas la véritable relation entre les variables. Par exemple, si nous omettons des variables importantes, ou si nous incluons des variables qui ne sont pas applicables. On peut utiliser le test RESET de Ramsey pour détecter une mauvaise spécification du modèle.

L'interaction entre les variables :

Dans certains cas, l'effet d'une variable indépendante sur la variable dépendante peut dépendre de la valeur d'une autre variable indépendante. Dans ce cas, nous pouvons inclure un terme d'interaction dans notre modèle, qui est simplement le produit de deux variables indépendantes. Par exemple, si nous avons deux variables indépendantes X_1 et X_2 , le terme d'interaction serait $X_1 * X_2$.

Variables indicatrices :

Ce sont des variables qui prennent la valeur 1 ou 0 pour indiquer la présence ou l'absence d'une certaine caractéristique ou d'un certain attribut. Les variables indicatrices peuvent être utilisées pour modéliser des effets non linéaires et sont souvent utilisées pour inclure des variables catégorielles dans un modèle de régression.

Vérification des hypothèses du modèle :

Après avoir révélé le modèle, il est important de vérifier que les hypothèses sur lesquelles repose le modèle sont respectées. Cela comprend l'examen des résidus pour vérifier la linéarité, l'homoscédasticité, l'indépendance et la normalité. Des graphiques résiduels peuvent être utiles pour cette tâche.



Modèle Log-Lin et Lin-Log :

Dans certains cas, une transformation logarithmique de la variable dépendante et/ou des variables indépendantes peut fournir un meilleur ajustement du modèle. Un modèle Log-Lin est un modèle où la variable dépendante est transformée par le log, alors que dans un modèle Lin-Log, ce sont les variables indépendantes qui sont transformées. Ces transformations peuvent aider à gérer l'hétéroscédasticité et à modéliser des relations non-linéaires.

Le modèle et ses hypothèses

Ecriture du modèle

Soit un modèle linéaire de la forme :

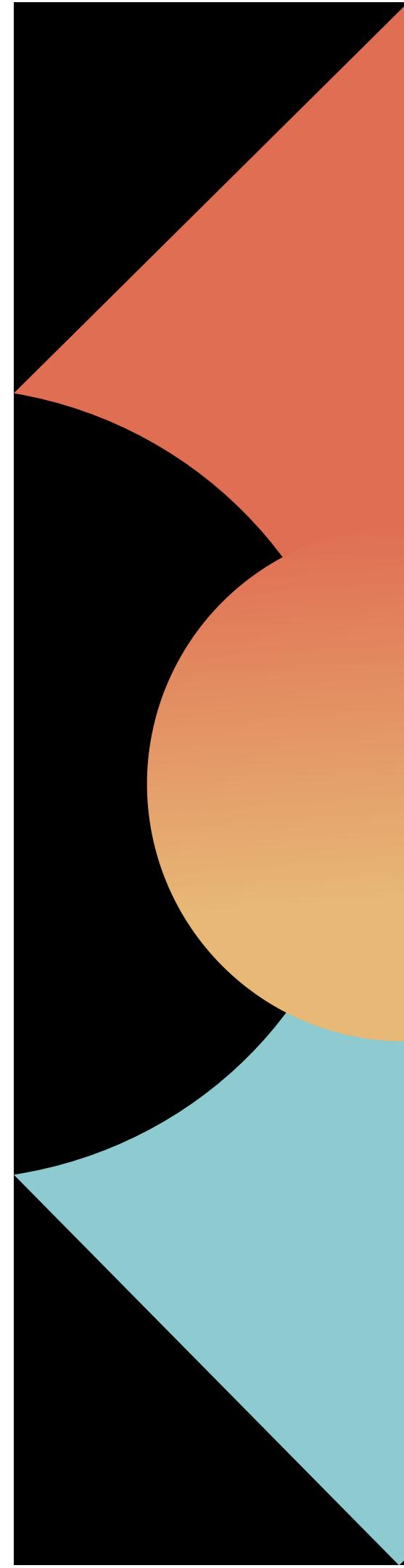
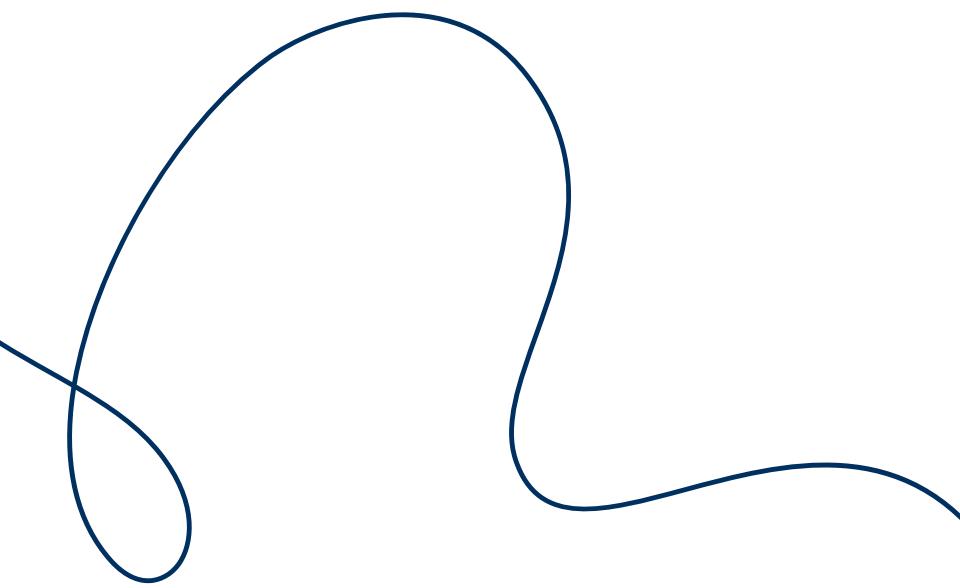
$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \dots + b_kx_{ki} + u_i, i = 1, \dots, N$$

Pour estimer ce modèle on dispose de N observations et k variables.

Pour pouvoir estimer ce modèle il est important de respecter la règle suivante : $k \leq N$

Pour aboutir à une écriture matricielle on empile les N observations de la façon suivante : $y_1 = b_0 + b_1x_{11} + b_2x_{21} + \dots + b_kx_{k1} + u_1$

$$y_2 = b_0 + b_1x_{12} + b_2x_{22} + \dots + b_kx_{k2} + u_2 \dots y_N = b_0 + b_1x_{1N} + b_2x_{2N} + \dots + b_kx_{kN} + u_N$$



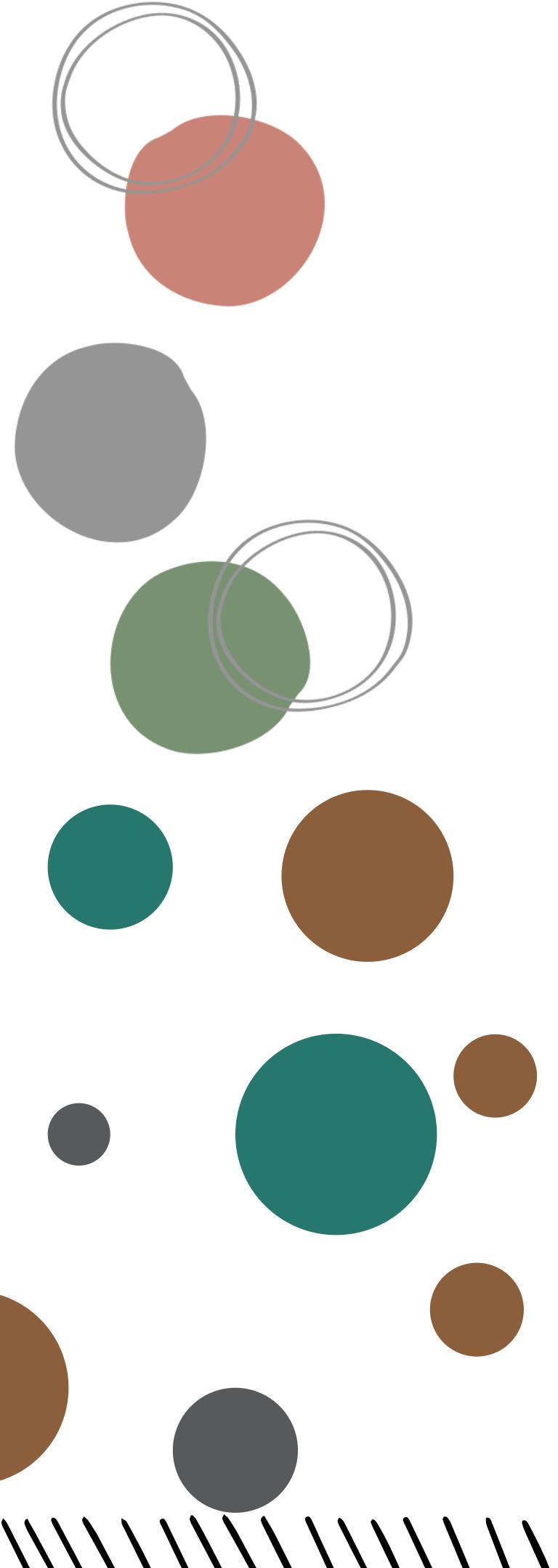


Comparaison de modèles :

Pour choisir le modèle de régression le plus adapté, on utilise généralement le R-carré, le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC). Le R-carré ajusté prend en compte le nombre de prédicteurs dans le modèle, avec une préférence pour un modèle ayant un R-carré ajusté élevé. L'AIC et le BIC consomment également en compte le nombre de paramètres, en privilégiant les modèles qui minimisent ces critères.

$$\text{R-carré ajusté} = 1 - (1 - R^2) * (n - 1) / (n - p - 1)$$
$$AIC = 2k - 2\ln(L)$$
$$BIC = n * \ln(\text{SSR}/n) + k * \ln(n)$$

Où n est le nombre d'observations, p est le nombre de prédicteurs, k est le nombre de paramètres révélés dans le modèle, L est la vraisemblance maximale du modèle, SSR est la somme des carrés des résidus.



Méthodes de sélection de variables :

Des techniques telles que les méthodes "forward", "backward" et "stepwise" peuvent être utilisées pour sélectionner les variables à inclure dans un modèle. La sélection forward commence avec aucun prédicteur, ajoute le prédicteur qui améliore le plus le modèle, et répète ce processus jusqu'à ce qu'aucun prédicteur ne puisse améliorer le modèle.

La sélection backward commence avec tous les prédicteurs, supprime le prédicteur qui améliore le plus le modèle lorsqu'il est supprimé, et répète ce processus jusqu'à ce qu'aucun prédicteur ne puisse être supprimé pour améliorer le modèle. La sélection stepwise est une combinaison des deux autres méthodes.



Utilisation de logiciels d'économétrie :

L'économétrie s'appuie fortement sur des logiciels spécialisés pour l'analyse des données. Des outils tels que STATA et EViews ont traditionnellement dominé le domaine. Cependant, de plus en plus, Python s'impose comme le langage de choix pour l'économétrie. Dans ce cours, nous mettrons un accent particulier sur Python pour plusieurs raisons.

Python en économétrie :

- Python est un langage de programmation puissant qui dépasse les capacités des logiciels d'analyse de données traditionnels. Il est capable de manipuler des données volumineuses et d'implémenter des modèles économétriques avancés.
- Grâce à sa syntaxe claire et concise, Python facilite la lecture, l'écriture et la compréhension du code.
- Python est extrêmement flexible, capable d'effectuer diverses analyses et d'être appliqué à divers domaines, ce qui est idéal pour les économistes.

- Soutenu par une communauté dynamique, Python possède de nombreuses bibliothèques dédiées à l'analyse économétrique (par exemple, StatsModels et Pandas) qui sont constamment mises à jour et améliorées.
- En tant que logiciel open-source, Python est accessible à tous, ce qui le rend excellent pour l'enseignement de l'économétrie, garantissant une participation équitable de tous les étudiants.

Par exemple, pour ajuster un modèle de régression multiple à un ensemble de données dans Python, vous pourriez utiliser le code suivant:

python

 Copy code

```
import statsmodels.api as sm

X = sm.add_constant(df[['Variable1', 'Variable2']])

Y = df['Outcome']

model = sm.OLS(Y, X)
results = model.fit()

print(results.summary())
```

“

Ce code crée un modèle de **régression OLS** en utilisant 'Variable1' et 'Variable2' pour prédire 'Outcome'. Le tableau de résumé généré par la méthode **summary()** donne un aperçu détaillé des résultats de la régression, y compris les coefficients de régression, les erreurs standards, les statistiques t et les valeurs p, entre autres informations utiles.

”

Il est également possible de visualiser les relations entre les variables en utilisant des bibliothèques comme Matplotlib ou Seaborn. Par exemple, pour créer un nuage de points des résidus versus les valeurs ajustées, vous pouvez utiliser le code suivant:

python

 Copy code

```
import matplotlib.pyplot as plt

plt.scatter(results.fittedvalues, results.resid)
plt.xlabel('Fitted values')
plt.ylabel('Residuals')
plt.title('Residuals vs Fitted Values')
plt.show()
```

Modèles non linéaires :

Parfois, une relation linéaire ne décrit pas adéquatement les données. Dans ces cas, des modèles non linéaires peuvent être utilisés. Par exemple, un modèle de régression logarithmique peut être utilisé si l'effet d'un prédicteur sur la réponse change en fonction du niveau du prédicteur. Un modèle de ce type pourrait ressembler à

$$\mathbf{Y} = \mathbf{a} + \mathbf{b} \ln(\mathbf{X}) + \mathbf{e}.$$

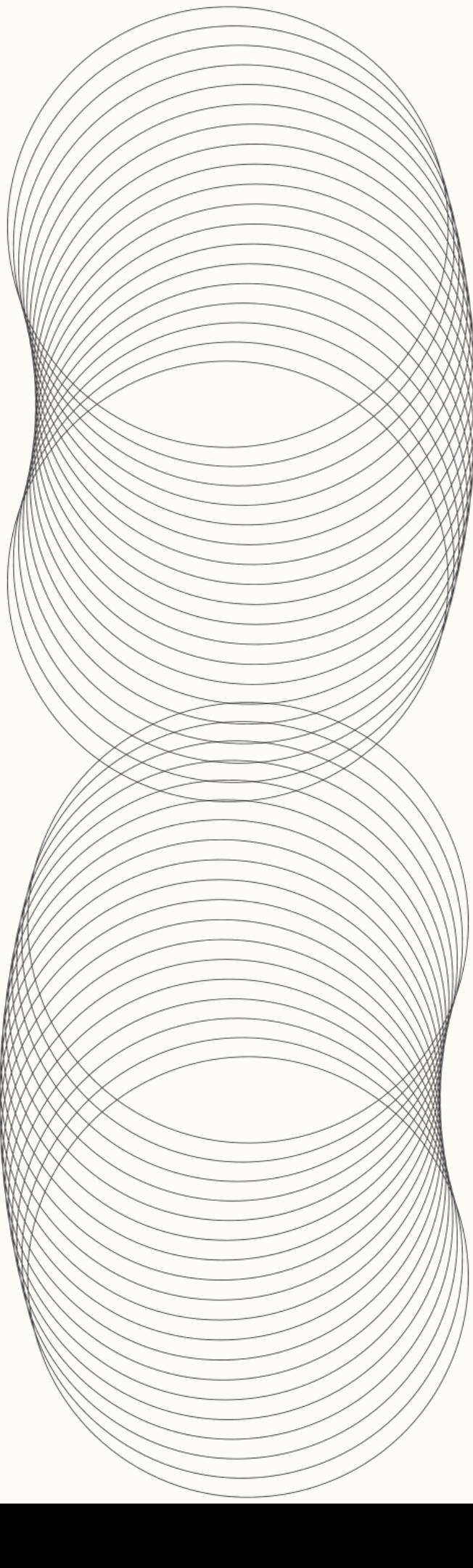
Partie 4: Test d'hypothèse et diagnostic des modèles

Introduction

L'économétrie est une branche de l'économie qui utilise des techniques statistiques pour analyser les relations économiques. Les tests d'hypothèse et les diagnostics des modèles sont des outils fondamentaux pour évaluer la robustesse des résultats économétriques et assurer la validité des conclusions tirées à partir des données. Cette présentation vise à fournir une compréhension approfondie de ces concepts clés.

Rappel des concepts fondamentaux de l'économétrie

Un modèle économétrique de base peut être exprimé sous la forme suivante : **$y = X\beta + \epsilon$** , où y est la variable dépendante, X est la matrice des variables indépendantes, β est le vecteur des coefficients et ϵ est le terme d'erreur. L'estimation des paramètres se fait généralement par la méthode des moindres carrés ordinaires (MCO). Les hypothèses de Gauss-Markov jouent un rôle crucial dans la validité des résultats économétriques.



Les tests d'hypothèse

Les tests d'hypothèse permettent de prendre des décisions statistiques concernant les paramètres du modèle. Ils comparent les valeurs estimées avec des valeurs théoriques ou des hypothèses nulles pour déterminer si ces valeurs diffèrent de manière significative. Les tests t et les tests F sont couramment utilisés.

Les tests t sont utilisés pour évaluer la signification individuelle des coefficients. La statistique de test t est calculée en divisant l'estimation du coefficient par son erreur standard. Si la statistique de test t est significativement différente de zéro, cela indique que le coefficient est statistiquement différent de zéro.

Un exemple illustratif serait un modèle de régression simple : $\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$. Les hypothèses nulles pourraient être $H_0 : \beta_1 = 0$ (aucun effet de x sur y) et $H_1 : \beta_1 \neq 0$ (effet significatif de x sur y). Le calcul de la statistique de test t et de la p-value permet de déterminer si l'hypothèse nulle est rejetée ou non.



Les tests F sont utilisés pour évaluer la significativité globale du modèle. Ils comparent la somme des carrés des résidus (SSR) entre deux modèles : un modèle restreint (qui omet certaines variables) et un modèle non restreint (qui inclut toutes les variables). La statistique de test F est calculée en comparant la variation expliquée par les variables explicatives par rapport à la variation résiduelle.

Prenons l'exemple d'un modèle de régression multiple : $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \varepsilon$. Supposons que nous voulions tester l'hypothèse nulle $H_0 : \beta_1 = \beta_2 = 0$, ce qui signifie que les variables x_1 et x_2 n'ont aucun effet significatif sur y . Nous pouvons calculer la statistique de test F en comparant le modèle complet avec le modèle restreint qui exclut ces variables. Si la statistique de test F est significativement différente de zéro, nous rejetons l'hypothèse nulle et concluons qu'il existe un effet significatif des variables x_1 et x_2 sur y .

Il est important de noter que la signification statistique des tests dépend du seuil de signification choisi, généralement $\alpha = 0,05$. Si la p-value associée au test est inférieure à α , nous rejetons l'hypothèse nulle.

Le diagnostic des modèles

Le diagnostic des modèles consiste à évaluer la validité des hypothèses sous-jacentes aux modèles économétriques. Cela permet de vérifier si les conditions requises pour les résultats économétriques sont respectées. Les diagnostics de spécification se concentrent sur les violations potentielles des hypothèses de base, telles que l'homoscédasticité, l'autocorrélation, la normalité des résidus, etc.

Un premier diagnostic essentiel est celui de l'homoscédasticité, qui suppose une variance constante des erreurs. La violation de cette hypothèse, appelée hétéroscédasticité, peut biaiser les estimations et les tests statistiques. Les tests de White et de Breusch-Pagan sont couramment utilisés pour détecter l'hétéroscédasticité. Ils comparent les variations des résidus en fonction des variables explicatives.



Un exemple concret serait d'appliquer le test de White à un modèle de régression simple : $y = \beta_0 + \beta_1 x + \varepsilon$. Nous estimons le modèle de régression des résidus au carré sur les variables explicatives et calculons la statistique de test de White, qui est une version ajustée du test de Breusch-Pagan. Si la statistique de test est significativement différente de zéro, cela indique la présence d'hétéroscédasticité.



Un autre diagnostic important est celui de l'autocorrélation, qui suppose l'absence de corrélation serielle entre les résidus. L'autocorrélation peut se produire lorsque les erreurs ne sont pas indépendantes au fil du temps, ce qui conduit à une inefficacité des estimations et à des erreurs de spécification. Le test de Durbin-Watson est largement utilisé pour détecter l'autocorrélation.

Dans un modèle de régression simple : $y = \beta_0 + \beta_1x + \varepsilon$, le test de Durbin-Watson calcule une statistique de test basée sur la corrélation entre les résidus adjacents. Si la statistique de test est proche de 2, cela suggère l'absence d'autocorrélation. Des valeurs significativement différentes de 2 indiquent l'existence d'autocorrélation positive ou négative. Dans ce cas, des ajustements tels que la transformation des variables ou l'inclusion de variables retardées peuvent être nécessaires pour corriger l'autocorrélation.

Outre l'hétéroscédasticité et l'autocorrélation, d'autres diagnostics de spécification incluent l'analyse de la normalité des résidus, la détection de la multicolinéarité et l'évaluation de la stabilité des coefficients dans le temps. Des tests statistiques tels que le test de Jarque-Bera pour la normalité, le test de variance inflation factor (VIF) pour la multicolinéarité et les tests de Chow pour la stabilité des coefficients peuvent être utilisés respectivement.

Interprétation des résultats des tests d'hypothèse et des diagnostics

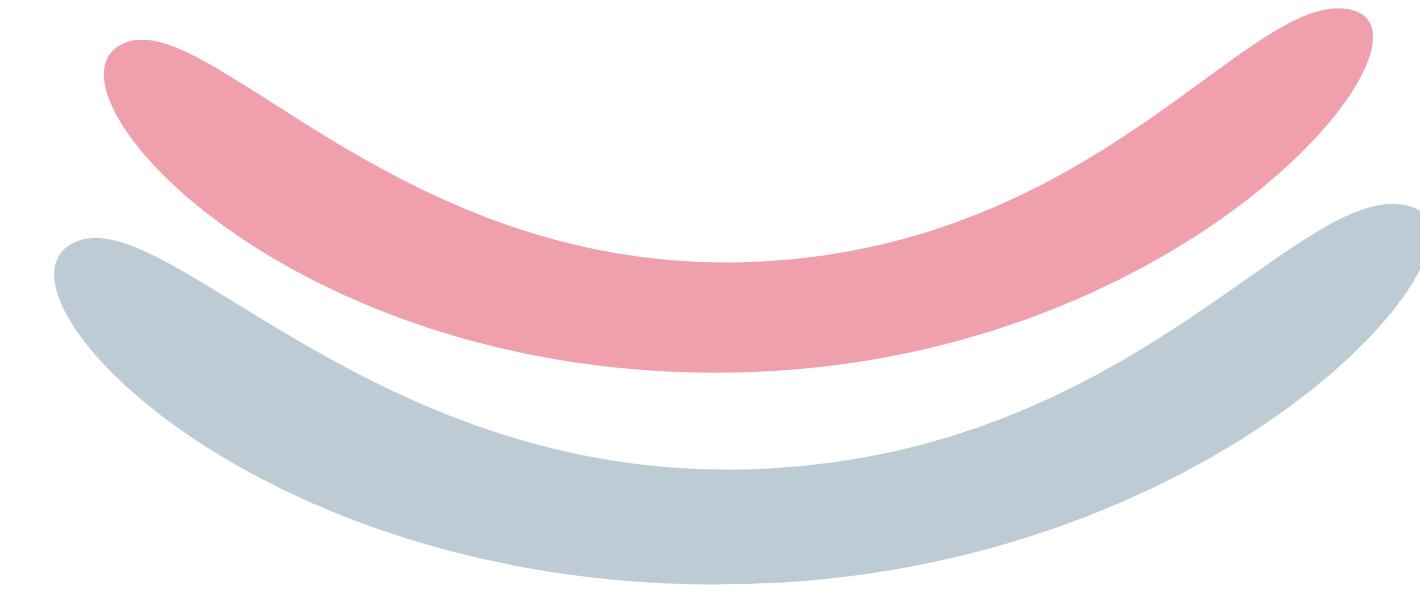
Il est important de noter que les tests d'hypothèse et les diagnostics ne fournissent pas de preuves définitives, mais plutôt des indications et des preuves statistiques pour soutenir ou réfuter certaines hypothèses. L'interprétation des résultats doit se faire avec prudence et en tenant compte du contexte économique et des objectifs de recherche.





Lorsque les tests d'hypothèse indiquent que l'hypothèse nulle peut être rejetée, cela suggère que les variables ont un effet significatif sur la variable dépendante. Cependant, la taille de l'effet et sa signification économique doivent également être évaluées.

De même, lorsqu'un diagnostic de spécification détecte une violation des hypothèses de base, il est nécessaire d'examiner la nature et l'ampleur de cette violation, ainsi que les actions correctives possibles pour améliorer la spécification du modèle.



Les tests d'hypothèse et les diagnostics des modèles sont des outils essentiels en économétrie pour évaluer la robustesse des résultats économétriques et s'assurer de la validité des conclusions. Ils permettent de vérifier les hypothèses sous-jacentes et de détecter les problèmes potentiels de spécification des modèles.

Il est crucial de comprendre les concepts, les méthodes et les limitations de ces outils pour une utilisation appropriée. Une utilisation judicieuse des tests d'hypothèse et des diagnostics peut renforcer la confiance dans les résultats économétriques et contribuer à des analyses économiques rigoureuses.

Partie 5: Régression non linéaire

Introduction

La régression non linéaire est une extension puissante du modèle de régression linéaire classique qui permet de modéliser des relations plus complexes et flexibles entre les variables. Alors que le modèle de régression linéaire suppose une relation linéaire entre la variable dépendante et les variables explicatives, le modèle de régression non linéaire offre la possibilité de capturer des relations fonctionnelles plus adaptées aux données.

Rappel du modèle de régression linéaire

Avant de plonger dans les modèles de régression non linéaire, rappelons brièvement le modèle de régression linéaire. Il s'exprime comme suit : $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k + \varepsilon$, où y est la variable dépendante, x_1, x_2, \dots, x_k sont les variables explicatives, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ sont les coefficients à estimer, et ε est le terme d'erreur.

Le modèle de régression linéaire est largement utilisé et repose sur l'hypothèse de linéarité de la relation entre les variables. Cependant, cette approche peut être limitée lorsque la relation réelle est non linéaire. C'est là que les modèles de régression non linéaire entrent en jeu.

Modèle de régression non linéaire : Formulation générale

Un modèle de régression non linéaire général peut être formulé comme suit : $y = f(x, \beta) + \varepsilon$, où $f(\cdot)$ est une fonction non linéaire de x et de β . La fonction $f(\cdot)$ peut prendre différentes formes, telles que des polynômes, des fonctions exponentielles, logarithmiques, ou encore des fonctions trigonométriques.

La spécification de la fonction $f(\cdot)$ dépend du contexte et des hypothèses théoriques sous-jacentes. Le choix d'une fonction appropriée est crucial pour capturer la relation non linéaire entre les variables de manière adéquate.

Exemple de modèle de régression non linéaire : Régression polynomiale

Un exemple courant de modèle de régression non linéaire est la régression polynomiale. Elle permet de modéliser des relations polynomiales entre la variable dépendante et les variables explicatives. Un modèle de régression polynomiale d'ordre k peut être écrit comme suit : $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k + \varepsilon$.

Dans ce modèle, y représente la variable dépendante, x est la variable explicative, $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ sont les coefficients à estimer, x^2, x^3, \dots, x^k sont les termes d'interaction et d'ordre supérieur, et ε est le terme d'erreur.

La régression polynomiale permet de capturer des relations non linéaires de manière flexible. Par exemple, un modèle de régression polynomiale d'ordre 2 (parabolique) peut capturer une courbure dans la relation entre les variables.

L'estimation des coefficients dans la régression polynomiale peut être réalisée à l'aide de la méthode des moindres carrés ordinaires (MCO) en minimisant la somme des carrés des résidus. L'estimation des coefficients β_0 , β_1 , β_2 , ..., β_k peut être effectuée en résolvant le système d'équations normales résultant de la minimisation de la somme des carrés des résidus.

Cependant, il est important de noter que des problèmes de surajustement (overfitting) peuvent survenir avec des ordres polynomiaux élevés. Il est donc essentiel de sélectionner l'ordre du polynôme de manière appropriée, en utilisant des méthodes de validation croisée ou des critères d'information tels que le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC).

Exemple de modèle de régression non linéaire : Régression exponentielle

Un autre exemple de modèle de régression non linéaire est la régression exponentielle. Elle est utilisée lorsque la relation entre les variables suit une croissance ou une décroissance exponentielle. Un modèle de régression exponentielle peut être écrit comme suit : $y = \beta_0 e^{(\beta_1 x)} + \varepsilon$, où y est la variable dépendante, x est la variable explicative, β_0 et β_1 sont les coefficients à estimer, e est la base du logarithme népérien, et ε est le terme d'erreur.

La régression exponentielle permet de capturer des tendances de croissance ou de décroissance exponentielle dans les données. Les coefficients β_0 et β_1 déterminent l'amplitude et la pente de la courbe exponentielle, respectivement.

L'estimation des coefficients dans la régression exponentielle peut être réalisée à l'aide de techniques d'optimisation, telles que la méthode des moindres carrés non linéaires ou la méthode du maximum de vraisemblance.

Méthode des moindres carrés non linéaires (MCNL) :

Cette méthode cherche à minimiser la somme des carrés des résidus en ajustant les paramètres du modèle de manière itérative. Elle utilise des techniques d'optimisation numérique, telles que la méthode de Gauss-Newton ou la méthode du gradient, pour trouver les valeurs des coefficients qui minimisent l'écart entre les valeurs observées et les valeurs prédites par le modèle.

Méthode du maximum de vraisemblance (MV) :

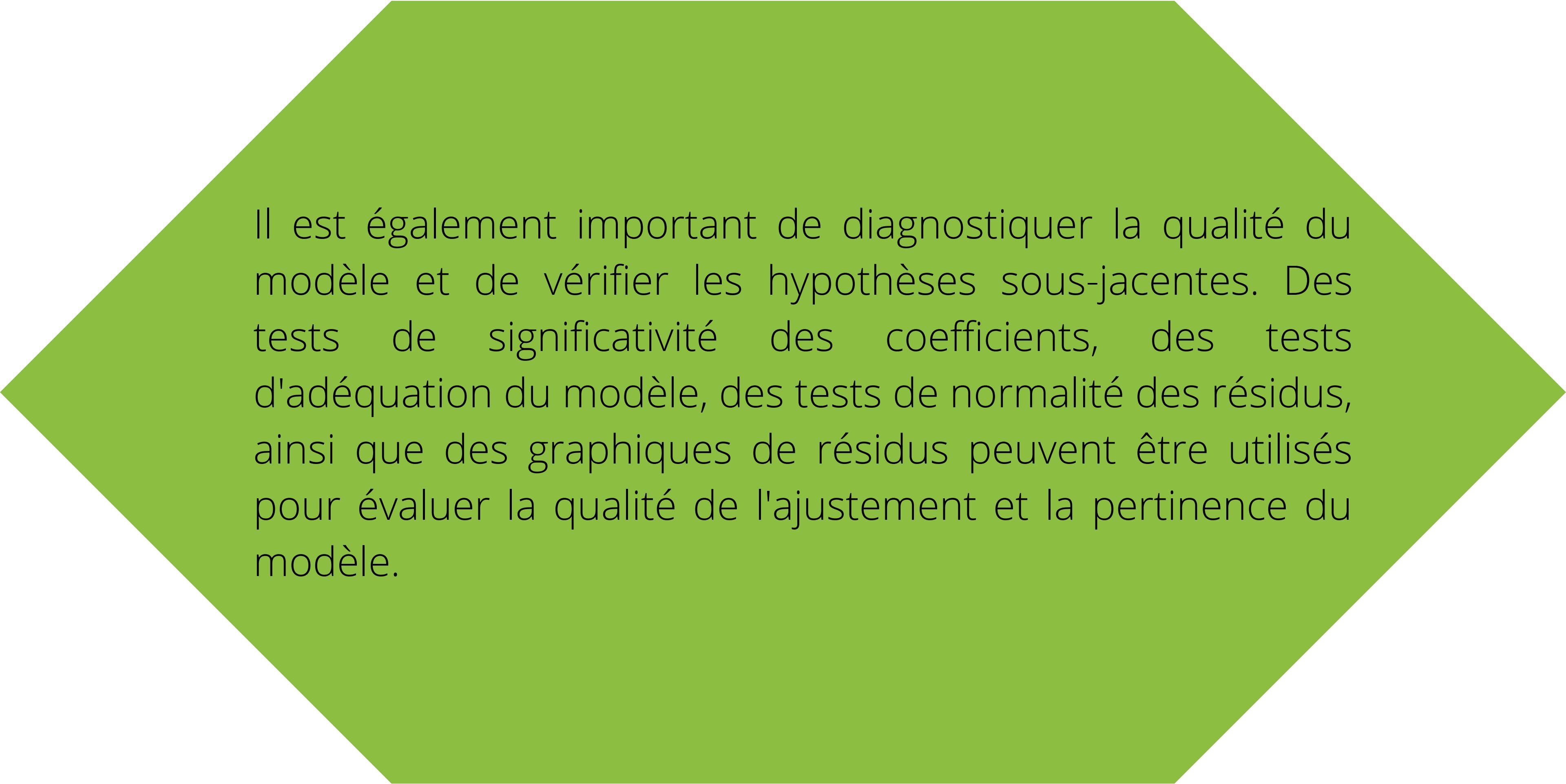
Cette méthode est utilisée lorsque les erreurs du modèle de régression non linéaire suivent une distribution probabiliste spécifique. Elle cherche à maximiser la vraisemblance des données observées en ajustant les paramètres du modèle. La vraisemblance est une mesure de la probabilité d'observer les données réelles compte tenu des paramètres du modèle. Des techniques d'optimisation, telles que l'algorithme de Newton-Raphson, sont souvent utilisées pour maximiser la vraisemblance.

Ces deux méthodes sont couramment utilisées pour estimer les paramètres des modèles de régression non linéaire. Le choix de la méthode dépend du contexte de recherche, des hypothèses sur les erreurs du modèle et des propriétés statistiques souhaitées.

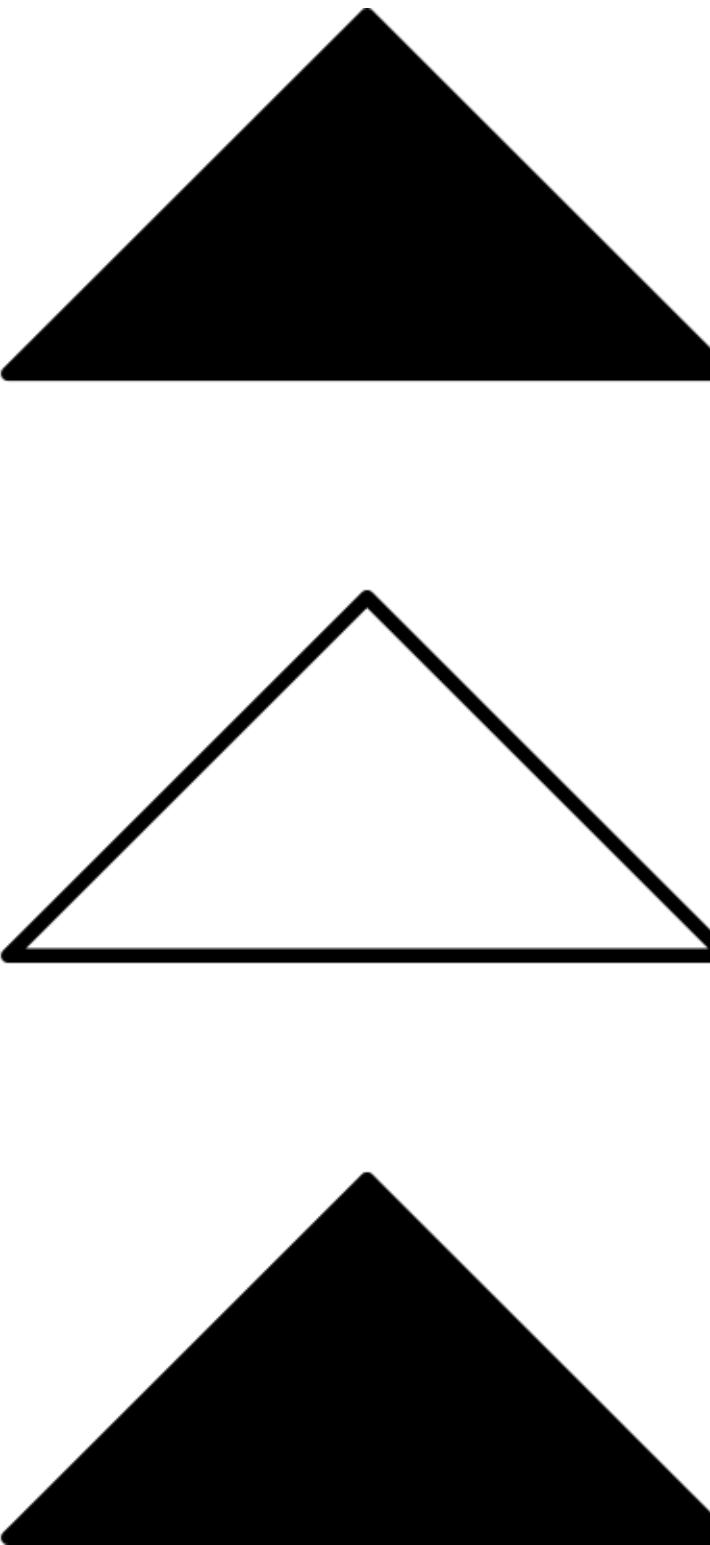
Interprétation des résultats et diagnostics

L'interprétation des résultats dans les modèles de régression non linéaire peut être plus complexe que dans les modèles linéaires en raison de la non-linéarité de la fonction de régression. Néanmoins, il est possible d'interpréter les coefficients de manière marginale, c'est-à-dire en mesurant les variations de la variable dépendante pour une unité de variation de la variable explicative, en tenant compte des autres variables explicatives constantes.

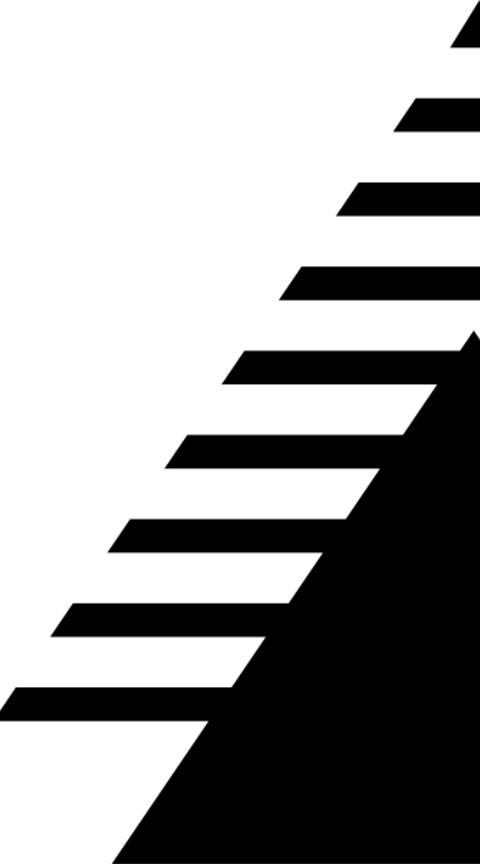
Par exemple, dans le cas d'une régression polynomiale d'ordre 2, le coefficient β_1 mesure la variation de la variable dépendante pour une unité de variation de la variable explicative x , tandis que le coefficient β_2 mesure l'effet d'une variation quadratique de x sur y .



Il est également important de diagnostiquer la qualité du modèle et de vérifier les hypothèses sous-jacentes. Des tests de significativité des coefficients, des tests d'adéquation du modèle, des tests de normalité des résidus, ainsi que des graphiques de résidus peuvent être utilisés pour évaluer la qualité de l'ajustement et la pertinence du modèle.



Il convient de noter que les modèles de régression non linéaire ne conviennent pas à toutes les situations. Il est important de prendre en compte la nature des données et de considérer d'autres approches statistiques si la relation entre les variables n'est pas clairement non linéaire.



Il est recommandé de se familiariser davantage avec les techniques d'estimation spécifiques à chaque type de modèle non linéaire et d'utiliser des logiciels statistiques appropriés pour leur application. De plus, il est essentiel de continuer à approfondir ses connaissances en économétrie et de rester à jour avec les développements récents dans le domaine.



Partie 6: Variable instrumentale et méthode des moments généralisés

Introduction à la variable instrumentale et à la méthode des moments généralisés

Définition de la variable instrumentale :

- Une variable instrumentale (VI) est une variable utilisée dans les modèles économétriques pour résoudre le problème d'endogénéité.
- Une VI doit être corrélée avec la variable endogène, mais ne doit pas être corrélée avec le terme d'erreur.



Comprendre l'endogénéité

L'endogénéité fait référence à la situation où une ou plusieurs des variables explicatives sont corrélées avec l'erreur de régression. Cela peut entraîner un biais dans les estimations des coefficients de régression.



Variable instrumentale

Une variable instrumentale est une variable qui est corrélée avec la variable explicative, mais pas avec l'erreur de régression. L'objectif est de surmonter le problème du biais d'endogénéité.

Formule générale : $Z = \text{Variable Instrumentale}$

Sélection d'une variable instrumentale

Choisir une bonne variable instrumentale peut être délicat. Elle doit être corrélée avec la variable explicative, mais elle ne doit pas être corrélée avec l'erreur. Cette deuxième condition est souvent difficile à vérifier.

Utilisation de variables instrumentales pour résoudre l'endogénéité

- Supposons un modèle de régression linéaire simple : $Y = \beta_0 + \beta_1 X + U$
- L'endogénéité se produit lorsque X est corrélée avec U , ce qui peut biaiser les estimations de β_1 .
- En introduisant une VI Z , le modèle devient : $Y = \beta_0 + \beta_1 X + \gamma Z + U$
- La VI Z doit être corrélée avec X mais ne doit pas être corrélée avec U .



Importance de la variable instrumentale dans les modèles économétriques :

- L'endogénéité peut entraîner un biais de causalité dans les estimations des coefficients.
- Les VI permettent de créer une situation d'expérience naturelle ou quasi-expérimentale pour établir une relation de causalité plus robuste.





Sélection d'une variable instrumentale

Choisir une bonne variable instrumentale peut être délicat. Elle doit être corrélée avec la variable explicative, mais elle ne doit pas être corrélée avec l'erreur. Cette deuxième condition est souvent difficile à vérifier.

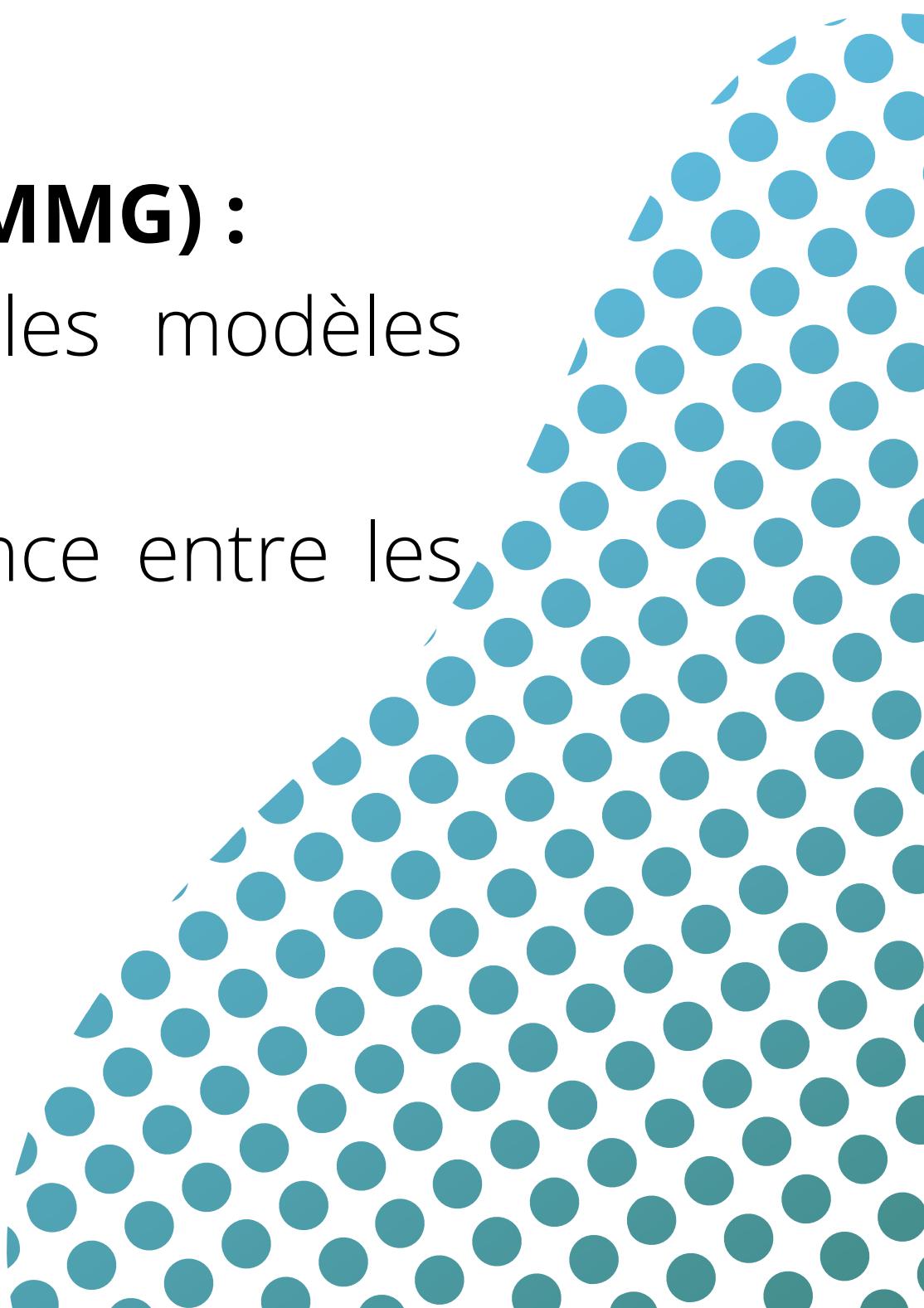
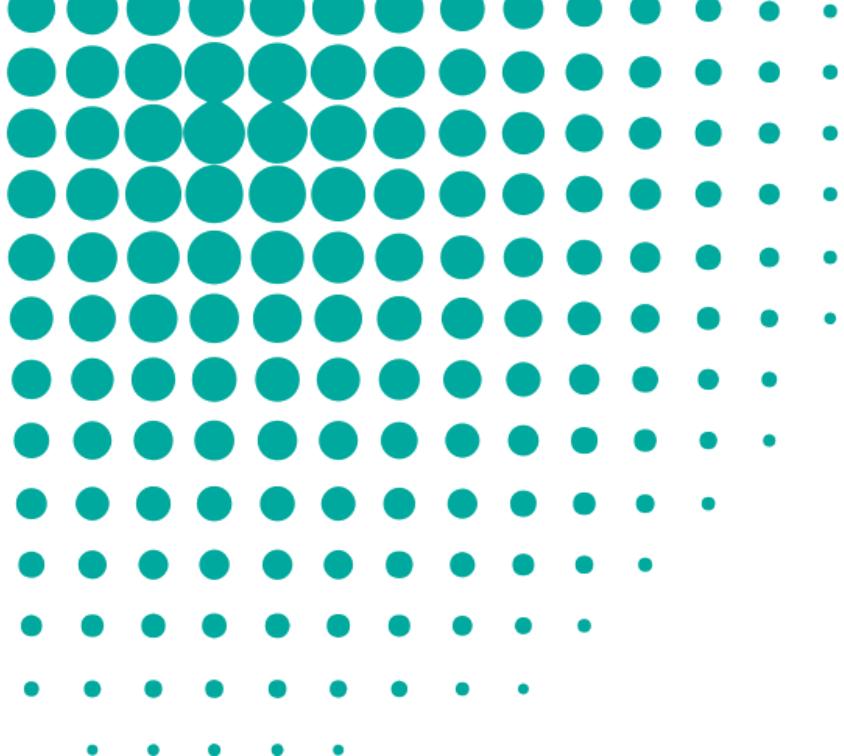


Sélection d'une variable instrumentale

Choisir une bonne variable instrumentale peut être délicat. Elle doit être corrélée avec la variable explicative, mais elle ne doit pas être corrélée avec l'erreur. Cette deuxième condition est souvent difficile à vérifier.

Méthode des Moments Généralisés (MMG)

La méthode des moments généralisés est une extension de la méthode des moments. Elle peut être utilisée pour estimer les paramètres d'un modèle économétrique lorsqu'il y a une violation de la condition d'indépendance entre les variables explicatives et l'erreur de régression.



Objectif de la méthode des moments généralisés (MMG) :

- La MMG est une technique statistique pour estimer les modèles économétriques avec des VI.
- Elle repose sur le principe de minimisation de la différence entre les moments observés et les moments théoriques du modèle.

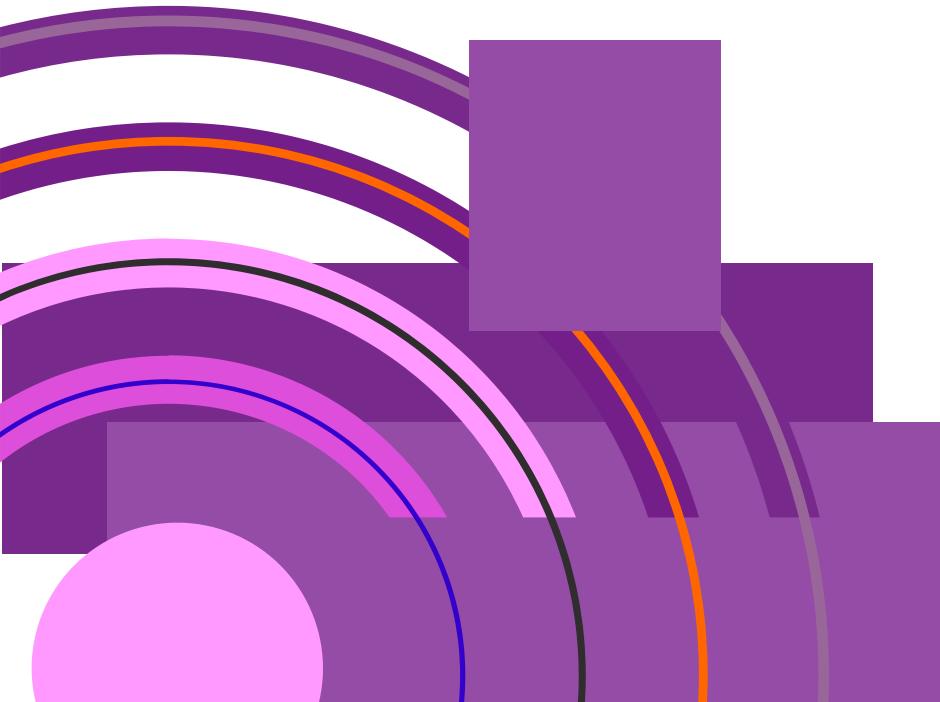


Estimation des coefficients avec la méthode des moments généralisés

La méthode des moments généralisés (MMG) est basée sur l'estimation des moments théoriques et observés du modèle.

Les moments théoriques sont des fonctions des paramètres inconnus du modèle.

Les moments observés sont des estimations des moments à partir des données.



Formulation de la méthode des moments généralisés

Soit le modèle économétrique : $Y = g(\beta, X) + U$

β représente les paramètres inconnus du modèle.

Les moments théoriques sont définis comme $E[h(Y, X, \beta)] = 0$, où $h(\cdot)$ est une fonction des variables et des paramètres.

Les moments observés sont estimés à partir des données.

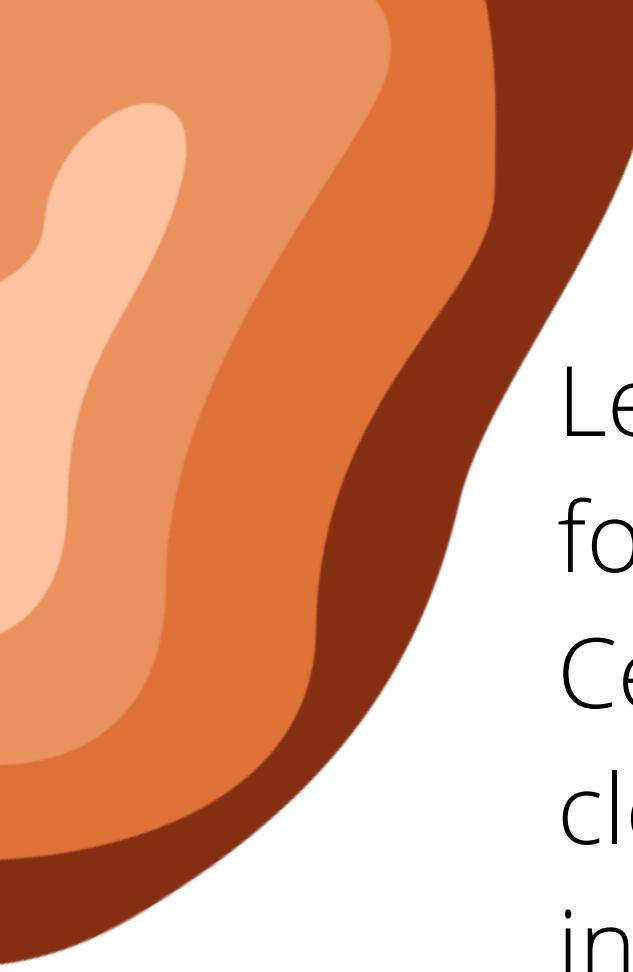


Résolution du modèle avec la méthode des moments généralisés

La MMG résout le modèle en minimisant la fonction d'objectif

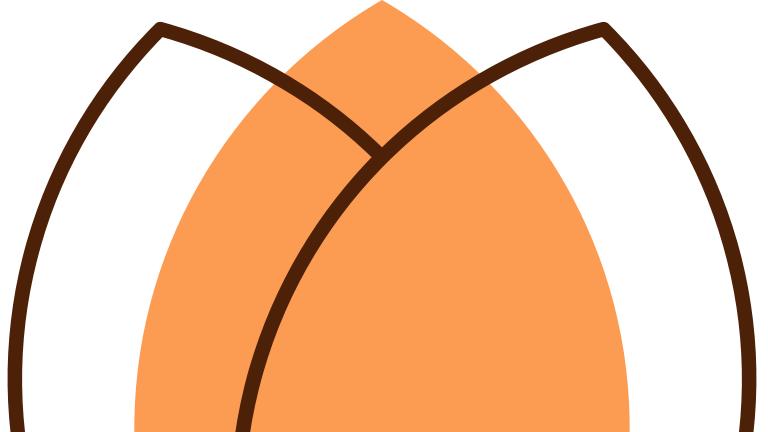
$$Q(\beta) = (1/N) * \sum [h(Y_i, X_i, \beta)]' W [h(Y_i, X_i, \beta)]$$

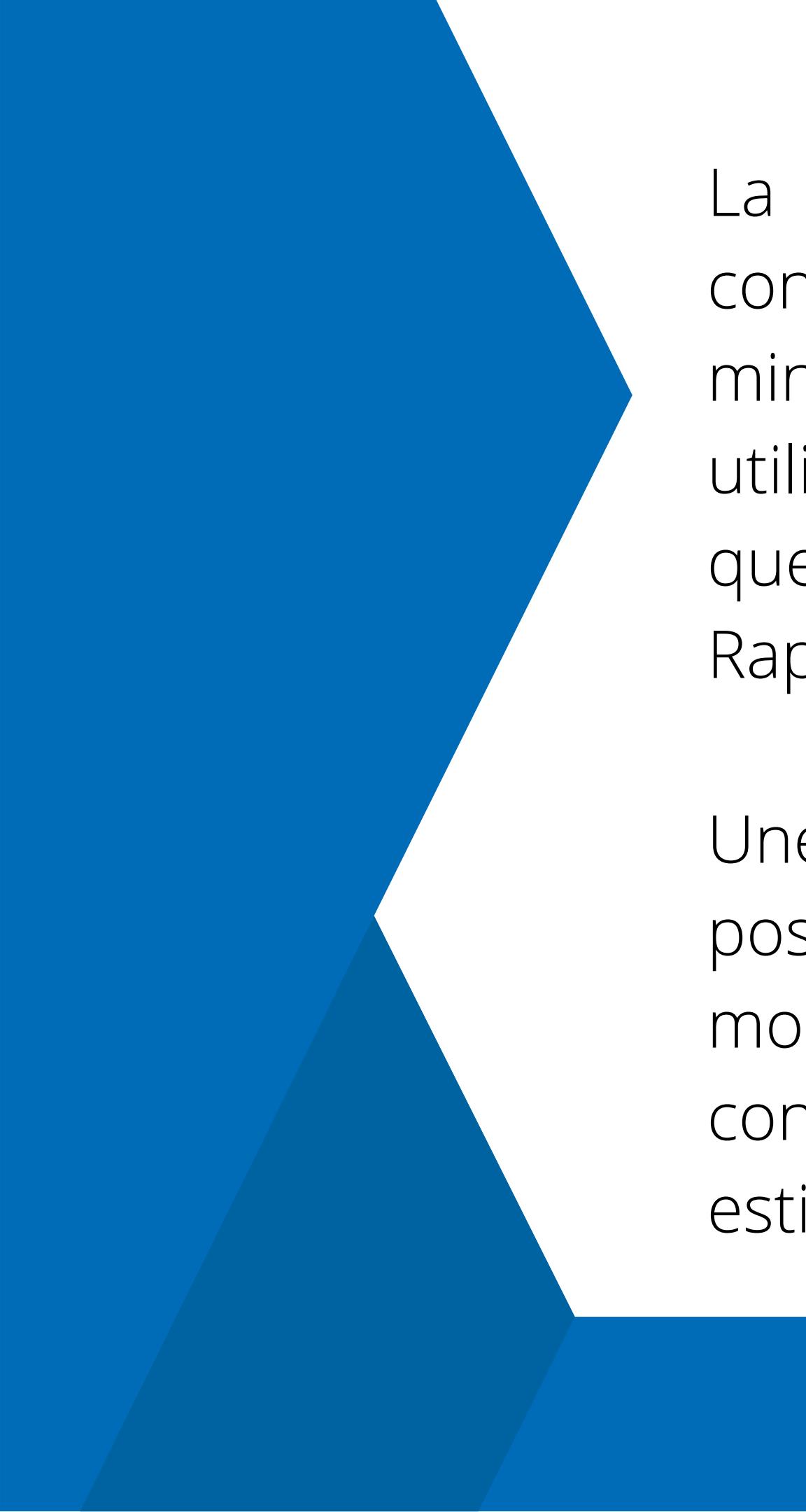
N est le nombre d'observations, $h(Y_i, X_i, \beta)$ est le vecteur des moments théoriques et la méthode des moments généralisés permet de résoudre le modèle en minimisant la fonction d'objectif $Q(\beta)$, qui est définie comme la moyenne des moments théoriques pondérés par la matrice de poids W.



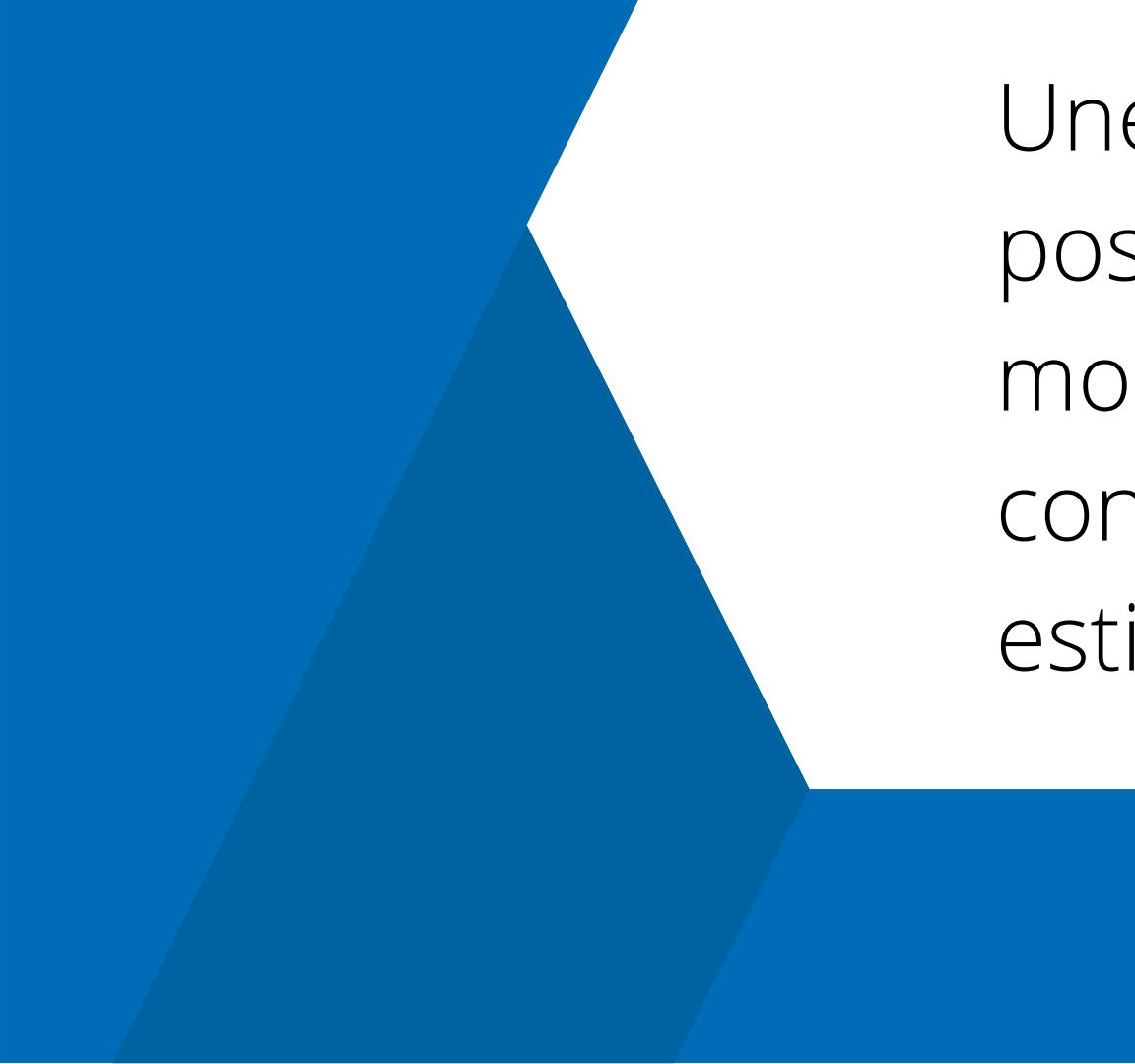
Le vecteur $h(Y_i, X_i, \beta)$ représente les moments théoriques, qui sont des fonctions des variables observées Y_i , X_i et des coefficients β à estimer. Ces moments sont choisis de manière à capturer les caractéristiques clés du modèle et à fournir des informations sur les paramètres inconnus.

La matrice de poids W est une matrice symétrique positive semi-définie qui permet de spécifier les pondérations pour chaque observation. Elle peut être utilisée pour donner plus de poids aux observations considérées comme plus fiables ou pour prendre en compte des structures de covariance spécifiques.





La solution de la méthode des moments généralisés consiste à trouver les estimations des coefficients β qui minimisent la fonction d'objectif $Q(\beta)$. Cela peut être fait en utilisant des techniques d'optimisation numérique, telles que la méthode du gradient ou la méthode de Newton-Raphson.



Une fois les estimations des coefficients obtenues, il est possible de tester l'hypothèse nulle sur les paramètres du modèle, d'évaluer la significativité des coefficients et de construire des intervalles de confiance pour les estimations.

Il est important de noter que la méthode des moments généralisés est une approche flexible qui peut être utilisée pour estimer des modèles avec des spécifications complexes, notamment lorsque les hypothèses classiques de régression linéaire ne sont pas entièrement satisfaites. Cependant, la mise en œuvre de cette méthode nécessite une attention particulière à la spécification des moments théoriques et à la sélection appropriée de la matrice de poids.



Exemple de résolution du modèle avec la méthode des moments généralisés

Considérons un modèle économétrique avec une variable endogène Y , une variable explicative X , et une variable instrumentale Z :

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + U$$

Pour estimer les coefficients β_0 , β_1 et β_2 , nous utilisons la méthode des moments généralisés en suivant les étapes suivantes :

1- Spécification des moments théoriques :

- Moment 1 : $E[XU] = 0$ (moment d'endogénéité)
- Moment 2 : $E[ZU] = 0$ (moment d'exogénéité des instruments)
- Moment 3 : $E[X^2] - E[X]E[X] = 0$ (moment d'orthogonalité des instruments)

2- Estimation des moments observés :

- Calcul des moments empiriques à partir des données disponibles.

3- Formation de la fonction d'objectif :

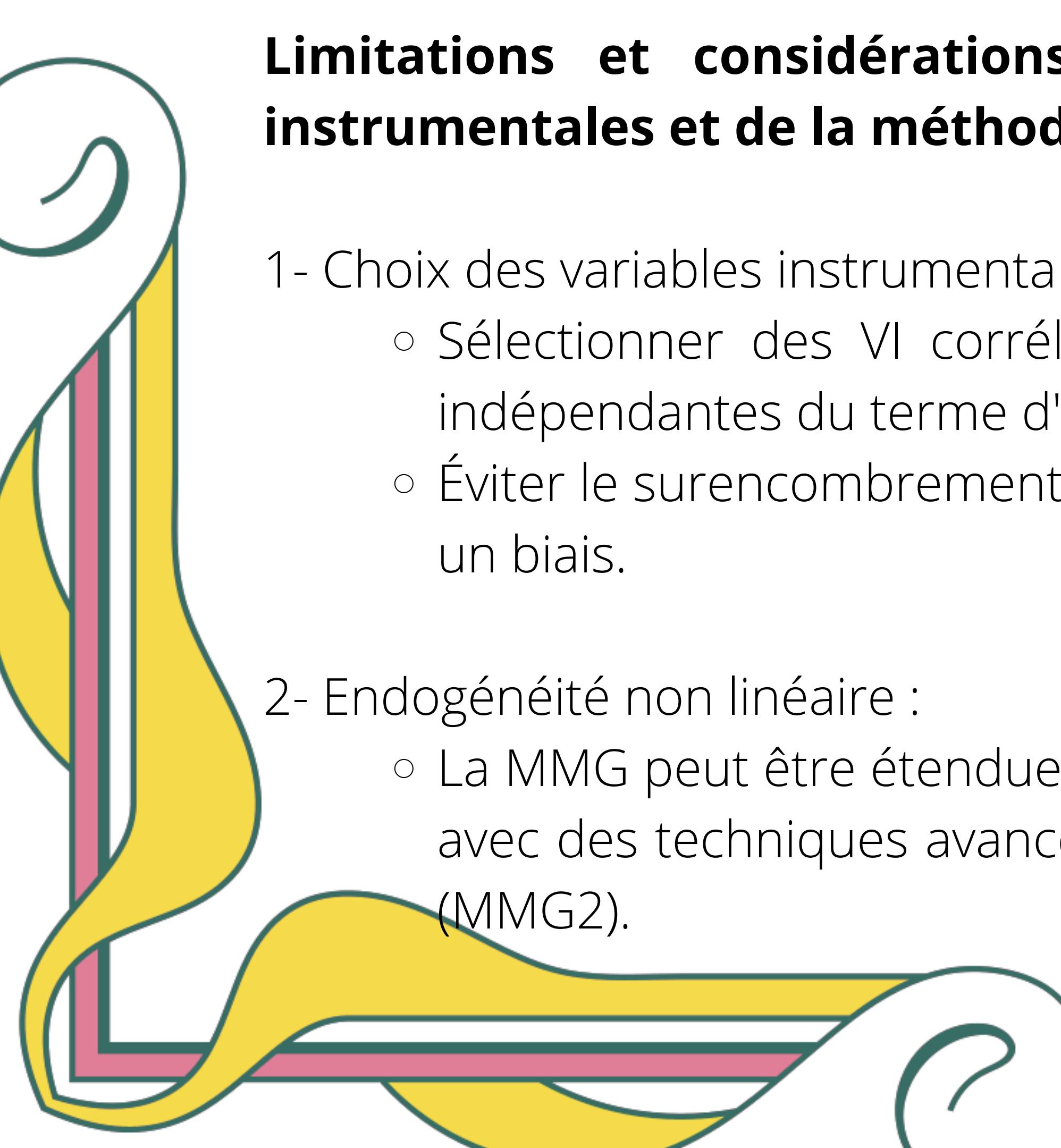
- $$Q(\beta) = (1/N) * \sum [h(Y_i, X_i, Z_i, \beta)]^T W [h(Y_i, X_i, Z_i, \beta)]$$

4- Minimisation de la fonction d'objectif :

- Utilisation des techniques d'optimisation pour trouver les coefficients β qui minimisent $Q(\beta)$.

5- Estimation des coefficients :

- Une fois que les coefficients β sont estimés, ils fournissent une estimation des relations entre les variables dans le modèle.



Limitations et considérations dans l'utilisation de variables instrumentales et de la méthode des moments généralisés

1- Choix des variables instrumentales :

- Sélectionner des VI corrélées avec la variable endogène, mais indépendantes du terme d'erreur.
- Éviter le surencombrement avec trop de VI qui peuvent introduire un biais.

2- Endogénéité non linéaire :

- La MMG peut être étendue pour traiter l'endogénéité non linéaire avec des techniques avancées telles que la MMG en deux étapes (MMG2).

3- Validité des instruments :

- Tester la validité des VI pour vérifier qu'elles satisfont les hypothèses requises et qu'elles ne sont pas faussées ou faibles.

4- Problème de surexclusion :

- L'utilisation excessive de VI peut conduire à un problème de surexclusion, où des instruments supplémentaires peuvent introduire un biais dans les estimations.



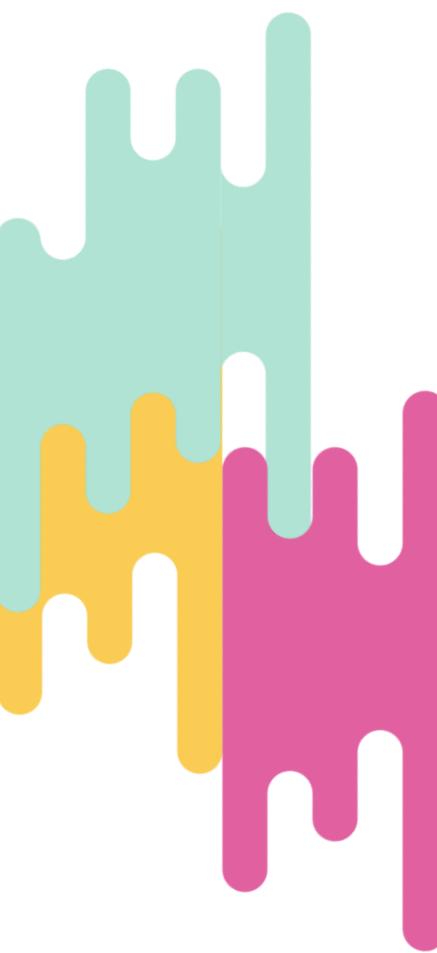
Illustration de l'utilisation de variables instrumentales et de la méthode des moments généralisés

Considérons un exemple où nous souhaitons estimer l'impact de l'éducation (X) sur les revenus (Y) en utilisant une variable instrumentale (Z) pour traiter l'endogénéité de X :

- Modèle : $Y = \beta_0 + \beta_1 X + \beta_2 Z + U$
- Variable endogène : X (corrélée avec U)
- Variable instrumentale : Z (corrélée avec X , indépendante de U)

L'estimation des coefficients β_0 , β_1 et β_2 est réalisée en utilisant la méthode des moments généralisés avec les moments théoriques et observés spécifiés précédemment.





Exemple avec Python

Présentation d'un exemple de mise en œuvre de la MMG et de

Logiciel Python pour MMG et VI

Python offre des packages tels que statsmodels et linearmodels qui peuvent être utilisés pour effectuer l'estimation par variables instrumentales et la méthode des moments généralisés.

Code Python pour la méthode VI

Démonstration de l'utilisation de Python pour l'estimation par variable instrumentale avec un exemple de code. Nous expliquons comment importer les données, comment effectuer l'estimation par variable instrumentale et comment interpréter les résultats.

Formule de code :

python

 Copy code

```
from linearmodels.iv import IV2SLS  
modèle = IV2SLS.from_formula('y ~ 1 + [x ~ z]', data).fit()  
print(model)
```

Code Python pour la méthode MMG

Démonstration de l'utilisation de Python pour la méthode des moments généralisés avec un exemple de code. Nous expliquons comment importer les données, comment effectuer l'estimation MMG et comment interpréter les résultats.

Formule de code :

```
python
from linearmodels.iv import GMM
modèle = GMM.from_formula('y ~ 1 + [x ~ z]', data).fit()
print(modèle)
```

Interprétation des résultats

Après avoir estimé les modèles, nous allons décrire comment interpréter les résultats de la sortie de Python, y compris les coefficients de régression, les statistiques de test et les diagnostics du modèle.

Test de validité de l'instrument

Présentation du test de Sargan (ou d'Hansen) pour vérifier la validité de l'instrument. Nous expliquons comment effectuer ce test en Python et comment interpréter les résultats.

Partie 7:

Série temporelle

Introduction aux séries temporelles

Les séries temporelles sont des ensembles de données organisées chronologiquement, où chaque observation est associée à un moment spécifique dans le temps. Elles sont souvent utilisées pour analyser et prévoir des données qui évoluent dans le temps, telles que les ventes mensuelles, les prix des actions, les températures quotidiennes, etc.

Les caractéristiques des séries temporelles comprennent la tendance, la saisonnalité et l'autocorrélation. La tendance représente la direction générale des données sur une période de temps. La saisonnalité se réfère à des variations périodiques régulières qui se répètent à intervalles fixes. L'autocorrélation indique la dépendance entre les observations successives.

Analyse exploratoire des séries temporelles

Avant de modéliser une série temporelle, il est important de procéder à une analyse exploratoire pour comprendre ses caractéristiques et ses comportements. Quelques outils couramment utilisés dans l'analyse exploratoire des séries temporelles sont les suivants :



- Graphiques de la série temporelle : Ils permettent de visualiser les tendances, les motifs saisonniers et les fluctuations aléatoires dans les données.
- Décomposition de la série : Elle permet de séparer la série en ses composantes de tendance, saisonnalité et résidus afin d'analyser séparément chaque composante.
- Autocorrélation et autocorrélogramme : Ils permettent d'identifier les dépendances entre les observations à différents retards temporels.
- Tests de stationnarité : Ils vérifient si les propriétés statistiques de la série ne varient pas au fil du temps.

Modèles ARIMA (AutoRegressive Integrated Moving Average)

Les modèles ARIMA sont largement utilisés pour modéliser les séries temporelles. Ils combinent les composantes auto-régressives (AR) et les composantes à moyenne mobile (MA) avec une différenciation intégrée (I) pour rendre la série stationnaire.

La notation générale d'un modèle ARIMA est ARIMA(p, d, q), où p est l'ordre de la composante AR, d est l'ordre de différenciation et q est l'ordre de la composante MA.

La modélisation ARIMA implique l'estimation des coefficients du modèle à l'aide de méthodes telles que la méthode des moindres carrés ou l'estimation de la vraisemblance. Les critères d'information tels que le critère d'information d'Akaike (AIC) ou le critère d'information bayésien (BIC) peuvent être utilisés pour sélectionner le meilleur modèle ARIMA.

L'équation d'un modèle ARIMA(p, d, q) est :

$$Y_t = \alpha + \sum(\varphi_i Y_{t-i}) + \sum(\theta_j \varepsilon_{t-j}) + \varepsilon_t$$

où Y_t est la valeur de la série temporelle à l'instant t , α est la constante, φ_i sont les coefficients auto-régressifs, θ_j sont les coefficients de la moyenne mobile, ε_t est le terme d'erreur à l'instant t .

Modèles ARCH/GARCH (AutoRegressive Conditional Heteroskedasticity/Generalized ARCH)

Les modèles ARCH/GARCH sont utilisés pour modéliser la volatilité hétéroscléastique conditionnelle dans les séries temporelles financières et économiques. Ils permettent de capturer les changements dans la variance des résidus au fil du temps.

Les modèles ARCH/GARCH utilisent des équations spécifiques pour modéliser la volatilité conditionnelle. Un modèle ARCH(p) est défini par l'équation : $\sigma_t^2 = \omega + \sum(a_i \varepsilon_{t-i}^2)$ où σ_t^2 est la variance conditionnelle à l'instant t, ω est la constante, a_i sont les coefficients ARCH et ε_t est le terme d'erreur.

Un modèle GARCH(p, q) étend le modèle ARCH en incorporant également des termes de moyenne mobile. L'équation d'un modèle GARCH(p, q) est :

$$\sigma_t^2 = \omega + \sum(\alpha_i \varepsilon_{t-i}^2) + \sum(\beta_j \sigma_{t-j}^2)$$

où β_j sont les coefficients GARCH.

L'estimation des coefficients dans les modèles ARCH/GARCH peut être effectuée à l'aide de méthodes telles que la maximisation de la vraisemblance ou la méthode des moindres carrés généralisés.

Prévisions dans les séries temporelles

Une fois que nous avons construit un modèle pour une série temporelle, nous pouvons l'utiliser pour effectuer des prévisions. Les méthodes couramment utilisées pour les prévisions dans les séries temporelles comprennent :

- Prévisions naïves : Elles consistent à utiliser la dernière observation comme prévision pour les périodes futures. Cette méthode est simple mais ne prend pas en compte les modèles ou les tendances.



- Méthodes de moyenne mobile : Elles utilisent les moyennes des observations passées pour effectuer les prévisions.
- Modèles ARIMA : Les modèles ARIMA peuvent être utilisés pour générer des prévisions en extrapolant les tendances et les motifs présents dans les données.
- Modèles ARCH/GARCH : Les modèles ARCH/GARCH peuvent également être utilisés pour prévoir la volatilité future des séries temporelles financières.





Les séries temporelles sont un domaine important de l'analyse des données, permettant de modéliser et de prévoir les données qui évoluent dans le temps. La compréhension des concepts clés tels que la tendance, la saisonnalité, l'autocorrélation et la volatilité conditionnelle est essentielle pour une modélisation efficace des séries temporelles.

Les modèles ARIMA, ARCH/GARCH et d'autres méthodes offrent des outils puissants pour analyser et prévoir les séries temporelles. L'analyse exploratoire, l'estimation des coefficients et les prévisions sont des étapes essentielles dans l'analyse des séries temporelles.



Validation des modèles et évaluation des prévisions

Une fois que nous avons construit un modèle de série temporelle et effectué des prévisions, il est crucial de les valider et d'évaluer leur performance. Quelques méthodes courantes pour la validation des modèles et l'évaluation des prévisions sont les suivantes :

- Validation croisée : Elle consiste à diviser les données en ensembles d'apprentissage et de test, en ajustant le modèle sur l'ensemble d'apprentissage et en évaluant les prévisions sur l'ensemble de test. Cela permet d'estimer la performance du modèle sur des données non utilisées lors de l'ajustement.

Validation des modèles et évaluation des prévisions

Une fois que nous avons construit un modèle de série temporelle et effectué des prévisions, il est crucial de les valider et d'évaluer leur performance. Quelques méthodes courantes pour la validation des modèles et l'évaluation des prévisions sont les suivantes :

- Validation croisée : Elle consiste à diviser les données en ensembles d'apprentissage et de test, en ajustant le modèle sur l'ensemble d'apprentissage et en évaluant les prévisions sur l'ensemble de test. Cela permet d'estimer la performance du modèle sur des données non utilisées lors de l'ajustement.

- Erreurs de prévision : Les erreurs de prévision, telles que l'erreur quadratique moyenne (RMSE) ou l'erreur absolue moyenne (MAE), mesurent l'écart entre les prévisions et les observations réelles. Des erreurs plus faibles indiquent une meilleure performance du modèle.
- Comparaison des modèles : Lorsque plusieurs modèles sont considérés, il est important de les comparer à l'aide de critères tels que l'AIC, le BIC ou le critère de log-vraisemblance afin de sélectionner le modèle le plus approprié.
- Tests de diagnostic : Les tests de diagnostic vérifient si les résidus du modèle présentent des motifs systématiques ou des violations des hypothèses sous-jacentes. Des résidus sans autocorrélation significative et sans structure particulière indiquent un bon ajustement du modèle.

Applications des séries temporelles

Les séries temporelles ont de nombreuses applications dans divers domaines :

- Finance : Prévisions des prix des actions, modélisation de la volatilité financière, prévisions des rendements.
- Économie : Analyse des cycles économiques, prévisions des indicateurs économiques, modélisation de l'inflation.

- Sciences environnementales : Modélisation des données météorologiques, prévisions des niveaux d'eau, analyse de la pollution atmosphérique.
- Marketing : Prévisions des ventes, analyse des tendances saisonnières, évaluation de l'impact des campagnes publicitaires.
- Santé : Prévisions des admissions à l'hôpital, modélisation des épidémies, analyse des tendances des données médicales.

Les séries temporelles fournissent des outils essentiels pour analyser et prévoir des données qui évoluent dans le temps. En comprenant les caractéristiques des séries temporelles, en utilisant des modèles appropriés tels que ARIMA et ARCH/GARCH, et en validant les modèles et les prévisions, nous pouvons obtenir des informations précieuses pour la prise de décision dans de nombreux domaines.

Il est important de rappeler que la modélisation des séries temporelles est un processus itératif qui nécessite une analyse approfondie, des ajustements de modèles et une évaluation rigoureuse. En exploitant les avantages des séries temporelles, nous pouvons obtenir des prévisions précises, identifier les tendances et les motifs clés, et prendre des décisions éclairées basées sur les données temporelles.



Structure des données

Il existe 3 types de données. Chaque type de données peut appeler des techniques économétriques particulières.

- Données “Cross-section”.
- Séries temporelles.
- Données de Panel.

Données “Cross-section”

- Échantillon d'individus, ménages, firmes, ..., pris à un point du temps donné.
- Important: on peut souvent supposer que les obs. = échantillon aléatoire → simplifie l'analyse.
- Données très utilisées en économie et sciences sociales → micro appliquée: marché du travail, finances publiques, organisation industrielle, économie spatiale, démographie, économie de la santé, etc.
- Example: Wage1.wf1.

Séries temporelles

- Séries chronologiques. Ex: PNB, importations, indices de prix, etc.
- Important: Rarement indépendantes au court du temps
→ complexifie l'analyse.
- Différentes fréquences: annuel, trimestriel, mensuel, hebdomadaire, journalier, intra-journalier.
- Données très utilisées en macro-économie et en finance.
- Example: PRMINWGE.wf1.

Données de panel

- Série temporelle pour chaque unité/individu.
- Important: la même unité est observée plusieurs fois au court du temps.
- Example: **WAGEPAN.wf1**.

Rappels Modèle de régression simple: Chapitre 2

Modèle de régression simple

- Objectif: estimer un modèle du type

$$wage = \beta_0 + \beta_1 educ + u. \quad (1)$$

- De manière générale:

$$y = \beta_0 + \beta_1 x + u. \quad (2)$$

- (?) est supposé tenir sur la population d'intérêt.
- u est le terme d'erreur (aléas) = facteurs *non-observés* autres que x qui affectent y .
- u et x sont des variables aléatoires.
- Interprétation *Ceteris Paribus*: Si $\Delta u = 0$ (autres facteurs inchangés) $\rightarrow \Delta y = \beta_1 \Delta x$.

Modèle de régression simple

- Pour estimer β_0 et β_1 et garder cette interprétation CP il faut faire certaines hypothèses.
- $E(u) = 0$: normalisation \rightarrow on ne perd rien.
- $E(u|x) = E(u)$: pour toute valeur de x , la moyenne des u correspondantes est la même.
 \rightarrow implique la non-corrélation (linéaire). Ex:
 $wage = \beta_0 + \beta_1 educ + u \rightarrow E(abil|8) = E(abil|16)$.
- En combinant ces 2 hypothèses:
 $\rightarrow E(u|x) = E(u) = 0$: hyp. moyenne cond. nulle.
 $\rightarrow E(y|x) = \beta_0 + \beta_1 x$: fonction de régression de la population.

Estimation de β_0 et β_1

$(x_i, y_i) : i = 1, \dots, n$ = échantillon aléatoire de taille n tiré de la population.

- $y_i = \beta_0 + \beta_1 x_i + u_i, \forall i.$
- $E(u) = 0 \rightarrow E(y - \beta_0 - \beta_1 x) = 0.$
- $E(u|x) = 0 \rightarrow \text{Cov. nulle entre } u \text{ et } x$
 $\rightarrow E[(y - \beta_0 - \beta_1 x)x] = 0.$
- Pour obtenir $\hat{\beta}_0$ et $\hat{\beta}_1$ on va résoudre ce système en remplaçant $E(\cdot)$ par son équivalent empirique $1/n \sum_{i=1}^n (\cdot).$

Estimation de β_0 et β_1

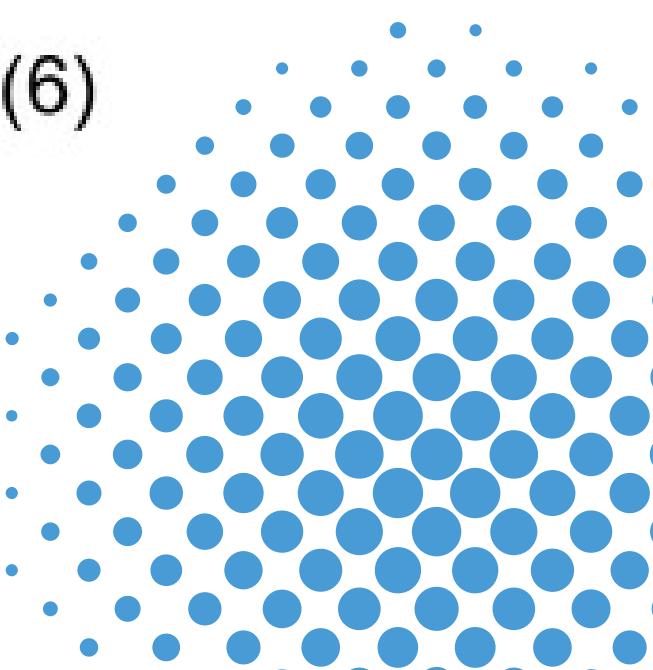
$$n^{-1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (3)$$

$$n^{-1} \sum_{i=1}^n [(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i] = 0 \quad (4)$$

$$\rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (5)$$

$$\rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ si } > 0. \quad (6)$$

Estimation de β_0 et β_1 par la méthode des moments.



Moindres carrés ordinaires (MCO)

$\min_{b_0, b_1} \sum_{i=1}^n (y_i - b_0 - b_1 x_i)^2 \rightarrow$ 2 équations “de premier ordre”:

$$-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (7)$$

$$-2 \sum_{i=1}^n [(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) x_i] = 0 \quad (8)$$

$$\hookrightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (9)$$

$$\hookrightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \text{ si } > 0. \quad (10)$$

Estimation de β_0 et β_1 par la méthode des MCO donne les mêmes $\hat{\beta}_0$ et $\hat{\beta}_1$.

Propriétés des estimateurs MCO

SLR=Simple Linear Regression.

- **SLR.1** $y = \beta_0 + \beta_1 x + u \rightarrow$ linéaire en les paramètres.
- **SLR.2** $(x_i, y_i) : i = 1, \dots, n \rightarrow$ échantillon aléatoire de taille n tiré de la population.
- **SLR.3** $E(u|x) = 0 \rightarrow$ moyenne conditionnelle nulle.
Permet de dériver les propriétés des MCO *conditionnellement* aux valeurs de x_i dans notre échantillon. Techniquement identique à supposer x_i fixes dans des échantillons répétés (pas très réaliste).
- **SLR.4** $\sum_{i=1}^n (x_i - \bar{x})^2 \neq 0 \rightarrow$ variation dans les x .
- **Théorème 2.1:** Sous les hypothèses SLR.1 - SLR.4,
 $E(\hat{\beta}_0) = \beta_0$ et $E(\hat{\beta}_1) = \beta_1$.

Propriétés des estimateurs MCO

- **SLR.5** $Var(u|x) = \sigma^2 \rightarrow$ homoscédasticité.
- **Théorème 2.2:** Sous les hypothèses SLR.1 - SLR.5,

$$Var(\hat{\beta}_0) = \frac{\sigma^2 n^{-1} \sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

- **Théorème 2.3:** Sous les hypothèses SLR.1 - SLR.5,
 $E(\hat{\sigma}^2) = \sigma^2$, où $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$.

Rappels Modèle de régression multiple: Chapitre 3

Modèle de régression multiple

- Objectif: estimer un modèle du type

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u. \quad (11)$$

- De manière générale:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u. \quad (12)$$

- OLS: $\min_{b_0, b_1, \dots, b_k} \sum_{i=1}^n (y_i - b_0 - b_1 x_{i1} - \dots - b_k x_{ik})^2.$

Modèle de régression multiple

Exemple: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u.$

→ Comment obtenir β_1 ?

- Régresser x_1 sur x_2 : $\hat{x}_1 = \hat{\alpha}_0 + \hat{\alpha}_1 x_2$.
- Calculer $\hat{r}_1 = x_1 - \hat{x}_1$.
- $\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$.
- β_1 est bien l'effet net de x_1 sur y , où net signifie après avoir tenu compte de l'effet des autres variables.

Goodness-of-Fit

Il est possible de décomposer la variabilité observée sur y , c-à-d $SST \equiv \sum_{i=1}^n (y_i - \bar{y})^2$, en 2 quantités:

- $SSE \equiv \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$: variabilité expliquée par le modèle;
- $SSR \equiv \sum_{i=1}^n \hat{u}_i^2$: variabilité **non**-expliquée par le modèle.

$$\rightarrow SST = SSE + SSR.$$

On peut donc définir une mesure de “qualité” de la régression (GoF):

$$R^2 = SSE/SST \text{ compris entre 0 et 1.}$$

Propriétés des estimateurs MC0

MLR=Multiple Linear Regression.

- **MLR.1** $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u \rightarrow$ linéaire en les paramètres.
- **MLR.2** $(x_{1i}, \dots, x_{ki}, y_i) : i = 1, \dots, n \rightarrow$ échantillon aléatoire de taille n tiré de la population.
- **MLR.3** $E(u|x_1, \dots, x_k) = 0 \rightarrow$ moyenne conditionnelle nulle.
- **MLR.4** $\sum_{i=1}^n (x_{ji} - \bar{x})^2 \neq 0 \forall j = 1, \dots, k$ et pas de relation linéaire parfaite entre les $x_j \rightarrow$ variation dans les x_j et pas de collinéarité parfaite.
- **Théorème 3.1:** Sous les hypothèses **MLR.1 - MLR.4**,
 $E(\hat{\beta}_j) = \beta_j \forall j = 0, \dots, k.$

Propriétés des estimateurs MCO

- **MLR.5** $Var(u|x_1, \dots, x_k) = \sigma^2 \rightarrow$ homoscédasticité.
- **Théorème 3.2:** Sous les hypothèses **MLR.1 - MLR.5**,

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{(1 - R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}$$

où R_j^2 est le R^2 de la régression de x_j sur les autres x (+ une constante).

- **Théorème 3.3:** Sous les hypothèses **MLR.1 - MLR.5** (appelées hypothèses de **Gauss-Markov**),
 $E(\hat{\sigma}^2) = \sigma^2$, où $\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^n \hat{u}_i^2$.
- **Théorème 3.4 où théorème de Gauss-Markov:** Sous les hypothèses **MLR.1 - MLR.5**, $\hat{\beta}_j$ est **BLUE**,
 $\forall i = 0, \dots, k$.

Rappels Inférence: Chapitre 4

Hypothèse de normalité

Pour faire de l'inférence statistique, il faut ajouter une hypothèse supplémentaire:

- **MLR.6** $u \sim N(0, \sigma^2)$ et indépendant des $x_j \rightarrow$ **normalité**.
- Les hypothèses **MLR.1 - MLR.6** sont appelées hypothèses *classiques* du modèle linéaire (CLM) = Gauss-Markov + normalité.
- CLM \rightarrow MCO à la variance minimale.
- Comment justifier cette hypothèse de normalité des résidus ?
- Si u est la somme de beaucoup de facteurs non-observés différents, on peut appeler le théorème central limite pour justifier la normalité.

Théorème Central Limite

- Soit Y_1, Y_2, \dots, Y_n un échantillon de variables aléatoires de moyenne μ et de variance σ^2 .
- $\rightarrow Z_n = \frac{\bar{Y}_n - \mu}{\sigma/\sqrt{n}}$ suit asymptotiquement une distribution $N(0,1)$.
- Si $Y \sim \chi^2(1)$, $\mu = 1$ et $\sigma^2 = 2$.
- Exemple.

Inférence

- **Théorème 4.1:** Sous les hypothèses **MLR.1 - MLR.6**,
 $\hat{\beta}_j \sim N(\beta_j, Var(\hat{\beta}_j)), \forall i = 0, \dots, k,$ où
 $Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2) \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2}.$
- **Exemple.**
- **Théorème 4.2:** Sous les hypothèses **MLR.1 - MLR.6**,
 $\frac{\hat{\beta}_j - \beta_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}, \forall i = 0, \dots, k,$ où
 $se(\hat{\beta}_j) = \sqrt{Var(\hat{\beta}_j)}$ et σ^2 est remplacé par $\hat{\sigma}^2.$
- On peut donc tester $H_0 : \beta_j = a_j$ contre
 $H_1 : \beta_j <, \neq, > a_j.$

P-valeur

- P-valeur: “Quelle est le plus petit niveau de significativité auquel H_0 serait rejeté”.
- Comment la calculer? Prendre la t-stat et regarder à quel pourcentile elle correspond dans la distribution de Student appropriée.
- → Rejeter H_0 si la P-Valeur < au seuil fixé (ex: 5%).

Tests multiples de restrictions linéaires

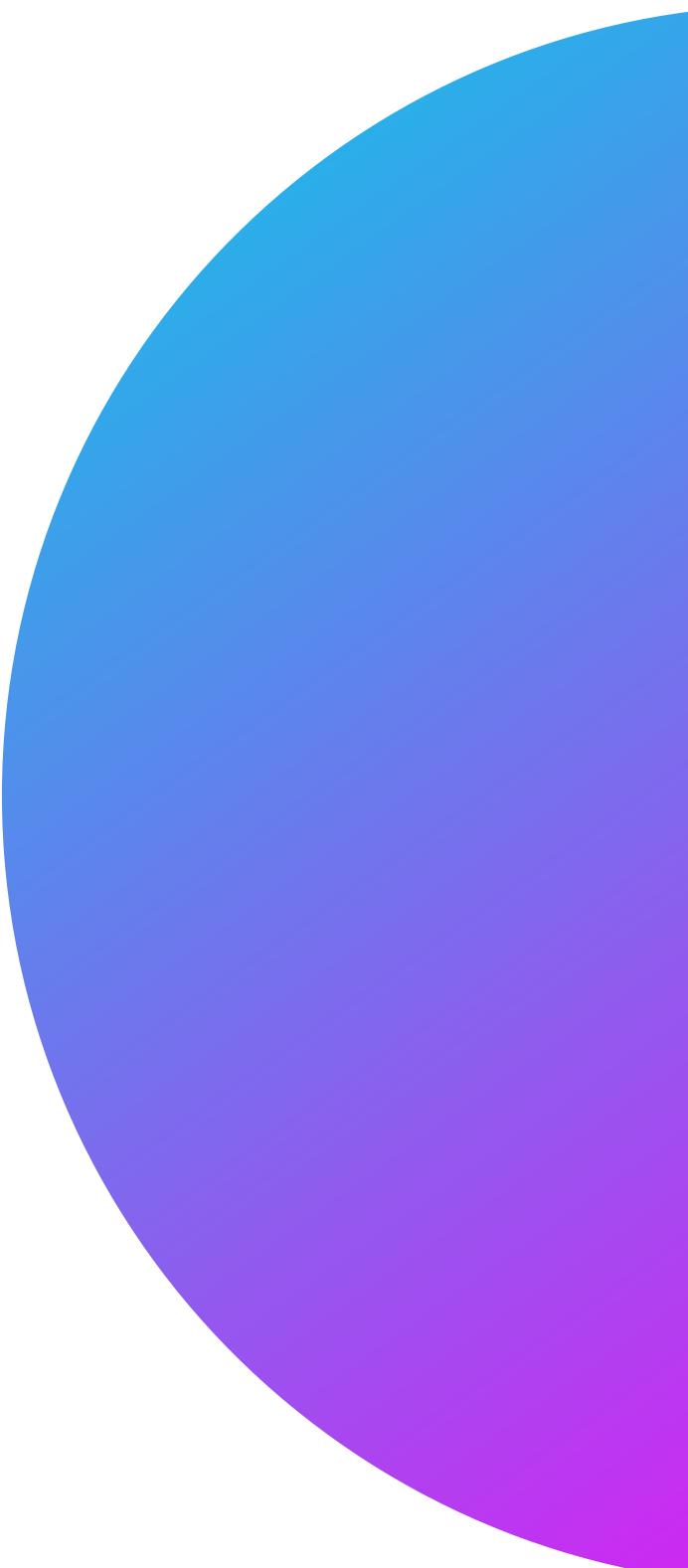
- Si $H_0 : \beta_j = 0$ pour $j \in 0, 1, \dots, k$. Ex: $H_0 : \beta_1 = \beta_2 = 0$.
- et $H_1 : H_0$ n'est pas vrai.
- → 1) Estimer le modèle non contraint.
- → 2) Estimer le modèle contraint.
- → 3) Calculer la statistique $F \equiv \frac{(SSR_c - SSR_{nc})/q}{SSR_{nc}/(n-k-1)}$, où SSR dénote la somme des carrés des résidus et les indices c et nc signifient respectivement constraint et non-constraint.
- Sous H_0 , $F \sim F_{q,n-k-1}$, où F = Fisher.
- **Rejet** H_0 , si $F > c_\alpha$, où $c = F_{q,n-k-1}^\alpha$.

Exercices récapitulatifs

- Exercice 4.12.
- Exercice 4.17.
- Exercice 4.19.



Rappels propriétés asymptotiques des MCO: Chapitre 5



Propriétés exactes des MCO

- Pour prouver le caractère non-biaisé et BLUE des MCO nous avons imposé des hypothèses assez fortes (**MLR.1-MLR.5**).
- Idem pour effectuer de l'inférence statistique (**MLR.6**: normalité).
- Ces propriétés sont vraies quel que soit la taille de l'échantillon ($\forall n$).
- → On parle dès lors de propriété exacte, d'échantillon fini, ou encore de petit échantillon.

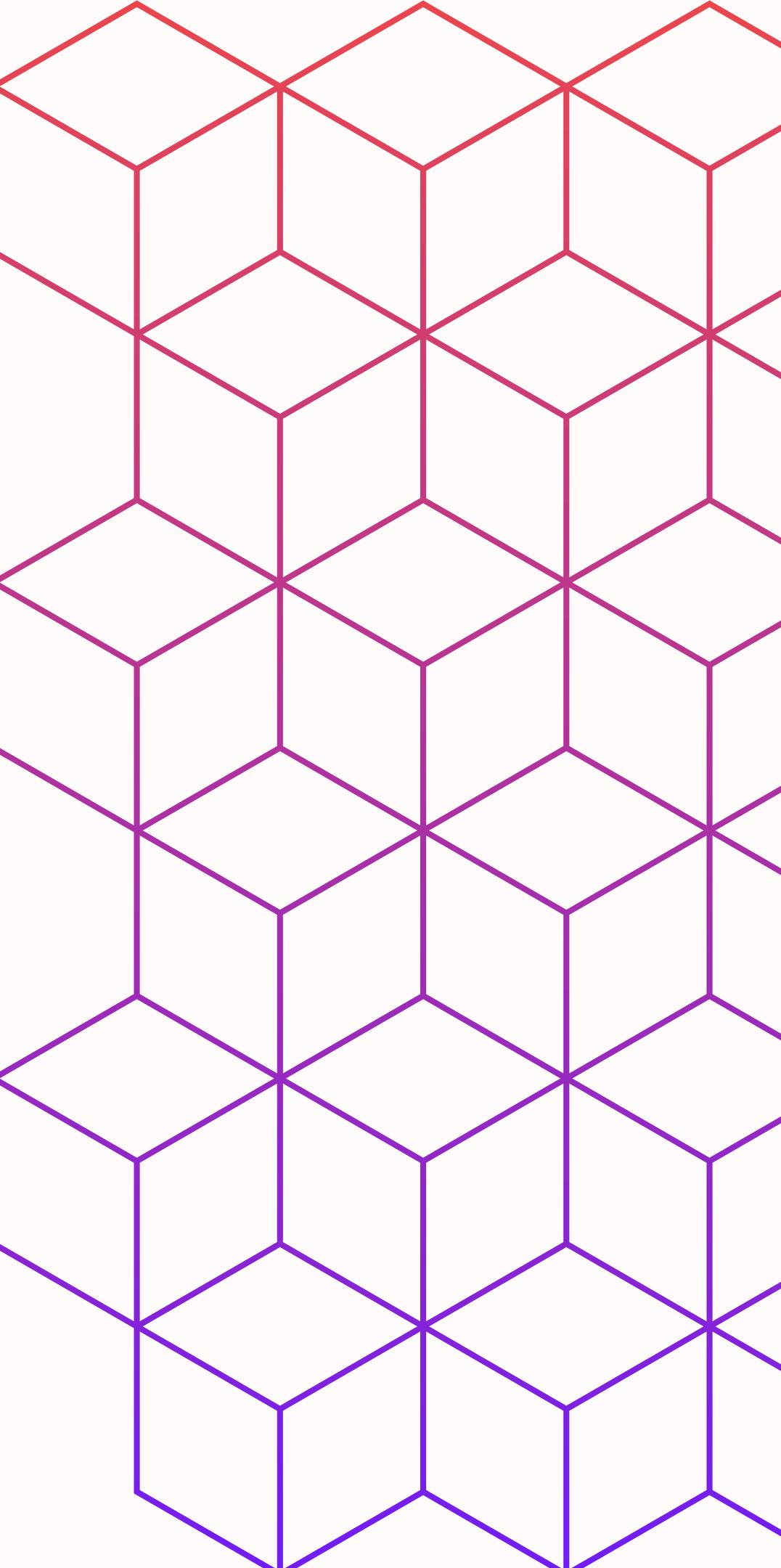
Propriétés asymptotiques des MCO

- Dans certains cas, le rejet de certaines hypothèses ne signifie pas que les MCO sont invalides.
- Ex: non-normalité de u .
- → En effet, les MCO peuvent être encore valides en grand échantillon *sous des hypothèses plus faibles*.
- → Étudier les propriétés statistiques pour n grand = étudier les propriétés asymptotiques.
- Nouveau concept: CONVERGENCE.

Propriétés asymptotiques des MCO

- Pour rappel. **Théorème 3.1:** Sous les hypothèses **MLR.1 - MLR.4**, $E(\hat{\beta}_j) = \beta_k \forall j = 0, \dots, k.$
- Pour rappel. **MLR.3** $E(u|x_1, \dots, x_k) = 0 \rightarrow$ **moyenne conditionnelle nulle**.
- On a vu que $E(u|x_1) = 0$ implique $Cov(u, x_1) = 0$.
- **Définition:** *plim = Limite en probabilité = valeur vers laquelle un estimateur converge lorsque la taille de l'échantillon tend vers l'infini.* Voir page 741.
- Hors, il est possible de montrer que:

$$\begin{aligned} \text{plim } \hat{\beta}_1 &= \beta_1 + \frac{Cov(x_1, u)}{Var(x_1)} \\ &= \beta_1, \text{ si } Cov(u, x_1) = 0. \end{aligned}$$



Propriétés asymptotiques des MCO

- → Il est possible de relâcher **MLR.3** pour prouver tout de même que les MCO sont convergents.
- **MLR.3'** $E(u) = 0$ et $Cov(x_j, u) = 0 \quad \forall j = 1, \dots, k \rightarrow$ moyenne nulle et correlation nulle.
- → Sous les hypothèses **MLR.1 - MLR.2 - MLR.3' - MLR.4**, $\hat{\beta}_j$ est un estimateur convergent de β_j ,
 $\forall j = 1, \dots, k$.

Normalité asymptotique

- La normalité ne joue aucun rôle dans le caractère non-biasés des MCO ni dans leur caractère BLUE.
- Par contre, pour effectuer de l'inférence statistique nous avons supposé que $u \sim N(0, \sigma^2)$ et donc que la distribution de $y|x_1, \dots, x_k$ est normale.
 - → distribution symétrique autour de sa moyenne.
 - → peut prendre des valeurs sur \mathbb{R} .
 - → plus de 95% des observations sont comprises entre 2 écart-types.

Normalité asymptotique

- Devons nous abandonner les t-stats si u n'est pas normalement distribué ?
- Non: on fait appel au théorème central limite pour conclure que les MCO satisfont la normalité asymptotique.
- Si n grand, $\hat{\beta}_j$ est *approximativement* $N(\beta_j, Var(\hat{\beta}_j))$.
- Illustration via une Simulation de Monte-Carlo: **Prg, Output n=2000, Output n=30.**

Rappels Hétéroscédasticité: Chapitre 8

Conséquences de l'hétéroscédasticité

- Une fois de plus l'hétéroscédasticité n'a pas d'influence sur le caractère non-biaisé ou convergent des MCO.
- Par contre la formule traditionnelle de $Var(\hat{\beta}_j)$ donne un estimateur biaisé de la variance des estimateurs si l'hypothèse d'homoscédasticité est violée.
 - → Inférence incorrecte si on utilise cette formule.
 - → MCO ne sont plus "BLUE".
- *Solution 1:* Corriger cette formule → **Écart-types robustes à l'hétéroscédasticité** (White, 1980).
- *Solution 2:* Déterminer la source de cette hétéroscédasticité et la prendre en compte dans l'estimation → **Moindres Carrés Pondérés**.

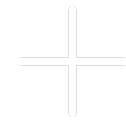
Régression avec des séries temporelles: Chapitre 10

Séries temporelles

Les séries chronologiques diffèrent des séries “cross-section” pour plusieurs raisons:

- → l'ordre (temporel) importe;
- → le passé influence souvent le futur;
- → la notion d'échantillon aléatoire est plus discutable car on n'a qu'une seule réalisation (sauf si on pense que des conditions initiales \neq auraient donné une réalisation \neq).
- Example: Phillips.wf1.
- Question: Peut-on toujours utiliser les MCO avec ces séries ?

Séries temporelles



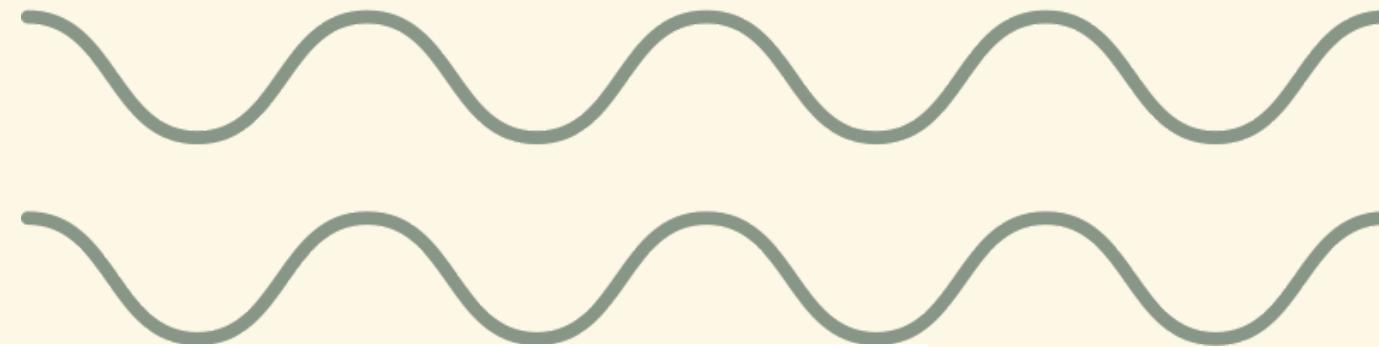
Il y a deux types principaux de modèles de séries temporelles simples:

- → les modèles statiques du type:

$$y_t = \beta_0 + \beta_1 z_t + u_t, t = 1, \dots, n.$$

- → les modèles dynamiques du type:

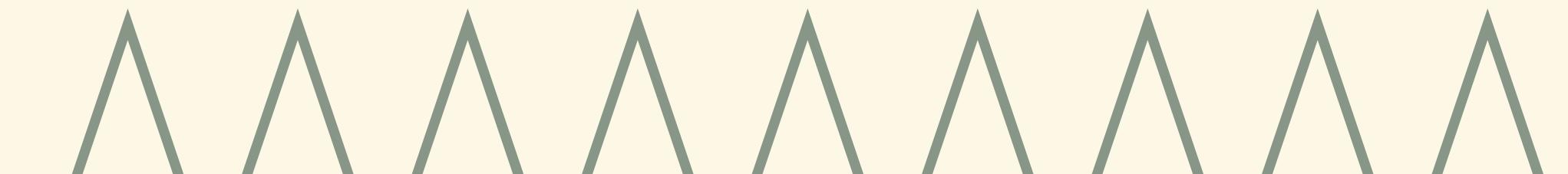
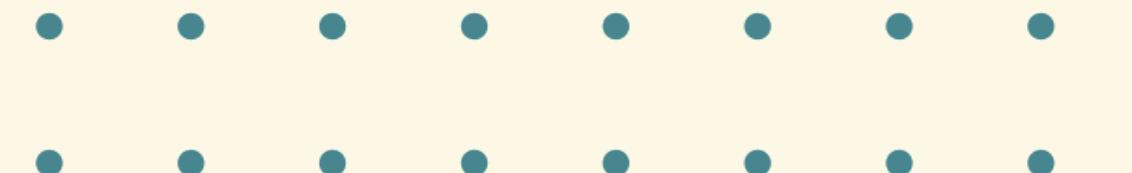
$$y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \dots + \delta_q z_{t-q} + u_t, t = 1, \dots, n.$$



Modèles statiques

$$y_t = \beta_0 + \beta_1 z_t + u_t, t = 1, \dots, n.$$

- $\Delta y_t = \beta_1 \Delta z_t$ si $\Delta u_t = 0 \rightarrow$ **Effet immédiat.**
- Example 1. Courbe de Phillips (annuel):
 $inft = \beta_0 + \beta_1 unemt + u_t.$
- Example 2. Consommation de tabac (mensuel):
 $ConsTabact = \mu_0 + \mu_1 DepPubl_t + u_t.$



Modèles dynamiques

FDL(2): $y_t = \alpha_0 + \delta_0 z_t + \delta_1 z_{t-1} + \delta_2 z_{t-2} + u_t, t = 1, \dots, n.$

- Supposons que $z = c, \forall$ observation avant t .
- En $t \rightarrow z = c + 1$.
- En $t + 1, t + 2, \dots, n \rightarrow z = c$.
- Pour simplifier supposons que $u_t = 0$.

Modèles dynamiques: Δ temporaire

$$\begin{aligned}y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c \\y_t &= \alpha_0 + \delta_0(c+1) + \delta_1 c + \delta_2 c \\y_{t+1} &= \alpha_0 + \delta_0 c + \delta_1(c+1) + \delta_2 c \\y_{t+2} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2(c+1) \\y_{t+3} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c\end{aligned}$$

Multiplicateurs d'impact

Après un changement temporaire d'une unité de z (c-à-d $\Delta z = 1$)

→ $y_t - y_{t-1} = \delta_0$ = effet immédiat sur y .

→ $y_{t+1} - y_{t-1} = \delta_1$ = effet dans 1 période sur y .

→ $y_{t+2} - y_{t-1} = \delta_2$ = effet dans 2 périodes sur y .

Modèles dynamiques: Δ permanente

$$\begin{aligned}y_{t-1} &= \alpha_0 + \delta_0 c + \delta_1 c + \delta_2 c \\y_t &= \alpha_0 + \delta_0(c+1) + \delta_1 c + \delta_2 c \\y_{t+1} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2 c \\y_{t+2} &= \alpha_0 + \delta_0(c+1) + \delta_1(c+1) + \delta_2(c+1)\end{aligned}$$

Multiplicateur de long-terme

Après un changement permanent d'une unité de z (c-à-d $\Delta z = 1$)

→ $y_t - y_{t-1} = \delta_0$ = effet immédiat sur y .

→ $y_{t+1} - y_{t-1} = \delta_0 + \delta_1$ = effet dans 1 période sur y .

→ $y_{t+2} - y_{t-1} = \delta_0 + \delta_1 + \delta_2$ = effet dans 2 périodes sur y .

→ $\sum_{i=0}^q \delta_i$ = multiplicateur de long-terme.

Propriétés des MCO

Nous allons montrer que les MCO sont encore valables avec des séries temporelles et faire le parallèle avec les séries “cross-section”. **Hypothèses de base:**

- **MLR.1** $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u \rightarrow$ linéaire en les paramètres.



TS.1 $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t \rightarrow$ linéaire en les paramètres. Ex: $x_{t3} = \text{salaire}_{t-2}$.

- **MLR.2** $(x_{1i}, \dots, x_{ki}, y_i) : i = 1, \dots, n \rightarrow$ échantillon aléatoire de taille n tiré de la population.



Pas nécessaire.

Propriétés des MCO

- **MLR.3** $E(u|x_1, \dots, x_k) = 0 \rightarrow$ moyenne conditionnelle nulle.



TS.2 $E(u_t|\mathbf{X}) = 0 \forall t \rightarrow$ moyenne conditionnelle nulle étant donné les variables explicatives $\forall t$
= exogénéité stricte.

→ implique exogénéité contemporaine (hypothèse plus faible) : $E(u_t|\mathbf{x}_t) = 0$.

Pour des séries “cross-section”, pas nécessaire de vérifier $E(u_t|\mathbf{X}) = 0 \forall t$, c-à-d $Corr(u_i, x_j)$ pour $i \neq j$ car **MLR.2**.

! Ne dit rien sur $Corr(u_t, u_s)$ ou $Corr(y_t, y_s)$ pour $t \neq s$!

Causes du rejet de TS.2

Modèle statique simple: $y_t = \beta_0 + \beta_1 z_t + u_t$.

TS.2 $\rightarrow \text{Corr}(u_t, z_t) = 0$ et $\text{Corr}(u_t, z_s) = 0 \forall t \neq s$.

- Variables omises $z_{t-1} \rightarrow \text{Corr}(u_t, z_{t-1}) \neq 0$.
- Erreurs de mesure dans des variables exogènes.
- $\text{Corr}(u_t, z_{t+1}) \neq 0$.

Ex: $mrdrtet = \beta_0 + \beta_1 polpct + u_t$ où $mrdrt$ = taux de criminalité dans une ville et $polpc$ = nombre de policiers par personne.

Quand $\Delta^+ mrdrtet$ non expliquée $\rightarrow \Delta^+ u_t$.

Si dans ce cas la ville engage plus de policiers ($\Delta^+ polpct$) $\rightarrow \text{corr}(polpct_{t+1}, u_t) > 0$.

Propriétés des MCO

- **MLR.4** $\sum_{i=1}^n (x_{ji} - \bar{x})^2 \neq 0 \forall j = 1, \dots, k$ et pas de relation linéaire parfaite entre les $x_j \rightarrow$ variation dans les x_j et pas de collinéarité parfaite.



TS.3 Pas de variable explicative constante et pas de collinéarité parfaite.

- **Théorème 10.1:** Sous les hypothèses **TS.1 - TS.3**,
 $E(\hat{\beta}_j) = \beta_j \forall j = 0, \dots, k$ (conditionnellement à \mathbf{X}).

Propriétés des MCO

- **MLR.5** $Var(u|x_1, \dots, x_k) = \sigma^2 \rightarrow$ homoscédasticité.
↓
TS.4 $Var(u_t|\mathbf{X}) = Var(u) = \sigma^2 \rightarrow \forall t = 1, \dots, n$ homoscédasticité.
Exemple de violation: **NASDAQ**, données journalières (+ de 4000 observations).
- **TS.5** $Corr(u_t, u_s|\mathbf{X}) = 0, \forall t \neq s \rightarrow$ pas de corrélation sérielle.
Exemple de violation: $Corr(u_t, u_{t-1}) = 0.5$.
- **Théorème 10.2:** Sous les hypothèses **TS.1 - TS.5** (appelées hypothèses de **Gauss-Markov** des séries temporelles), $Var(\hat{\beta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{t=1}^n(x_{tj}-\bar{x}_j)^2}$, où R_j^2 est le R^2 de la régression de x_j sur les autres x (+ une cst).

Propriétés des MCO

- **Théorème 10.3:** Sous les hypothèses **TS.1 - TS.5**,
 $E(\hat{\sigma}^2) = \sigma^2$, où $\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{t=1}^n \hat{u}_t^2$.
- **Théorème 10.4 où théorème de Gauss-Markov:**
Sous les hypothèses **TS.1 - TS.5**, $\hat{\beta}_j$ est **BLUE**,
 $\forall i = 0, \dots, k$ (conditionnellement à \mathbf{X}).
- → même propriétés qu'avec des séries “cross-section”.
- Pour effectuer de l'inférence statistique il faut également supposer:
MLR.6 $u \sim N(0, \sigma^2)$ et indépendant des $x_j \rightarrow$ **normalité**.
↓
TS.6 $u \sim N(0, \sigma^2)$ et indépendant de $\mathbf{X} \rightarrow$ **normalité**.
- L'inférence classique est d'application.

Exemple 10.2 et exercice 10.7

- En utilisant la base de données **INTDEF.wf1**, estimatez le modèle de tx d'intérêt suivant:

$$i3_t = \alpha_0 + \alpha_1 inf_t + \alpha_2 def_t + u_t,$$

où $i3$ est le tx d'intérêt obligataire à 3 mois, inf est le tx d'inflation annuel basé sur l'indice des prix à la consommation et def est le déficit budgétaire en % du PNB.

- En Octobre 1979, la FED a changé sa politique pour modifier la masse monétaire en utilisant directement l'instrument du tx d'intérêt.
 - Comment tenir compte de cette information ?
 - Est-ce que cela change les résultats ?

Trend linéaire

- Beaucoup de séries ont un trend. Ex:
Output par heure aux USA (47-87).
- Modèle du type: $y_t = \alpha_0 + \alpha_1 t + e_t$, $t = 1, \dots, n$.
- → Quand $\Delta e_t = 0$, $\Delta y_t = y_t - y_{t-1} = \alpha_1$.
- → Ceteris paribus, α_1 mesure la Δy_t due au passage du temps.
- → $E(y_t) = \alpha_0 + \alpha_1 t$. Croissant ou décroissant en fonction de α_1 .
- → $Var(y_t) = Var(e_t)$. Le trend n'a pas d'effet sur la variance de y_t .
- → On appelle ce type de trend: **trend linéaire**.

Trend exponentiel

- Beaucoup de séries sont mieux représentées par des **trend exponentiels**.
- Ex: Série mensuelle représentant le nombre de demandeurs d'emploi dans le canton d'Anderson dans l'Indiana sur la période 1980.1 à 1988.11. Figures: **Chômeurs et $\log(\text{Chômeurs})$** .
- Appelle plutôt une modélisation du type $\log(y_t) = \beta_0 + \beta_1 t + e_t, t = 1, \dots, n.$
- → comme $\Delta \log(y_t) \approx \frac{y_t - y_{t-1}}{y_{t-1}}$ pour $\Delta \log(y_t)$ petit, $\Delta \log(y_t) = \beta_1$ est approximativement le taux de croissance moyen par période de y_t .

Trend quadratique

- Autre modèle: $y_t = \alpha_0 + \alpha_1 t + \alpha_2 t^2 + e_t$, $t = 1, \dots, n$.
- $\rightarrow \frac{\Delta y_t}{\Delta t} = \alpha_1 + 2\alpha_2 t$. Si $\alpha_1 > 0$ et $\alpha_2 < 0$, le trend a une forme en bosse.

Exemple

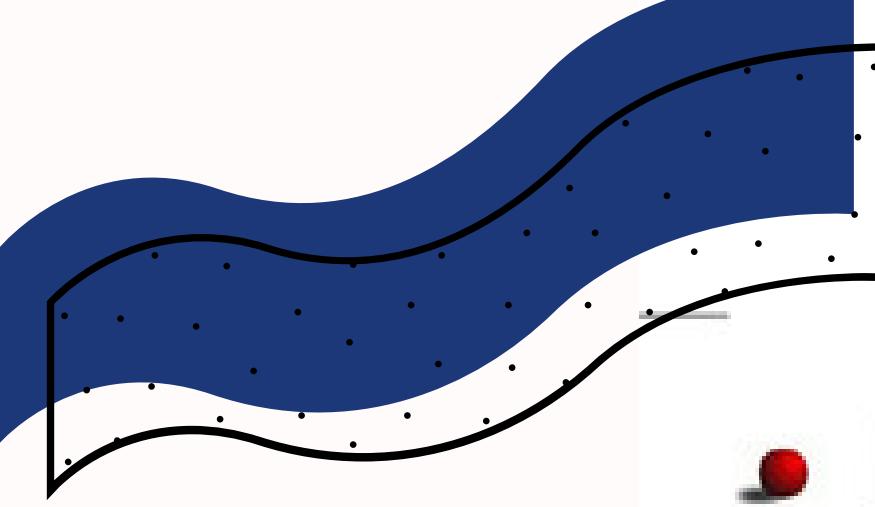
- Données annuelles (1947-1988) sur l'investissement immobilier (*invpc*) et sur un indice des prix immobilier (*price*, 1982 = 1). **Graphique des données.**
- **Modèle simple sans trend.**
- L'élasticité Investissement-Indice de prix est très élevée: 1.241 et statistiquement > 0.
- **Modèle simple avec trend.**
- L'élasticité Investissement-Indice de prix est négative mais non significative. Par contre le trend est significatif et implique environ une hausse de 1% par an de *invpc*. De plus le R^2 est passé de 0.208 à 0.341.
- → relation fallacieuse expliquée par un trend commun, c-à-d l'existence d'une variable omise qui explique l'évolution conjointe des deux séries.



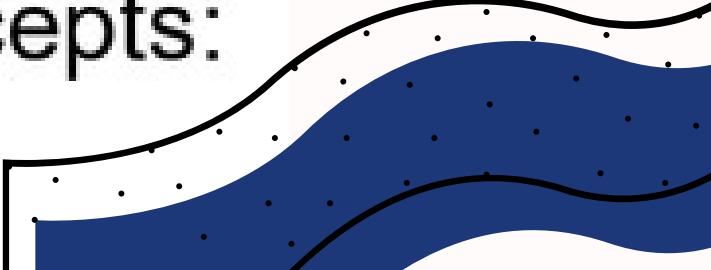
Autre interprétation

- Modèle estimé: $\hat{y}_t = \hat{\beta}_0 + \hat{\beta}_1 x_{t1} + \hat{\beta}_2 x_{t2} + \hat{\beta}_3 t.$
- Il est facile de montrer que l'on peut obtenir également $\hat{\beta}_1$ et $\hat{\beta}_2$ de cette manière.
- Régresser par MCO:
 - y_t sur *cst* et trend; → sauver les résidus: y'_t .
 - x_{t1} sur *cst* et trend; → sauver les résidus: x'_{t1} .
 - x_{t2} sur *cst* et trend; → sauver les résidus: x'_{t2} .
- Régresser par MCO y'_t sur x'_{t1} et x'_{t2} (avec ou sans constante).
- Revenons à l'exemple précédent. Graphiques log(invpc) et log(price).**
- Important: le $R^2 = 0.008$ → sur les variables “détrendées”, $\log(price)$ n’explique rien de l’évolution de $\log(invpc)$.

Séries temporelles (plus) avancées: Chapitre 11



Rappels



- Nous avons étudié les propriétés des MCO en échantillon fini.
- Nous avons obtenu les mêmes propriétés avec des séries temporelles qu'avec des séries "cross-section" mais au prix d'hypothèses très fortes.
- Nous allons tenter de relâcher certaines hypothèses comme celle d'exogénéité stricte, hypothèse violée si par exemple \mathbf{X} comprend y_{t-i} avec $i > 0$.
- On ne pourra plus étudier le comportement en échantillon fini des MCO mais bien le comportement asymptotique (n grand).
- On va devoir introduire pour cela 2 nouveaux concepts: la **stationnarité** et la **dépendance faible**.

Stationnarité

- Définition formelle: *Le processus $\{x_t : t = 1, 2, \dots\}$ est strictement stationnaire si pour chaque collection d'indices de temps $1 \leq t_1 < t_2 < \dots < t_m$, la distribution jointe de $(x_{t_1}, x_{t_2}, \dots, x_{t_m})$ est la même que la distribution jointe de $(x_{t_1+h}, x_{t_2+h}, \dots, x_{t_m+h})$, \forall entier $h \geq 1$.*
- Les x_t peuvent être corrélés mais la corrélation doit être constante.
- Exemple de processus non-stationnaire:
 $y_t = \alpha_0 + \alpha_1 t + e_t$, $t = 1, \dots, n$ et $e_t \sim iid(0, 1)$.
 $\rightarrow E(y_t) = \alpha_0 + \alpha_1 t$.
- Par contre, $y_t - \alpha_0 - \alpha_1 t$ est stationnaire.
- Il est très difficile de tester l'hypothèse de stationnarité.

Processus Covariance-stationnaire

- Définition: *Un processus stochastique* $\{x_t : t = 1, 2, \dots\}$ avec *second moment fini* $[E(x_t^2) < \infty]$ est **covariance-stationnaire** si:
 - $E(x_t)$ est constante;
 - $Var(x_t)$ est constante;
 - $\forall t, h \geq 1, Cov(x_t, x_{t+h})$ dépend seulement de h et non de t .
- \rightarrow implique uniquement les 2 premiers moments.
- Si un processus stationnaire a $E(x_t^2) < \infty \Rightarrow$ il est covariance-stationnaire.
- L'inverse n'est pas vrai.
- Exemples : **Rejet et non rejet** de l'hypothèses de covariance-stationnarité.

Dépendance faible

- Définition 1: Un processus stationnaire $\{x_t : t = 1, 2, \dots\}$ est faiblement dépendant si x_t et x_{t+h} sont ± indépendant au fur et à mesure que $h \rightarrow \infty$.
- Définition 2: Un processus covariance-stationnaire $\{x_t : t = 1, 2, \dots\}$ est faiblement dépendant si $\text{Corr}(x_t, x_{t+h}) \rightarrow 0$ suffisamment rapidement si $h \rightarrow \infty$.
- Cette hypothèse remplace celle d'échantillon aléatoire permettant au théorème centrale limite et à la loi des grands nombres de s'appliquer aux séries temporelles.

Exemple: MA(1)

- $MA(1) : x_t = e_t + \alpha_1 e_{t-1}$, où $e_t \sim i.i.d.(0, \sigma_e^2)$.
- $\rightarrow Corr(x_t, x_{t+1}) = \frac{\alpha_1}{1+\alpha_1^2}$.
- $\rightarrow Corr(x_t, x_{t+h}) = 0 \quad \forall h > 1$.
- $MA(1)$ est faiblement dépendant.
- Exemple de processus $MA(1)$.

Exemple: AR(1)

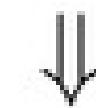
- $AR(1) : y_t = \rho_1 y_{t-1} + e_t$, où $e_t \sim i.i.d.(0, \sigma_e^2)$.
- Si $|\rho_1| < 1 \rightarrow \text{Corr}(y_t, y_{t+h}) = \rho_1^h \forall h \geq 1$.
- $AR(1)$ est faiblement dépendant si $|\rho_1| < 1$ car dans ce cas, $\rho_1^h \rightarrow 0$ si $h \rightarrow \infty$.
- Exemple de processus $AR(1)$ stable.
- Voir les détails mathématiques dans le bouquin.

Propriétés asymptotiques des MCO

Nous allons montrer que les MCO sont encore valables avec des hypothèses plus faibles que celles postulées au Chapitre 10.

Hypothèses de base:

- **TS.1** $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t \rightarrow$ linéaire en les paramètres.



- **TS.1'** $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t \rightarrow$ linéaire en les paramètres et $(\mathbf{x}_t, y_t) : t = 1, 2, \dots$ est stationnaire et faiblement dépendant.

Propriétés asymptotiques des MCO

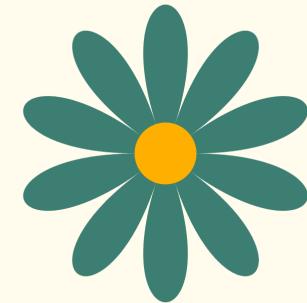
- **TS.2** $E(u_t | \mathbf{X}) = 0 \quad \forall t \rightarrow$ moyenne conditionnelle nulle étant donné les variables explicatives $\forall t$
= exogénéité stricte.



TS.2' $E(u_t | \mathbf{x}_t) = 0 \rightarrow$ moyenne conditionnelle nulle étant donné les variables explicatives contemporaines
= exogénéité contemporaine
 \rightarrow implique que $E(u_t) = 0$ et $Cov(x_{tj}, u_t) = 0,$
 $\forall j = 1, \dots, k.$



Propriétés asymptotiques des MCO



- **TS.3** Pas de variable explicative constante et pas de collinéarité parfaite.



TS.3' Idem.

- **Théorème 11.1:** Sous les hypothèses **TS.1'**, **TS.2'** et **TS.3'**, $\text{plim} \hat{\beta}_j = \beta_j \quad \forall j = 0, \dots, k$.
- **AR(1):** $y_t = \beta_0 + \beta_1 y_{t-1} + u_t$, où $E(u_t | y_{t-1}, y_{t-2}, \dots) = 0$.
 $\rightarrow E(y_t | y_{t-1}, y_{t-2}, \dots) = E(y_t | y_{t-1}) = \beta_0 + \beta_1 y_{t-1}$.
 \rightarrow seulement y_{t-1} affecte la valeur espérée de y_t .
 \rightarrow la valeur espérée de y_t est une fonction linéaire de y_{t-1} .

Estimation du modèle AR(1)

- Comme \mathbf{x}_t contient seulement y_{t-1} , $E(u_t|y_{t-1}, y_{t-2}, \dots) = 0$ implique que l'hypothèse **TS.2'** tient.
- Par contre **TS.2** ne tient pas. En effet cette hypothèse implique que $\forall t$, u_t est non corrélé avec $x_1, \dots, x_t, \dots, x_n$. Ceci est faux pour un processus AR(1) car
$$\begin{aligned} \text{Cov}(x_{t+1}, u_t) &= \text{Cov}(y_t, u_t) = \text{Cov}(\beta_0 + \beta_1 y_{t-1} + u_t, u_t) = \\ &= \beta_1^2 \text{Cov}(y_{t-1}, u_t) + \text{Cov}(u_t, u_t) = 0 + \sigma^2 > 0, \text{ alors que} \\ \text{Cov}(y_{t-1}, u_t) &= 0 \rightarrow \text{exogénéité contemporaine.} \end{aligned}$$
Pour que l'hypothèse de dépendance faible tienne il faut que $|\beta_1| < 1$. \rightarrow Théorème 11.1: Convergence des MCO.
- Illustration: Modèle classique et Modèle AR(1).

Propriétés asymptotiques des MCO

- **TS.4** $Var(u_t|\mathbf{X}) = Var(u) = \sigma^2, \forall t = 1, \dots, n \rightarrow$ homoscédasticité.



TS.4' $Var(u_t|\mathbf{x}_t) = \sigma^2, \forall t = 1, \dots, n \rightarrow$ homoscédasticité contemporaine.

- **TS.5** $Corr(u_t, u_s|\mathbf{X}) = 0, \forall t \neq s \rightarrow$ pas de corrélation serielle.



TS.5' $E(u_t u_s | \mathbf{x}_t, \mathbf{x}_s) = 0, \forall t \neq s \rightarrow$ pas de corrélation serielle.

En pratique on omet souvent le conditionnement sur \mathbf{x}_t et \mathbf{x}_s et on vérifie que u_t et u_s sont non-correlés.

Propriétés asymptotiques des MCO

- **Théorème 11.2:** Sous les hypothèses **TS.1' - TS.5'**, les estimateurs MCO sont asymptotiquement normalement distribués. De plus, les écart-types standards sont asymptotiquement valides ainsi que les t , F et LM tests.
- Illustration: **Modèle classique** et **Modèle AR(1)**.
- Exemple: Tester l'hypothèse d'efficience des marchés. y_t est le rendement hebdomadaire en % (le mercredi) du **New York Stock Exchange composite index**. Une forme faible de l'hypothèse d'efficience des marché stipule que l'information publique disponible en $t - 1$ ne doit pas permettre pour prédire y_t
 $\rightarrow E(y_t|y_{t-1}, y_{t-2}, \dots) = E(y_t).$
- \rightarrow Estimer un $AR(q)$ et tester H_0 : coefficient $AR(q)$ sont nuls.

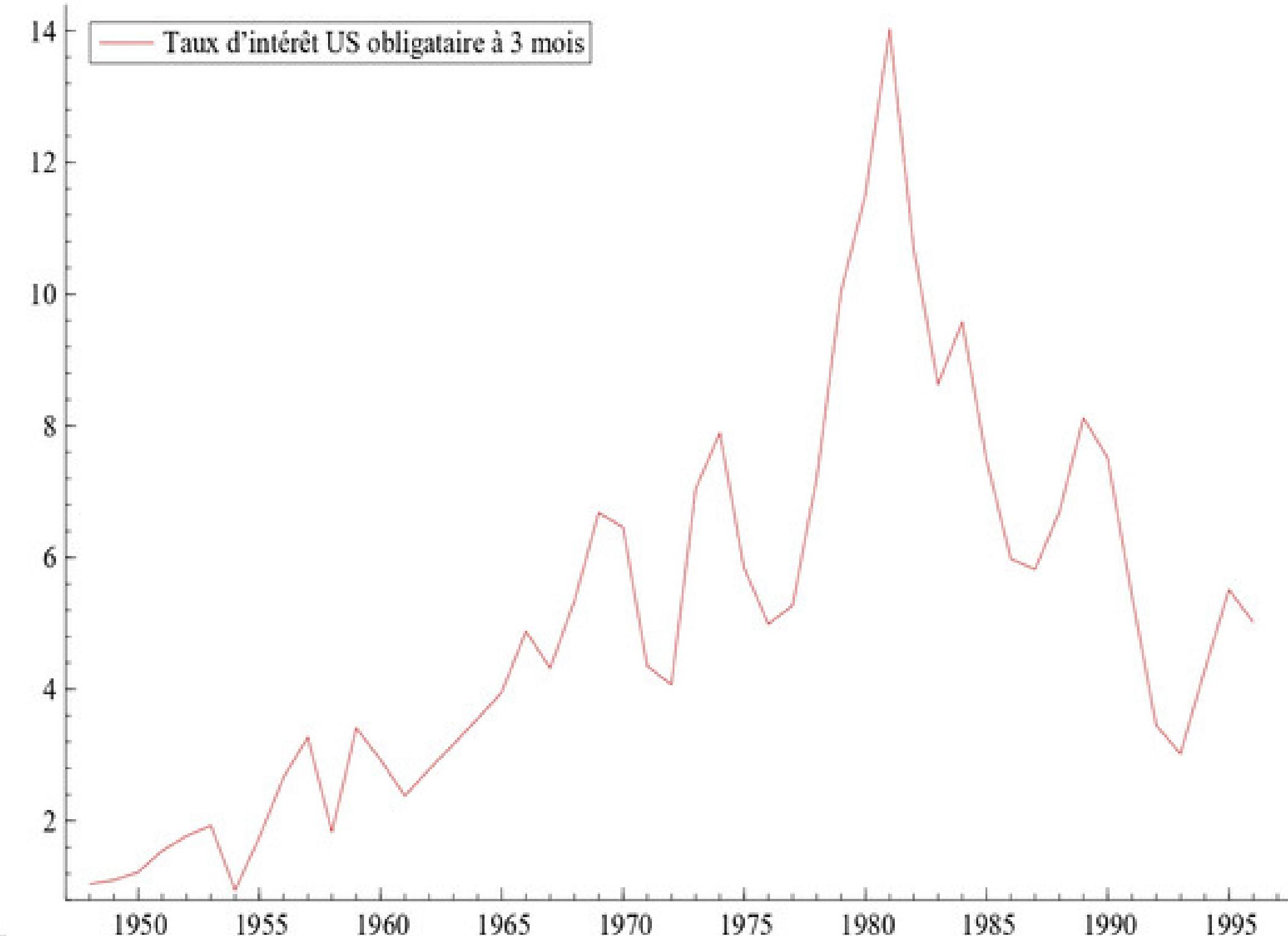
Séries temporelles très persistantes

- Beaucoup de séries chronologiques sont très persistantes, ce qui peut mener au rejet de l'hypothèse de dépendance faible évoquée plus haut.
- L'exemple le plus courant est celui d'une marché aléatoire.
 - $\rightarrow y_t = y_{t-1} + e_t, t = 1, 2, \dots$, où $e_t \sim iid(0, \sigma_e^2)$.
 - $\rightarrow AR(1)$ avec $\rho_1 = 1$.
 - $\rightarrow y_t = e_t + e_{t-1} + \dots + e_1 + y_0$.
 - $\rightarrow E(y_t) = E(e_t) + E(e_{t-1}) + \dots + E(e_1) + E(y_0)$.
 - $\rightarrow E(y_t) = E(y_0), \forall t \geq 1$.
 - $\rightarrow Var(y_t) =$
 $Var(e_t) + Var(e_{t-1}) + \dots + Var(e_1) + Var(y_0) = t\sigma_e^2$.

Séries temporelles très persistantes

- Persistance ?
- $\rightarrow y_{t+h} = e_{t+h} + e_{t+h-1} + \dots + e_{t+1} + y_t.$
- $\rightarrow E(y_{t+h}|y_t) = y_t, \forall h \geq 1.$
- Alors que pour un $AR(1)$: $E(y_{t+h}|y_t) = \rho_1^h y_t.$
- $\rightarrow \text{Corr}(y_t, y_{t+h}) = \sqrt{t/(t+h)} \approx 1$ pour t grand.
- \rightarrow une marche aléatoire n'est pas covariance-stationnaire.
- \rightarrow une marche aléatoire n'est pas faiblement dépendante.
- Illustration.

Exemple de marche aléatoire



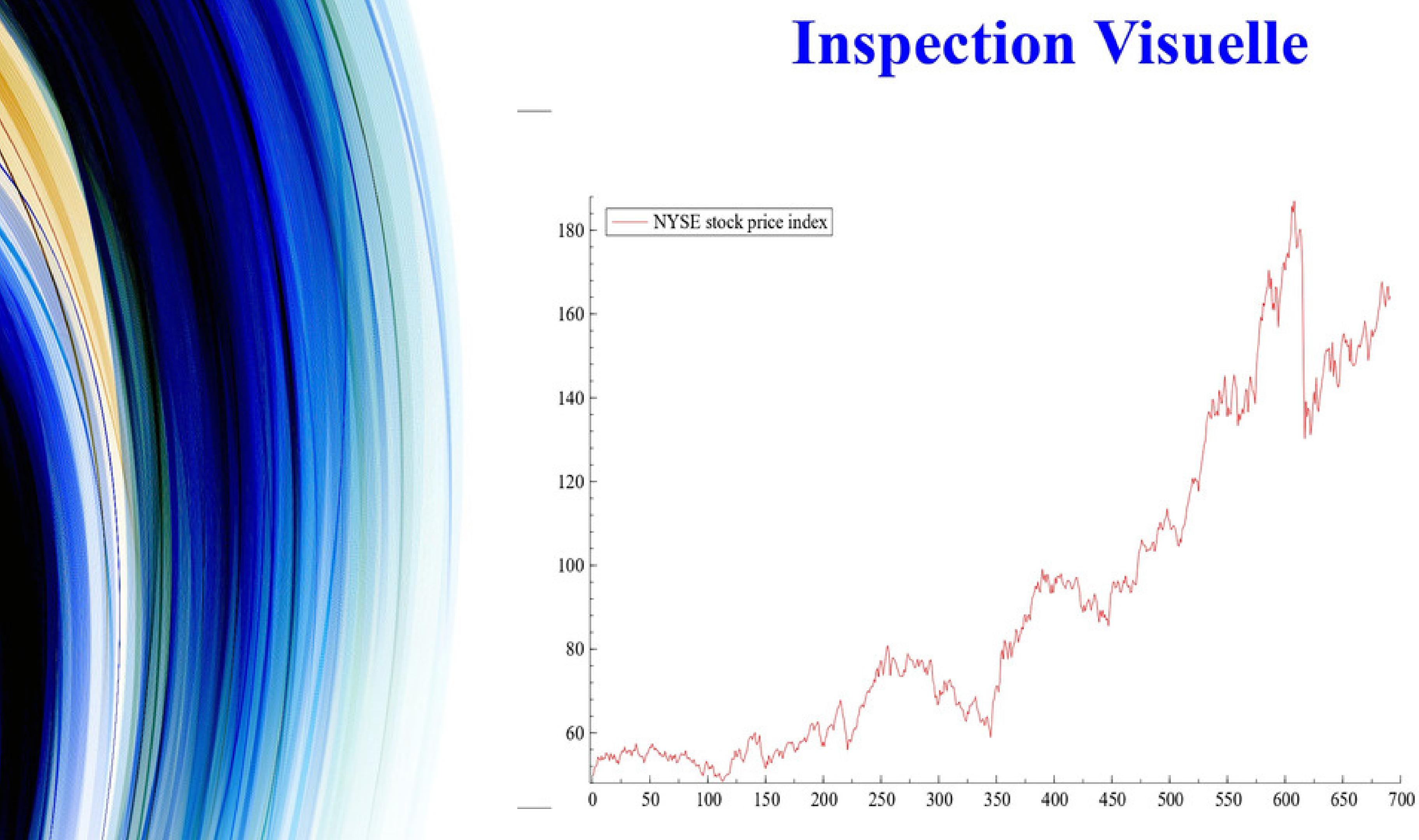
Implication économique

- Si y est faiblement dépendant, une Δ de politique économique affectant y (ex: PNB) aujourd’hui aura peu d’effet dans quelques années.
- Par contre, si y suit une marche aléatoire (ou plus généralement a une racine unitaire, c-à-d $\rho_1 = 1$), une Δ de politique économique affectant y aujourd’hui aura un effet important dans plusieurs années.
- Tester la présence d’une racine unitaire a donc une implication économique importante.

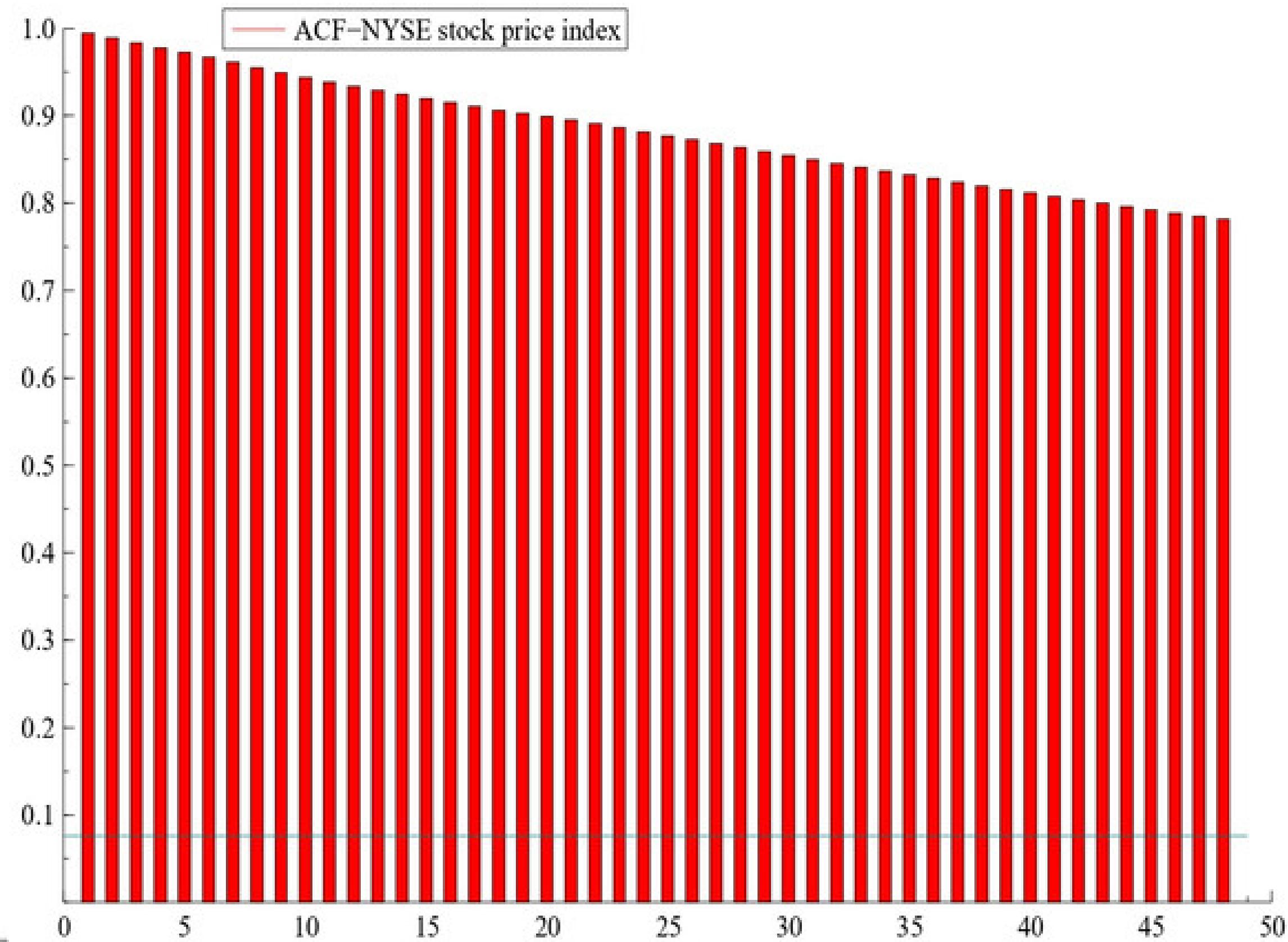
Marche aléatoire avec tendance (drift)

- $y_t = \alpha_0 + y_{t-1} + e_t$: AR(1) avec constante.
- $\rightarrow y_t = \alpha_0 t + e_t + e_{t-1} + \dots + e_1 + y_0$.
- $\rightarrow E(y_t) = \alpha_0 t$ si $E(y_0) = 0$.
- $\rightarrow E(y_{t+h}|y_t) = \alpha_0 h + y_t, \forall h \geq 1$.
- $\rightarrow Var(y_t) = t\sigma_e^2$.
- Exemple de marche aléatoire avec tendance.
- Comment détecter une RU ?

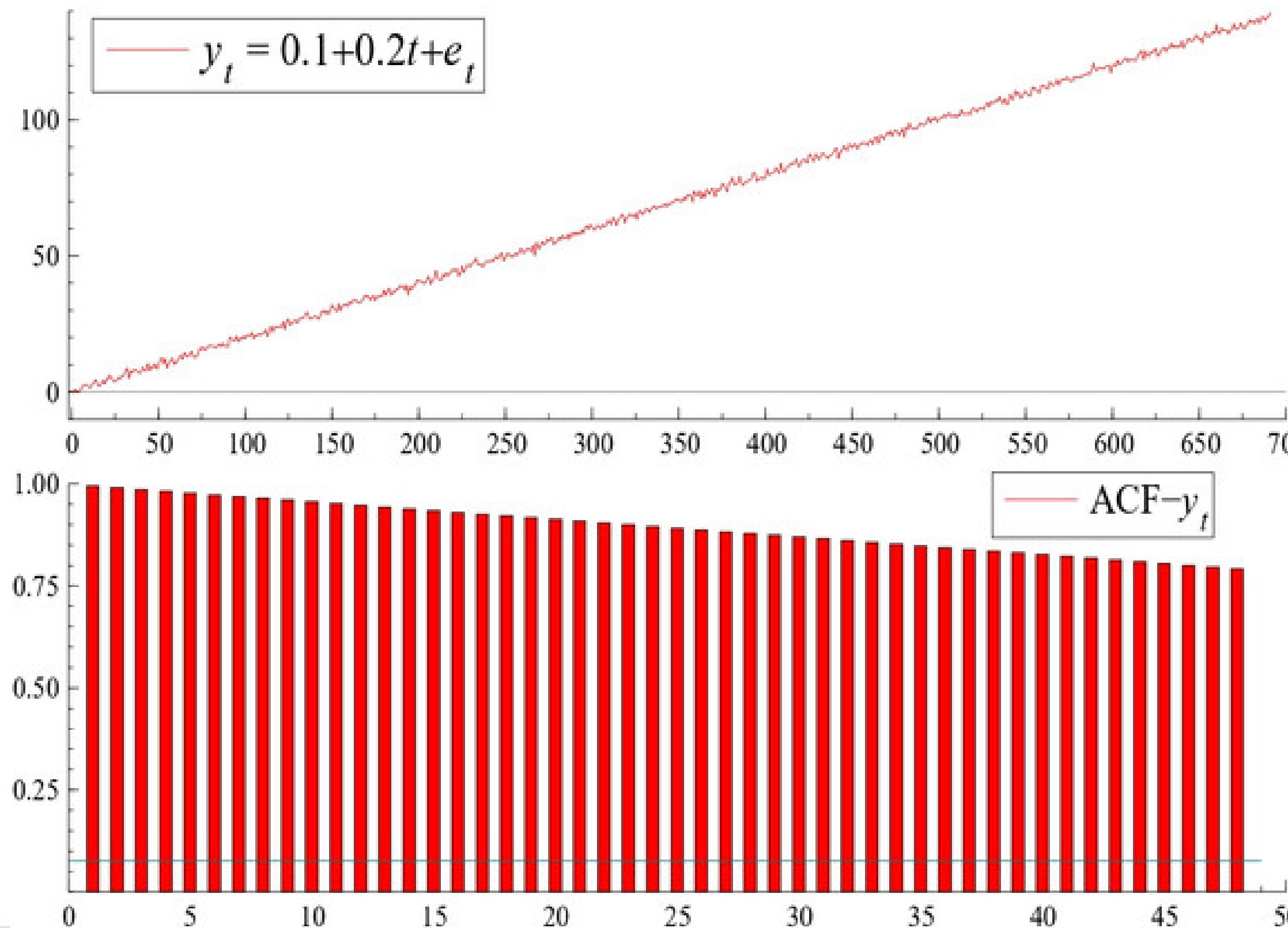
Inspection Visuelle



Correlogramme



Difficile de distinguer RU et trend



Tester la présence d'une RU

- Modèle $AR(1)$: $y_t = \alpha + \rho y_{t-1} + e_t$.
- y_0 est supposé observé et $E(e_t | y_{t-1}, y_{t-2}, \dots, y_0) = 0$.
- RU si $\rho = 1 \rightarrow H_0 : \rho = 1$ et $H_1 : \rho < 1$.
- Notons que: $y_t - y_{t-1} = \alpha + (\rho - 1)y_{t-1} + e_t$.
- $\Leftrightarrow \Delta y_t = \alpha + \theta y_{t-1} + e_t$.
- RU si $H_0 : \theta = 0$ et $H_1 : \theta < 0$.
- Problème sous H_0 , y_{t-1} n'est pas faiblement dépendant et donc les t-stats n'ont pas une distribution standard (théorème centrale limite pas applicable, même pour n grand).
- → Il faut utiliser des valeurs critiques tabulées par **Dickey Fuller (1979)**.

Tester la présence d'une RU

- Valeurs critiques asymptotiques pour un test de RU avec constante mais sans trend:

| α | 1% | 2.5% | 5% | 10% |
|-----------------|-------|-------|-------|-------|
| Valeur critique | -3.43 | -3.12 | -2.86 | -2.57 |

- Exemple: **Taux d'intérêt US obligataire à 3 mois.**
- Notez que à 5% la valeur critique traditionnelle pour n grand est -1.65 contre -2.86.
- Limite: dans le test DF usuel, on suppose sous H_0 que Δy_t n'a pas d'autocorrélation.

Test ADF

- Le test ADF (Augmenté) généralise le test de DF:
$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \dots + \gamma_p \Delta y_{t-p} + e_t.$$
- Les valeurs critiques sont les mêmes que le test de DF.
- Les t-stats de $\gamma_1, \dots, \gamma_p$ sont standards asymptotiquement.
- Le F-test $\gamma_1 = \dots = \gamma_p = 0$ est valide asymptotiquement.
- Inclure le bon nombre de lags est important car cela pourrait affecter la fiabilité du test de RU. En effet, les valeurs critiques sont calculées en supposant que toute la dynamique de Δy_t a été modélisée.
- Exemple: **Taux d'intérêt US obligataire à 3 mois.**

Test ADF avec trend

- Nous avons évoqué le fait qu'il est parfois difficile de distinguer RU et tendance.
- → une série peut donc être stationnaire autour d'un trend, c-à-d $y_t - \hat{\delta}t$ est faiblement dépendant.
- Il est donc important d'inclure un trend dans le test ADF si celui-ci est apparent (visuellement).
- Le test ADF (Augmenté) généralise le test de DF:
$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \dots + \gamma_p \Delta y_{t-p} + e_t.$$
- Sous H_0 : $\theta = 0$ et sous H_1 : y_t est stationnaire autour d'un trend.

Test ADF avec trend

- Les valeurs critiques sont différentes.
Valeurs critiques asymptotiques pour un test de RU avec constante et avec trend:

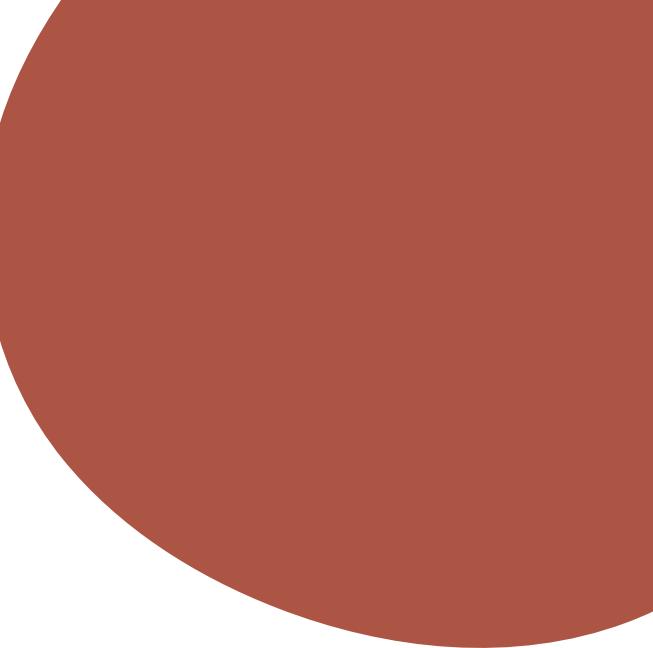
| α | 1% | 2.5% | 5% | 10% |
|-----------------|-------|-------|-------|-------|
| Valeur critique | -3.96 | -3.66 | -3.41 | -3.12 |

- Le t-stat relatif au trend n'a pas une distribution standard.
- Exemple: **Taux d'intérêt US obligataire à 3 mois.**
- Il y a beaucoup d'autres tests de RU qui ont tous leurs avantages et inconvénients.

Que faire si y_t a une RU ?

- Si $y_t = \alpha + y_{t-1} + e_t \rightarrow \Delta y_t \equiv y_t - y_{t-1} = \alpha + e_t$.
- Si par contre Δy_t est stationnaire, y_t est dit **Intégré d'ordre 1** ou $I(1)$ et Δy_t est dit **Intégré d'ordre 0** ou $I(0)$.
- Si l'application le permet, on peut donc estimer par MCO un modèle sur Δy_t .
- Dans beaucoup d'applications économiques, on définit des rendements en %: $r_t = 100\Delta \log(y_t) \simeq 100 \frac{y_t - y_{t-1}}{y_{t-1}}$.
- Si différentier les séries n'est pas possible de part la nature de l'application, on peut avoir recourt à des techniques plus avancées connues sous le nom **Cointégration** (voir section 18.4), permettant de modéliser y_t et non r_t .
- Qu'est-ce qu'une régression fallacieuse (ou spurious regression) ?

Corrélation sérielle et hétéroskedasticité: Chapitre 12



Biais des MCO et autocorrélation

- Supposons que $y_t = \beta_0 + \beta_1 x_t + u_t$.
- Supposons que le terme d'erreur suit un $AR(1)$: $u_t = \rho u_{t-1} + e_t$ avec $|\rho| < 1$ et $e_t \sim iid(0, \sigma_e^2)$.
- Chapitre 10: **TS.1-TS.3** → MCO non-biaisés pour autant que x soit strictement exogène.
- Chapitre 11: **TS.1'-TS.3'** → MCO consistant pour autant que y_t soit faiblement dépendant et $E(u_t|x_t) = 0$.
- Rien n'est dit sur le fait que u_t ne puisse pas suivre un $AR(1)$.
- Par contre les MCO ne sont plus BLUE car violation de **TS.5** et **TS.5'**.

Efficiency-Inférence des MCO

- Estimation du modèle $y_t = \beta_0 + \beta_1 x_t + u_t$ par MCO, où u_t supposé $iid(0, \sigma^2)$ et pour simplifier $\bar{x} = 0$.
- Rappelons que u_t suit en réalité un $AR(1)$.
- $\hat{\beta}_1 = \beta_1 + SST_x^{-1} \sum_{i=1}^n x_t u_t$, où $SST_x = \sum_{i=1}^n x_t^2$.
-

$$\begin{aligned} Var(\hat{\beta}_1) &= SST_x^{-2} Var\left(\sum_{i=1}^n x_t u_t\right) \\ &= SST_x^{-2} \left(\sum_{i=1}^n x_t^2 Var(u_t) + 2 \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} x_t x_{t+j} E(u_t u_{t+j}) \right) \\ &= \frac{\sigma^2}{SST_x} + 2 \frac{\sigma^2}{SST_x^2} \sum_{t=1}^{n-1} \sum_{j=1}^{n-t} \rho^j x_t x_{t+j} \end{aligned}$$

Efficiency-Inference des MCO

- La formule traditionnelle ignore le second terme.
- Si $\rho > 0 \rightarrow$ on sous-estime $Var(\hat{\beta}_1)$.
- Si $\rho < 0 \rightarrow$ on sur-estime $Var(\hat{\beta}_1)$.
- Question: Supposons que $u_t = e_t + \alpha e_{t-1}$. Montrez que la formule traditionnelle pour calculer $Var(\hat{\beta}_1)$ est incorrecte si $\alpha \neq 0$.

Tester la présence d'autocorrélation

- Dans le modèle général:

$y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t \rightarrow$ comment tester la présence de corrélation sérielle ?

- Deux types de tests:

- 1. Tests basés sur l'hypothèse que X est strictement exogène: t-test ou DW.
- 2. Tests basés sur l'hypothèse que X n'est pas strictement exogène.

Si X est strictement exogène

- Supposons que $y_t = \beta_0 + \beta_1 x_t + u_t$.
- Nous voulons tester si le terme d'erreur suit un $AR(1)$: $u_t = \rho u_{t-1} + e_t$ avec $|\rho| < 1$.
- $H_0 : \rho = 0$.
- Développons tout d'abord un test valide quand n est grand et quand X est strictement exogène.
- Hypothèses supplémentaires disant en quelque sorte que e_t est *iid*:
 $E(e_t | u_{t-1}, u_{t-2}, \dots) = 0$ et $Var(e_t | u_{t-1}) = Var(e_t) = \sigma_e^2$.
- Si les u_t étaient observés on pourrait utiliser le **Théorème 11.2** pour valider l'utilisation asymptotique des t-tests dans la régression $u_t = \rho u_{t-1} + e_t$.
- Que ce passe-t-il si on remplace u_t par \hat{u}_t ? A noter que \hat{u}_t dépend de $\hat{\beta}_0$ et de $\hat{\beta}_1$.

Marche à suivre pour le t-test

- Estimer $y_t = \beta_0 + \beta_1 x_t + u_t$ et obtenir \hat{u}_t .
- Estimer $\hat{u}_t = \rho \hat{u}_{t-1} + e_t$.
- Utiliser un t-test pour tester $H_0 : \rho = 0$ contre $H_1 : \rho < \neq > 0$.

Test de Durbin-Watson

- Le test de Durbin-Watson est une autre test pour tester la corrélation sérielle d'ordre 1 sous l'hypothèse que X est strictement exogène.
- La statistique $DW = \frac{\sum_{t=2}^n (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^n \hat{u}_t^2}$, où \hat{u}_t est le résidu d'une régression du type
$$y_t = \beta_0 + \beta_1 x_{1t} + \dots + \beta_k x_{kt} + u_t.$$
- Il est facile de montrer que $DW \approx 2(1 - \rho)$.
- Durbin et Watson (1951) ont dérivé la distribution de DW conditionnellement à X sous l'hypothèse que les hypothèses du modèle linéaire classique sont vérifiées (incluant la normalité).
- La distribution de DW dépend de n, k , et du fait qu'on a inclut une constante ou pas.

Test de Durbin-Watson

- Notons que

$$\sum_{t=2}^n (\hat{u}_t - \hat{u}_{t-1})^2 = \sum_{t=2}^n \hat{u}_t^2 + \sum_{t=2}^n \hat{u}_{t-1}^2 - 2 \sum_{t=2}^n \hat{u}_t \hat{u}_{t-1}.$$

- Si $\hat{\rho} = 1 \rightarrow DW = 0$.
- Si $\hat{\rho} = -1 \rightarrow DW = 4$.
- Si $\hat{\rho} = 0 \rightarrow DW = 2$.
- $H_0 : \rho = 0$ contre généralement $H_1 : \rho > 0$.
- Durbin et Watson (1951) reportent des valeurs critiques inférieures d_L et supérieures d_U .
- Si $d_L \leq DW \leq d_U$, on ne rejette pas H_0 .
- La plupart des logiciels économétriques reportent DW mais pas les valeurs critiques.

Table de Durbin-Watson pour $H_1 : \rho > 0$

Table E.8 Critical Values d_L and d_U of the Durbin-Watson Statistic D (Critical Values Are One-Sided)^a

| $\alpha = .05$ | | | | | | | | | | $\alpha = .01$ | | | | | | | | | | |
|----------------|-------|---------|-------|---------|-------|---------|-------|---------|-------|----------------|-------|---------|-------|---------|-------|---------|-------|---------|------|------|
| $P = 1$ | | $P = 2$ | | $P = 3$ | | $P = 4$ | | $P = 5$ | | $P = 1$ | | $P = 2$ | | $P = 3$ | | $P = 4$ | | $P = 5$ | | |
| n | d_L | d_U | d_L | d_U | d_L | d_U | d_L | d_U | d_L | d_U | | |
| 15 | 1.08 | 1.36 | .95 | 1.54 | .82 | 1.75 | .69 | 1.97 | .56 | 2.21 | .81 | 1.07 | .70 | 1.25 | .59 | 1.46 | .49 | 1.70 | .39 | 1.96 |
| 16 | 1.10 | 1.37 | .98 | 1.54 | .86 | 1.73 | .74 | 1.93 | .62 | 2.15 | .84 | 1.09 | .74 | 1.25 | .63 | 1.44 | .53 | 1.66 | .44 | 1.90 |
| 17 | 1.13 | 1.38 | 1.02 | 1.54 | .90 | 1.71 | .78 | 1.90 | .67 | 2.10 | .87 | 1.10 | .77 | 1.25 | .67 | 1.43 | .57 | 1.63 | .48 | 1.85 |
| 18 | 1.16 | 1.39 | 1.05 | 1.53 | .93 | 1.69 | .82 | 1.87 | .71 | 2.06 | .90 | 1.12 | .80 | 1.26 | .71 | 1.42 | .61 | 1.60 | .52 | 1.80 |
| 19 | 1.18 | 1.40 | 1.08 | 1.53 | .97 | 1.68 | .86 | 1.85 | .75 | 2.02 | .93 | 1.13 | .83 | 1.26 | .74 | 1.41 | .63 | 1.58 | .56 | 1.77 |
| 20 | 1.20 | 1.41 | 1.10 | 1.54 | 1.00 | 1.68 | .90 | 1.83 | .79 | 1.99 | .95 | 1.15 | .86 | 1.27 | .77 | 1.41 | .68 | 1.57 | .60 | 1.74 |
| 21 | 1.22 | 1.42 | 1.13 | 1.54 | 1.03 | 1.67 | .93 | 1.81 | .83 | 1.96 | .97 | 1.16 | .89 | 1.27 | .80 | 1.41 | .72 | 1.55 | .63 | 1.71 |
| 22 | 1.24 | 1.43 | 1.15 | 1.54 | 1.05 | 1.66 | .96 | 1.80 | .86 | 1.94 | 1.00 | 1.17 | .91 | 1.28 | .83 | 1.40 | .75 | 1.54 | .66 | 1.69 |
| 23 | 1.26 | 1.44 | 1.17 | 1.54 | 1.08 | 1.66 | .99 | 1.79 | .90 | 1.92 | 1.02 | 1.19 | .94 | 1.29 | .86 | 1.40 | .77 | 1.53 | .70 | 1.67 |
| 24 | 1.27 | 1.45 | 1.19 | 1.55 | 1.10 | 1.66 | 1.01 | 1.78 | .93 | 1.90 | 1.04 | 1.20 | .96 | 1.30 | .88 | 1.41 | .80 | 1.53 | .72 | 1.66 |
| 25 | 1.29 | 1.45 | 1.21 | 1.55 | 1.12 | 1.66 | 1.04 | 1.77 | .95 | 1.89 | 1.05 | 1.21 | .98 | 1.30 | .90 | 1.41 | .83 | 1.52 | .75 | 1.65 |
| 26 | 1.30 | 1.46 | 1.22 | 1.55 | 1.14 | 1.65 | 1.06 | 1.76 | .98 | 1.88 | 1.07 | 1.22 | 1.00 | 1.31 | .93 | 1.41 | .85 | 1.52 | .78 | 1.64 |
| 27 | 1.32 | 1.47 | 1.24 | 1.56 | 1.16 | 1.65 | 1.08 | 1.76 | 1.01 | 1.86 | 1.09 | 1.23 | 1.02 | 1.32 | .95 | 1.41 | .88 | 1.51 | .81 | 1.63 |
| 28 | 1.33 | 1.48 | 1.26 | 1.56 | 1.18 | 1.65 | 1.10 | 1.75 | 1.03 | 1.85 | 1.10 | 1.24 | 1.04 | 1.32 | .97 | 1.41 | .90 | 1.51 | .83 | 1.62 |
| 29 | 1.34 | 1.48 | 1.27 | 1.56 | 1.20 | 1.65 | 1.12 | 1.74 | 1.05 | 1.84 | 1.12 | 1.25 | 1.05 | 1.33 | .99 | 1.42 | .92 | 1.51 | .85 | 1.61 |
| 30 | 1.35 | 1.49 | 1.28 | 1.57 | 1.21 | 1.65 | 1.14 | 1.74 | 1.07 | 1.83 | 1.13 | 1.26 | 1.07 | 1.34 | 1.01 | 1.42 | .94 | 1.51 | .88 | 1.61 |
| 31 | 1.36 | 1.50 | 1.30 | 1.57 | 1.23 | 1.65 | 1.16 | 1.74 | 1.09 | 1.83 | 1.15 | 1.27 | 1.08 | 1.34 | 1.02 | 1.42 | .96 | 1.51 | .90 | 1.60 |
| 32 | 1.37 | 1.50 | 1.31 | 1.57 | 1.24 | 1.65 | 1.18 | 1.73 | 1.11 | 1.82 | 1.16 | 1.28 | 1.10 | 1.35 | 1.04 | 1.43 | .98 | 1.51 | .92 | 1.60 |
| 33 | 1.38 | 1.51 | 1.32 | 1.58 | 1.26 | 1.65 | 1.19 | 1.73 | 1.13 | 1.81 | 1.17 | 1.29 | 1.11 | 1.36 | 1.05 | 1.43 | 1.00 | 1.51 | .94 | 1.59 |
| 34 | 1.39 | 1.51 | 1.33 | 1.58 | 1.27 | 1.65 | 1.21 | 1.73 | 1.15 | 1.81 | 1.18 | 1.30 | 1.13 | 1.36 | 1.07 | 1.43 | 1.01 | 1.51 | .95 | 1.59 |
| 35 | 1.40 | 1.52 | 1.34 | 1.58 | 1.28 | 1.65 | 1.22 | 1.73 | 1.16 | 1.80 | 1.19 | 1.31 | 1.14 | 1.37 | 1.08 | 1.44 | 1.03 | 1.51 | .97 | 1.59 |
| 36 | 1.41 | 1.52 | 1.35 | 1.59 | 1.29 | 1.65 | 1.24 | 1.73 | 1.18 | 1.80 | 1.21 | 1.32 | 1.15 | 1.38 | 1.10 | 1.44 | 1.04 | 1.51 | .99 | 1.59 |
| 37 | 1.42 | 1.53 | 1.36 | 1.59 | 1.31 | 1.66 | 1.25 | 1.72 | 1.19 | 1.80 | 1.22 | 1.32 | 1.16 | 1.38 | 1.11 | 1.45 | 1.06 | 1.51 | 1.00 | 1.59 |
| 38 | 1.43 | 1.54 | 1.37 | 1.59 | 1.32 | 1.66 | 1.26 | 1.72 | 1.21 | 1.79 | 1.23 | 1.33 | 1.18 | 1.39 | 1.12 | 1.45 | 1.07 | 1.52 | 1.02 | 1.58 |
| 39 | 1.43 | 1.54 | 1.38 | 1.60 | 1.33 | 1.66 | 1.27 | 1.72 | 1.22 | 1.79 | 1.24 | 1.34 | 1.19 | 1.39 | 1.14 | 1.45 | 1.09 | 1.52 | 1.03 | 1.58 |
| 40 | 1.44 | 1.54 | 1.39 | 1.60 | 1.34 | 1.66 | 1.29 | 1.72 | 1.23 | 1.79 | 1.25 | 1.34 | 1.20 | 1.40 | 1.15 | 1.46 | 1.10 | 1.52 | 1.05 | 1.58 |
| 45 | 1.48 | 1.57 | 1.40 | 1.62 | 1.38 | 1.67 | 1.34 | 1.72 | 1.29 | 1.78 | 1.29 | 1.38 | 1.24 | 1.42 | 1.20 | 1.48 | 1.16 | 1.53 | 1.11 | 1.58 |
| 50 | 1.50 | 1.59 | 1.46 | 1.63 | 1.42 | 1.67 | 1.38 | 1.72 | 1.34 | 1.77 | 1.32 | 1.40 | 1.28 | 1.45 | 1.24 | 1.49 | 1.20 | 1.54 | 1.16 | 1.59 |
| 55 | 1.53 | 1.60 | 1.49 | 1.64 | 1.45 | 1.68 | 1.41 | 1.72 | 1.38 | 1.77 | 1.36 | 1.43 | 1.32 | 1.47 | 1.28 | 1.51 | 1.25 | 1.55 | 1.21 | 1.59 |
| 60 | 1.55 | 1.62 | 1.51 | 1.65 | 1.48 | 1.69 | 1.44 | 1.73 | 1.41 | 1.77 | 1.38 | 1.45 | 1.35 | 1.48 | 1.32 | 1.52 | 1.28 | 1.56 | 1.25 | 1.60 |
| 65 | 1.57 | 1.63 | 1.54 | 1.66 | 1.50 | 1.70 | 1.47 | 1.73 | 1.44 | 1.77 | 1.41 | 1.47 | 1.38 | 1.50 | 1.35 | 1.53 | 1.31 | 1.57 | 1.28 | 1.61 |
| 70 | 1.58 | 1.64 | 1.55 | 1.67 | 1.52 | 1.70 | 1.49 | 1.74 | 1.46 | 1.77 | 1.43 | 1.49 | 1.40 | 1.52 | 1.37 | 1.55 | 1.34 | 1.58 | 1.31 | 1.61 |
| 75 | 1.60 | 1.65 | 1.57 | 1.68 | 1.54 | 1.71 | 1.51 | 1.74 | 1.49 | 1.77 | 1.45 | 1.50 | 1.42 | 1.53 | 1.39 | 1.56 | 1.37 | 1.59 | 1.34 | 1.62 |
| 80 | 1.61 | 1.66 | 1.59 | 1.69 | 1.56 | 1.72 | 1.53 | 1.74 | 1.51 | 1.77 | 1.47 | 1.52 | 1.44 | 1.54 | 1.42 | 1.57 | 1.39 | 1.60 | 1.36 | 1.62 |
| 85 | 1.62 | 1 | | | | | | | | | | | | | | | | | | |

Si X n'est pas strictement exogène

- Durbin (1970) propose un autre test valable si X n'est pas strictement exogène, par exemple si le modèle contient y_{t-1} comme variable explicative.
- i) Estimer le modèle $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t$ et obtenir \hat{u}_t , $\forall t = 1, 2, \dots, n$.
- ii) Estimer le modèle $\hat{u}_t = \gamma_0 + \gamma_1 x_{t1} + \dots + \gamma_k x_{tk} + \rho \hat{u}_{t-1}$, $\forall t = 2, 3, \dots, n$ et obtenir $\hat{\rho}$ ainsi que $t_{\hat{\rho}}$.
- iii) Utiliser $t_{\hat{\rho}}$ de la manière usuelle pour tester $H_0 : \rho = 0$ contre $H_1 : \rho < \neq > 0$.
- → On régresse \hat{u}_t sur x_t et \hat{u}_{t-1} et donc on permet à chaque x_{tj} d'être corrélé avec u_{t-1} .
- → $t_{\hat{\rho}}$ a approximativement une distribution en t si n est grand.

Tester un $AR(q)$

- i) Estimer le modèle $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t$ et obtenir \hat{u}_t , $\forall t = 1, 2, \dots, n$.
- ii) Estimer le modèle
$$\hat{u}_t = \gamma_0 + \gamma_1 x_{t1} + \dots + \gamma_k x_{tk} + \rho_1 \hat{u}_{t-1} + \dots + \rho_q \hat{u}_{t-q},$$
$$\forall t = q+1, q+2, \dots, n.$$
- iii) Effectuer le F-test suivant $H_0 : \rho_1 = \dots = \rho_q = 0$ contre $H_1 : \text{un des } \rho_j \neq 0, \forall j = 1, \dots, q$.
- Si les x_t sont supposés strictement exogènes, on peut les omettre dans les étapes i) et ii).
- A noter que ces tests supposent $Var(u_t | x_t, u_{t-1}, \dots, u_{t-q}) = \sigma^2$. Il existe une version de ces tests robuste à l'hétéroskedasticité comme on le verra plus loin.

Tester un $AR(q)$

- Une alternative à F -test est d'utiliser un Lagrange Multiplier (LM) Test: $LM = (n - q)R_{\hat{u}}^2$, où $R_{\hat{u}}^2$ est le R^2 de la régression
$$\hat{u}_t = \gamma_0 + \gamma_1 x_{t1} + \dots + \gamma_k x_{tk} + \rho_1 \hat{u}_{t-1} + \dots + \rho_q \hat{u}_{t-q}, \forall t = q + 1, q + 2, \dots, n.$$
- Sous H_0 , $LM \sim \chi_q^2$ asymptotiquement.
- Ce test est connu sous le nom de test de **Breusch-Godfrey**.
- Il est test est disponible en Eviews (avec le F -test).
- Exemple: **Taux d'intérêt US obligataire à 3 mois.**
$$ci3_t = \alpha_0 + \alpha_1 ci3_{t-1}.$$

Correction pour corrélation sérielle

- Si on détecte de la corrélation sérielle, on peut modifier le modèle initial pour tenter d'obtenir un modèle **dynamiquement complet** (ex: $AR(1) \rightarrow AR(2)$).
- Dans certains cas nous ne sommes pas intéressé par modéliser cette dynamique → l'intérêt réside plutôt dans les autres variables incluses dans le modèle.
- Mais l'inférence est compromise → Que faire ?
 - → Calculer des écart-types robustes à n'importe quelle forme de corrélation sérielle.

Écart-types robustes

- Considérons le modèle $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t$, $t = 1, \dots, n$.
- Comment obtenir un écart-type pour $\hat{\beta}_1$ robuste à la corrélation sérielle ?
- $x_{t1} = \delta_0 + \delta_2 x_{t2} \dots + \delta_k x_{tk} + r_t$, où $E(r_t) = 0$ et $Corr(r_t, x_{tj}) = 0, \forall j \geq 2$.
- Il est possible de montrer que $Avar(\hat{\beta}_1) = \frac{Var(\sum_{i=1}^n r_t u_t)}{\left(\sum_{i=1}^n E(r_t^2)\right)^2}$, où $Avar$ dénote la variance asymptotique.
- Sous l'hypothèse **TS.5'**, $a_t \equiv r_t u_t$ est non corrélé sériellement et donc la formule traditionnelle de $Var(\hat{\beta}_1)$ est valide. Par contre si **TS.5'** ne tient pas, $Avar(\hat{\beta}_1)$ doit tenir compte de la corrélation entre a_t et $a_s \ \forall t \neq s$.

Écart-types robustes

- Newey et West (1987) et Wooldridge (1989) ont montré que $Avar(\hat{\beta}_1)$ peut être estimé de la manière suivante.
- i) Estimer par MCO $y_t = \beta_0 + \beta_1 x_{t1} + \dots + \beta_k x_{tk} + u_t$,
 $t = 1, \dots, n$. $se(\hat{\beta}_1)$ dénote l'écart-type de $\hat{\beta}_1$ et $\hat{\sigma}$ l'écart-type de \hat{u}_t .
- ii) Estimer la régression auxiliaire:
 $x_{t1} = \delta_0 + \delta_2 x_{t2} \dots + \delta_k x_{tk} + r_t$.
- iii) Calculer $\hat{a}_t = \hat{r}_t \hat{u}_t$, $\forall t = 1, \dots, n$.
- iv) Pour une valeur $g > 0$ donnée, calculer:
$$\hat{v} = \sum_{t=1}^n \hat{a}_t^2 + 2 \sum_{h=1}^g [1 - h/(g+1)] \left(\sum_{t=h+1}^n \hat{a}_t \hat{a}_{t-h} \right).$$

→ g contrôle la "quantité" de corrélation sérielle que nous permettons.

Écart-types robustes

- Ex: $g = 1, \hat{v} = \sum_{t=1}^n \hat{a}_t^2 + \sum_{t=2}^n \hat{a}_t \hat{a}_{t-1}$.
- v) L'écart-type robuste à la corrélation sérielle de $\hat{\beta}_1$ est:
$$se^*(\hat{\beta}_1) = [se(\hat{\beta}_1)/\hat{\sigma}^2] \sqrt{\hat{v}}$$
.
- On peut montrer que cet estimateur est aussi robuste à toute forme d'hétéroskedasticité → cas plus général de ce qui est exposé au Chapitre 8.
- Comment choisir g ?
- La théorie nous dit que g doit croître avec n .

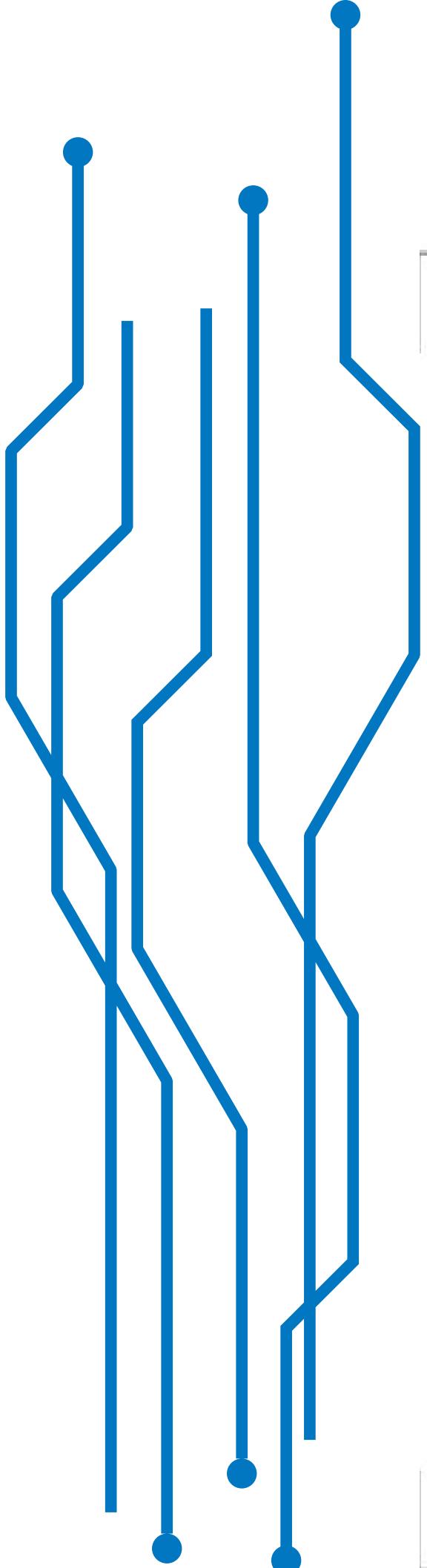
Choix de g

- Certains travaux ont suggéré pour:
 - des données annuelles $\rightarrow g = 1, 2$;
 - des données trimestrielles $\rightarrow g = 4, 8$;
 - des données mensuelles $\rightarrow g = 12, 24$.
- Newey et West (1987) recommandent de prendre la partie entière de $4(n/100)^{2/9} \rightarrow$ implémenté en Eviews.
- Exemple: **Taux d'intérêt US obligataire à 3 mois.**

$$ci3_t = \alpha_0 + \alpha_1 ci3_{t-1}.$$

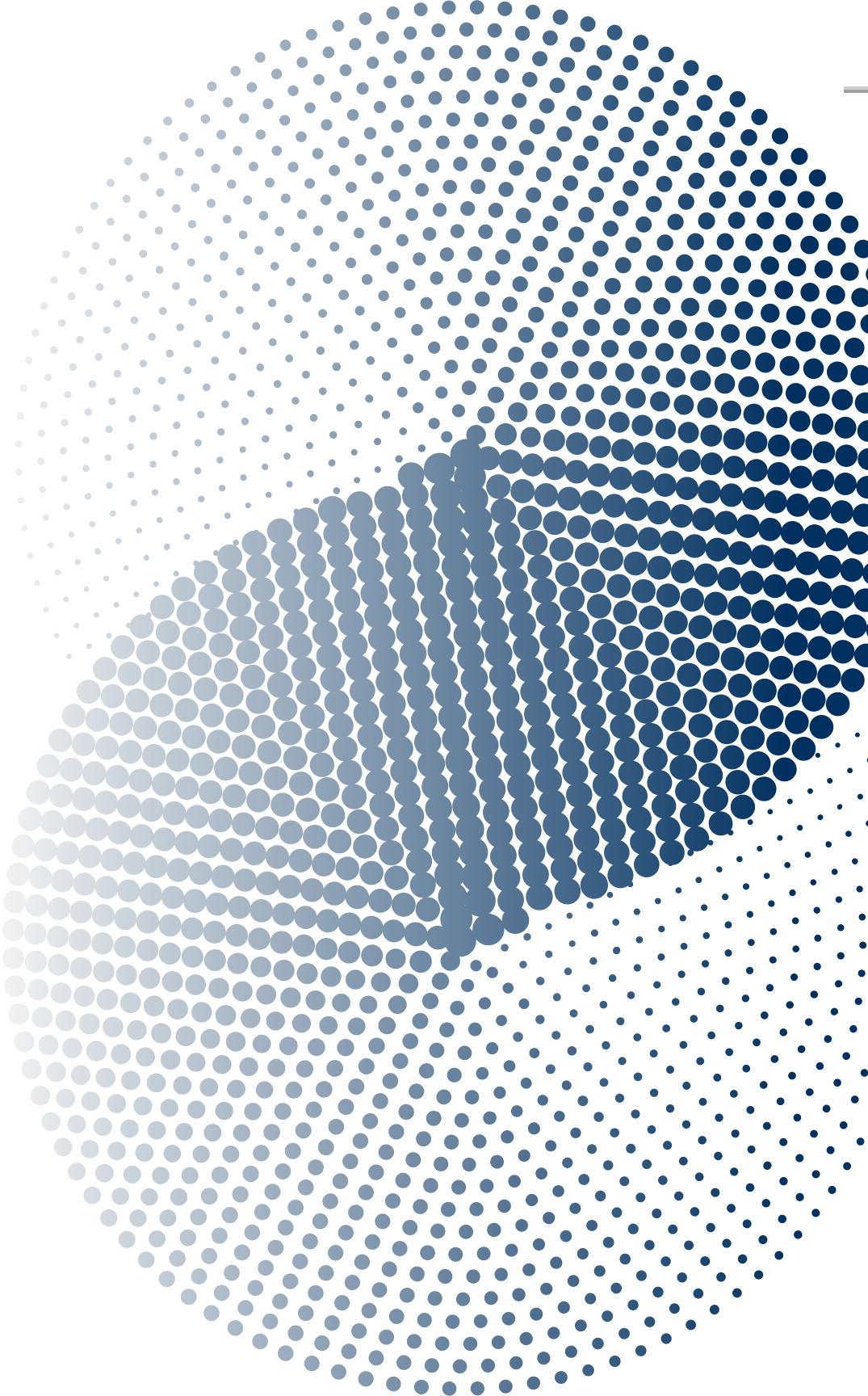
Hétéroscédasticité

- Pour beaucoup de séries temporelles, l'hypothèse **TS.4** ou **TS.4'** d'homoscédasticité est violée.
- Exemple: **Rendements journaliers du NYSE**
 - → n'affecte pas le caractère non-biaisé ou convergent des MCO mais invalide l'inférence statistique traditionnelle.
- Il existe deux manières d'aborder le problème lié à l'hétéroscédasticité.
 - i) Corriger les écart-types pour effectuer de l'inférence correctement.
 - ii) Modéliser la dynamique présente dans la variance.



Écart-types robustes à la White (1980)

- White (1980) propose une méthode permettant de rendre les écart-types robustes à toute forme d'hétéroscédasticité.
- → offre une solution intéressante, pour autant que l'intérêt ne se porte que sur la modélisation de la moyenne conditionnelle.
- Eviews, ainsi que beaucoup d'autres logiciels économétriques, offre cette option.
- Il est possible de montrer que la formule de White (1980) est un cas particulier de la formule de Newey et West (1987) qui permet de tenir compte également d'une possible autocorrélation des résidus.
- Exemple: AR(1) sur NYSE



Tester la présence d'hétéroscédasticité

- Avant de modéliser la dynamique présente dans la variance, il est judicieux de tester la présence d'hétéroscédasticité afin d'avoir une meilleure idée de la spécification à adopter.
- Pour appliquer les tests présentés ci-dessous il faut supposer que les résidus u_t sont non corrélé sériellement → tester avant.
- **Test de Breusch-Pagan:**
$$u_t^2 = \delta_0 + \delta_1 x_{t1} + \dots + \delta_k x_{tk} + v_t; H_0 : \delta_1 = \dots = \delta_k = 0.$$
- Pour utiliser un F -test, il faut que les écart-types des MCO soient valables et donc que v_t satisfasse **TS.4** ou **TS.4'** et **TS.5** ou **TS.5'**.
- Exemple:
— AR(1) sur NYSE: Estimer $\hat{u}_t^2 = \alpha_0 + \alpha_1 return_{t-1} + e_t$

Modèles ARCH

- Considérons un modèle simple: $y_t = \beta_0 + \beta_1 z_t + u_t$.
- Une caractéristique largement admise des séries financières à fréquence élevée est que la variance n'est pas constante au court du temps et qu'il existe des **grappes de volatilité**.
 - Si la variance en t est grande, elle le sera probablement demain et les jours qui suivent.
 - Si la variance en t est petite, elle le sera probablement demain et les jours qui suivent.
- Engle (1982) a proposé un modèle appelé ARCH: Autoregressive Conditional Heteroskedasticity.

Modèles ARCH

- La caractéristique du modèle ARCH(1) est que:
 $E(u_t^2|u_{t-1}, u_{t-2}, \dots) = E(u_t^2|u_{t-1}) = \alpha_0 + \alpha_1 u_{t-1}^2$, alors que $E(u_t^2|u_{t-1}, u_{t-2}, \dots) = 0$.
- Les u_t sont non corrélés sériellement alors que les u_t^2 le sont.
- Conditions de positivité: $E(u_t^2|u_{t-1}) > 0, \forall t \rightarrow \alpha_0 > 0$ et $\alpha_1 \geq 0$.
- Si $\alpha_1 = 0 \rightarrow$ homoscédasticité.
- on peut tester la présence d'effets ARCH en estimant ce modèle (sur \hat{u}_t), voir une version plus étendue (plus de retards).
- On peut tester $\alpha_1, \dots, \alpha_q = 0$ en utilisant un LM – test ou F – test.