



**Commission économique des Nations unies pour l'Afrique
(CEA)/
Institut des Nations unies pour le développement
économique et la planification (IDEP)**

**Titre du cours : Modélisation macroéconomique pour
le développement durable**

**MODULE IV : MODÉLISATION ET PRÉVISION DE
LA POLITIQUE ÉCONOMIQUE À L'AIDE DE
L'ANALYSE DE RÉGRESSION, DE SÉRIES
TEMPORELLES ET DE VAR**

Professeur Sylvain H. Boko

2025

Table des matières

| | |
|--|-----------|
| 1. Introduction | 4 |
| 1.1. Objectifs du module | 4 |
| 1.2. Résultats attendus de l'apprentissage | 5 |
| 2. Le modèle de régression linéaire simple | 5 |
| 2.1. Importance des données dans l'analyse de régression..... | 6 |
| 2.2. Application de l'analyse de régression linéaire : Une illustration | 7 |
| 2.2.1 Étapes de l'analyse de régression..... | 8 |
| 3. Régression multiple | 16 |
| 3.1. Le modèle | 16 |
| 3.2 Estimation | 17 |
| 3.3. Interprétation des résultats..... | 19 |
| 4. Analyse des séries temporelles | 20 |
| 4.1. La définition..... | 20 |
| 4.3. Types de modèles de séries temporelles | 20 |
| 4.4. Modèles d'autorégression | 25 |
| 4.4.1 Estimation des modèles autorégressifs (AR) | 25 |
| 4.5. Modélisation et prévision des séries temporelles | 29 |
| 4.5.1. Modèles autorégressifs (AR)..... | 29 |
| 4.5.2. Modèles de moyenne mobile (MA)..... | 29 |
| 4.5.3. Séries de données intégrées..... | 30 |
| 4.6. Étude de cas : PIB réel du Nigeria (RGDP) sur la période 1970-2017..... | 30 |
| 4.6.1. Visualisation et test des racines unitaires | 31 |
| 4.6.2. Modélisation du (log) RGDP du Nigeria en première différence comme ARMA(1,1) | 33 |
| 5. Modèles vectoriels autorégressifs (VAR) | 34 |
| 5.1. Contexte..... | 34 |
| 5.2 Description du modèle | 34 |
| 5.3 Estimation d'un modèle VAR | 35 |
| 5.4 Analyse des résidus et fonctions de réponse impulsionale | 36 |

| | |
|---|-----------|
| 5.5 Performance globale de l'approche VAR | 37 |
| Références | 38 |

1. Introduction

L'analyse de régression est un outil statistique puissant utilisé pour examiner la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Elle est largement utilisée dans divers domaines, notamment l'économie, la finance, la biologie et les sciences sociales, pour comprendre et prédire les modèles et les tendances.

En économie, l'analyse de régression est un instrument fondamental pour évaluer les théories, tester les hypothèses et éclairer les décisions politiques. Les économistes utilisent la régression pour étudier des phénomènes tels que l'inégalité des revenus, les modèles de consommation, le comportement du marché et les effets des interventions fiscales et monétaires. Qu'il s'agisse des subtilités microéconomiques ou des dynamiques macroéconomiques, l'application de l'analyse de régression permet aux économistes de faire le lien entre les constructions théoriques et les données du monde réel, ouvrant ainsi la voie à des conclusions fondées sur des preuves et à des solutions pratiques.

L'analyse de régression consiste essentiellement à adapter un modèle aux données observées. La forme la plus simple est le modèle de régression linéaire, qui suppose une relation linéaire entre la variable dépendante et les variables indépendantes. Ce modèle peut être étendu à la régression multiple, où plusieurs variables indépendantes sont prises en compte simultanément.

Les principaux objectifs de l'analyse de régression sont les suivants :

1. **Estimation** : Détermination des coefficients qui décrivent le mieux la relation entre les variables dépendantes et indépendantes.
2. **Prédiction** : Utilisation du modèle pour prévoir les valeurs futures de la variable dépendante sur la base des nouvelles valeurs des variables indépendantes.
3. **Inférence** : Tirer des conclusions sur la population dont les données de l'échantillon sont tirées, y compris les tests d'hypothèse et les intervalles de confiance.

L'analyse de régression implique également plusieurs hypothèses, telles que la linéarité, l'indépendance, l'homoscédasticité et la normalité des résidus. L'évaluation de ces hypothèses est cruciale pour garantir la validité et la fiabilité du modèle.

En comprenant et en appliquant l'analyse de régression, les chercheurs et les analystes peuvent développer un cadre analytique pour estimer et prévoir l'impact potentiel de divers facteurs, y compris les politiques gouvernementales.

Le module passera d'abord en revue les principes de la régression multiple, puis s'étendra à l'analyse des séries temporelles et se terminera par un examen des modèles VAR.

1.1. Objectifs du module

Les objectifs du module IV sont les suivants :

1. **Comprendre les bases de la régression** - Introduire les concepts fondamentaux de l'analyse de régression, y compris son objectif et ses applications dans divers domaines.

2. **Explorer les différents types de régression** - Examiner la régression linéaire simple et multiple, les séries chronologiques et le VAR.
3. **Construction et interprétation de modèles** - Apprenez à construire des modèles de régression, à interpréter les coefficients et à évaluer la performance des modèles à l'aide de métriques.
4. **Hypothèses et diagnostics** - Discuter des principales hypothèses des modèles de régression (par exemple, linéarité, homoscédasticité, indépendance, stationnarité) et des techniques permettant de vérifier si elles sont respectées.
5. **Appliquer la régression à des problèmes réels** - Utiliser des techniques de régression pour analyser des ensembles de données réels, faire des prédictions et en tirer des enseignements significatifs.

1.2. Résultats attendus de l'apprentissage

Les principaux résultats de l'apprentissage sont les suivants

1. **Expliquer les principes de l'analyse de régression** - Démontrer une bonne compréhension des concepts de régression, y compris les variables dépendantes et indépendantes, et les extensions aux séries temporelles et aux méthodes VAR.
2. **Appliquer différentes techniques de régression** - Utiliser différentes méthodes de régression pour modéliser les relations entre les données.
3. **Interpréter les résultats des modèles de régression** - Analyser les coefficients, les niveaux de significativité et les mesures d'adéquation pour obtenir des informations significatives.
4. **Évaluer et valider les modèles de régression** - Effectuer des tests de diagnostic pour vérifier les hypothèses telles que la multicollinéarité, l'hétéroscédasticité, la normalité, la stationnarité, etc.
5. **Traiter les problèmes du monde réel à l'aide de l'analyse de régression** - Appliquer des modèles de régression à des données du monde réel.

2. Le modèle de régression linéaire simple

Un modèle de régression linéaire évalue la relation entre une variable dépendante (ou endogène) et une ou plusieurs variables indépendantes (ou exogènes). Lorsque l'analyse de régression comprend une variable dépendante et une variable indépendante, on parle d'analyse de régression univariée. Inversement, lorsque le modèle comprend une variable dépendante mais plusieurs variables indépendantes, on parle de régression multiple. Nous commencerons notre discussion par la régression linéaire simple. Un modèle de régression linéaire univarié peut être représenté par l'équation suivante :

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon \quad (1)$$

où,

Y = Variable dépendante (endogène)

X_1 = Variable indépendante (exogène/explorative)

β_0 = Constante ou l'ordonnée à l'origine (mesure la valeur de Y lorsque $X_1 = 0$)

β_1 = Pente de la droite (mesure le changement de Y en réponse à un changement de X_1)

ϵ = Résidu ou terme d'erreur (mesure la proportion de la variable dépendante inexplicable par la variable exogène)

L'**équation 1** décrit une relation entre la variable endogène Y et la variable exogène X_1 . Dans l'analyse de régression, la variable endogène est la variable dépendante influencée par d'autres variables du modèle. En revanche, la variable exogène est une variable indépendante qui n'est pas affectée par les autres variables du modèle. La variable exogène est également appelée variable explicative car elle explique les variations de la variable dépendante. L'analyse de régression vise à déterminer, par des méthodes d'estimation, la direction et la force de la relation entre les variables endogènes et exogènes, mesurée par le coefficient β_1 .

2.1. Importance des données dans l'analyse de régression

Les données constituent l'épine dorsale de l'analyse de régression, car elles permettent d'identifier et de quantifier les relations entre les variables. En l'absence de données précises et pertinentes, toute analyse effectuée manquerait de validité et de significativité. Les points suivants soulignent l'importance cruciale des données :

1. **Fondement des modèles** : Les modèles de régression s'appuient sur des données pour définir la relation entre les variables dépendantes et indépendantes. Sans données, il n'y a pas de base pour construire ou tester le modèle.
2. **Perspectives et interprétations** : Les données fournissent les preuves nécessaires pour découvrir des modèles, des tendances et des liens de causalité. La richesse et la qualité des données ont un impact direct sur la fiabilité et la profondeur des conclusions tirées.
3. **Pouvoir prédictif** : Dans la régression, les données sont utilisées pour faire des prédictions sur les résultats futurs. La précision de ces prédictions dépend de la pertinence et de la précision des données d'entrée.
4. **Validation** : Les modèles de régression doivent être validés par rapport à des données réelles afin de garantir leur précision et leur applicabilité. Des données de mauvaise qualité peuvent conduire à des résultats invalides ou trompeurs.
5. **Applications économiques** : En économie, les données représentent des phénomènes réels tels que le revenu, l'emploi, l'inflation et le commerce. Des données précises permettent aux économistes de tester des théories, d'évaluer des politiques et de comprendre des systèmes complexes.

Essentiellement, les données transforment des concepts abstraits et des hypothèses en résultats mesurables et exploitables. Mais il est important de noter que des erreurs dans la collecte, la mesure ou la saisie des données peuvent fausser les résultats de la régression.

Vous trouverez ci-dessous un aperçu des types de données couramment utilisés :

1. **Données d'observation** : Il s'agit de données collectées à partir de scénarios réels, tels que des enquêtes, des dossiers administratifs ou des indicateurs économiques.
2. **Données expérimentales** : Dans certains cas, la régression utilise des données provenant d'expériences contrôlées où les variables sont manipulées afin d'observer leurs effets.
3. **Données de panel** : Données qui suivent les mêmes entités (par exemple, des individus, des entreprises ou des pays) au fil du temps, ce qui permet de mieux comprendre les changements et les tendances.
4. **Données transversales** : Un instantané de différentes entités à un moment donné (par exemple, les revenus des ménages en 2023).
5. **Données de séries temporelles** : Données collectées sur une séquence d'intervalles de temps (par exemple, le PIB trimestriel, les taux d'inflation mensuels).

2.2. Application de l'analyse de régression linéaire : Une illustration

Considérons un scénario dans lequel le gouvernement met en œuvre un programme de formation professionnelle dans le cadre d'une politique de relance. Le tableau 1 présente les dépenses mensuelles du gouvernement pour le programme de formation professionnelle et le nombre correspondant d'emplois créés en 2020. Comment pouvons-nous estimer la relation linéaire entre le nombre d'emplois jeunes créés et les dépenses du gouvernement sur la base de l'équation (1), et comment pouvons-nous mesurer la force de cette relation ? Les réponses à ces questions sont fournies dans la section ci-dessous.

Tableau 1. Dépenses mensuelles du gouvernement pour le programme de formation à l'emploi (Illustration)

| Période | Jobs_Y | G-Exp |
|---------|--------|-------|
| 2020::1 | 21 | 8350 |
| 2020::2 | 180 | 23755 |
| 2020::3 | 50 | 13455 |

| | | |
|-----------------|------------|--------------|
| 2020::4 | 195 | 21100 |
| 2020::5 | 98 | 15000 |
| 2020::6 | 44 | 12500 |
| 2020::7 | 171 | 20700 |
| 2020::8 | 135 | 19722 |
| 2020::9 | 120 | 16115 |
| 2020::10 | 75 | 13100 |
| 2020::11 | 106 | 15670 |
| 2020::12 | 198 | 25300 |

2.2.1 Étapes de l'analyse de régression

2.2.1.1 Visualisation des données et des droites de régression

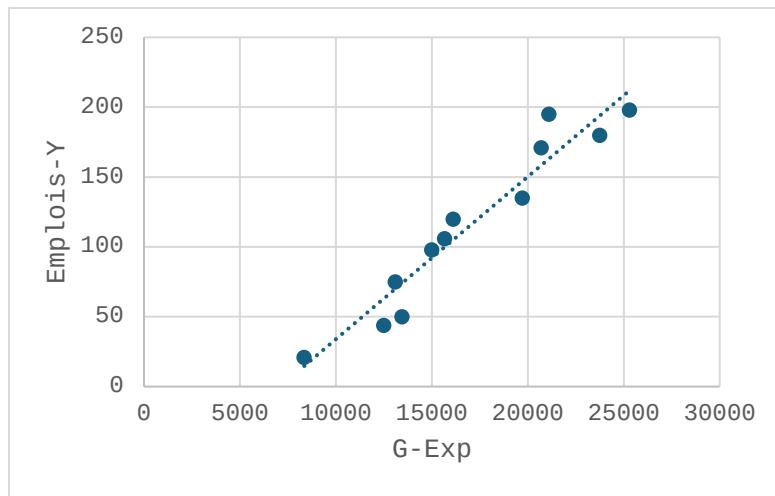
La visualisation des données joue un **rôle crucial** dans l'analyse de régression en aidant les analystes à explorer, comprendre et communiquer les relations entre les variables. Par exemple, les diagrammes de dispersion permettent d'**évaluer visuellement les corrélations** entre les variables indépendantes et dépendantes avant d'appliquer un modèle de régression, et aident à déterminer si la relation est **linéaire ou non linéaire**. La visualisation des données peut également mettre en évidence des modèles sous-jacents, des grappes et des tendances qui peuvent ne pas être immédiatement évidents à partir de données numériques brutes.

La figure 1 illustre la représentation graphique des données présentées dans le tableau 1, accompagnées d'une droite de tendance. L'axe horizontal représente la variable explicative, G-Exp, tandis que l'axe vertical désigne la variable dépendante, Jobs-Y. Le nuage de points indique une corrélation positive entre Jobs-Y et G-Exp, ce qui est corroboré par la droite de tendance à pente positive. Douze observations représentent les observations mensuelles des combinaisons Jobs-Y et G-Exp. Comme prévu, certaines observations sont positionnées le long de la droite de tendance, tandis que d'autres se situent au-dessus ou au-dessous de celle-ci.

2.2.1.2 Procédure d'estimation

L'équation (1) représente un modèle linéaire simple, et le coefficient β_1 peut être estimé à l'aide de la méthode des moindres carrés ordinaires (MCO). Les MCO sont une technique de base de l'analyse de régression utilisée pour estimer les coefficients d'un modèle linéaire. L'objectif des MCO est de trouver la droite qui minimise la somme des carrés des différences entre les valeurs observées et les valeurs prédites par le modèle. Cette méthode est couramment utilisée en raison de sa simplicité et de son efficacité à identifier la relation entre les variables dépendantes et indépendantes.

Figure 1. Représentation visuelle de Jobs-Y par rapport à G-Exp

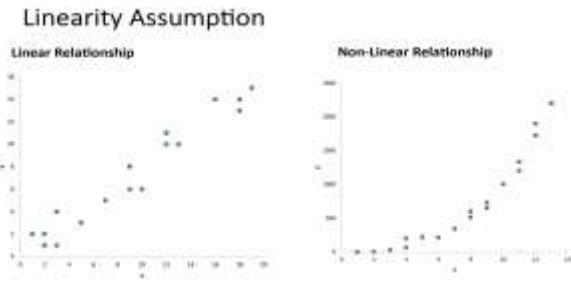


2.2.1.3 Hypothèses de l'analyse de régression par les MCO

La régression par les moindres carrés ordinaires (MCO) repose sur une série d'hypothèses pour garantir la validité de ses résultats. Les principales hypothèses sont les suivantes :

1. **Linéarité** : Le modèle suppose que la relation entre les variables indépendantes (X) et la variable dépendante (Y) est linéaire. La violation de cette hypothèse peut signifier qu'il n'y a pas de relation entre X et Y, ou que la relation est courbée. La **figure 2** présente des diagrammes de dispersion de deux relations entre deux variables : une relation linéaire et une relation non linéaire. Le panneau de gauche illustre une relation linéaire, tandis que celui de droite montre une relation non linéaire. Si l'on soupçonne une non-linéarité, une transformation est appliquée à la variable dépendante et/ou indépendante, en utilisant par exemple les méthodes du logarithme, de la racine carrée ou de la réciproque.

Figure 2. Relations linéaires et non linéaires



2. **Distribution des termes d'erreur** : Le modèle suppose que les observations sont indépendants les uns des autres, ce qui signifie qu'il n'y a pas de corrélation entre les résidus (erreurs) des observations. Plus précisément, on suppose que les termes d'erreur sont **indépendants et identiquement distribués (iid)**. L'hypothèse d'**indépendance des termes d'erreur** implique qu'aucun terme d'erreur n'est influencé par les autres, ce qui signifie qu'il n'y a pas de corrélation entre eux

Les termes d'erreur **distribués de manière identique** doivent suivre la même distribution de probabilité, ce qui implique qu'ils ont la même moyenne et la même variance. Dans les modèles de régression, l'hypothèse de *termes d'erreur identiques* garantit que les prédictions du modèle sont impartiales et fiables. Cette hypothèse simplifie l'analyse mathématique et facilite l'application de techniques statistiques telles que les tests d'hypothèse, car de nombreuses méthodes statistiques, y compris le **théorème de la limite centrale**, reposent sur des hypothèses *iid* pour obtenir des résultats significatifs.

3. **Homoscédasticité** : Cette hypothèse s'applique à la **variance** des résidus (termes d'erreur), qui est supposée être constante pour tous les niveaux des variables indépendantes. Plus précisément, le modèle suppose que la dispersion des termes d'erreur ne doit pas changer lorsque les valeurs prédites de (Y) augmentent ou diminuent

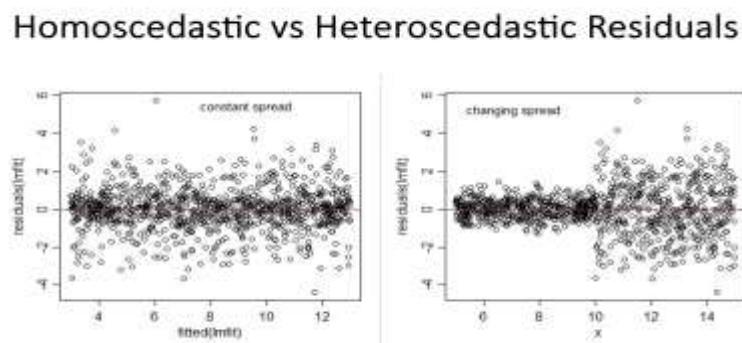
Cette hypothèse garantit que l'analyse de régression produit des estimations efficaces et sans biais. Lorsque l'homoscédasticité est respectée, les prédictions restent stables pour différentes valeurs de la variable indépendante. Les violations de l'homoscédasticité (connues sous le nom d'**hétéroscédasticité**) peuvent conduire à des conclusions erronées lors des tests d'hypothèses. Plus précisément, en présence d'hétéroscédasticité, résultats de la régression ne sont pas fiables. Par exemple, une variable estimée peut être considérée comme statistiquement significative dans le modèle alors qu'elle ne l'est pas (**voir figure 3**).

La **méthode de Breusch-Pagan** est généralement utilisée pour tester l'hétéroscédasticité. Ce test permet de déterminer si la variance des termes d'erreur dépend des variables indépendantes. L'**hypothèse nulle (H_0)** est que les termes d'erreur ont une **variance**

constante (homoscédasticité) ; l'**hypothèse alternative (H_1)** est que les termes d'erreur ont une **variance non constante** (hétéroscléasticité). Si la valeur p de la statistique du test du **Khi-deux(χ^2)** qui en résulte est **inférieure à 0,05**, l'hypothèse nulle est rejetée, ce qui indique l'hétéroscléasticité (voir la **section 2.2.1.6** sur les tests d'hypothèse).

En cas d'hétéroscléasticité, les solutions comprennent la transformation de la variable dépendante (par exemple en utilisant la transformation logarithmique), la redéfinition de la variable dépendante (**par exemple croissance contre stock**) ou l'utilisation d'une **méthode de régression pondérée**.

Figure 3. Comparaison des résidus homoscédastiques et hétéroscléastiques



4. **Multicollinéarité :** L'hypothèse de multicolinéarité est pertinente dans le cas de **multiples Régression** (voir section 3). Elle se produit lorsque deux ou plusieurs variables indépendantes d'un modèle de régression sont fortement corrélées, ce qui rend difficile la détermination de l'effet individuel de chaque variable sur la variable dépendante. La présence de multicolinéarité peut fausser l'estimation des coefficients. Par exemple, de petits changements dans les données peuvent entraîner de grandes fluctuations dans les coefficients de régression, ce qui les rend instables.

La multicolinéarité **réduit** également **l'interprétabilité** des résultats de la régression, car il devient difficile de déterminer quelle variable influence réellement la variable dépendante. Elle affaiblit la significativité statistique des résultats de l'estimation en gonflant les **erreurs types**, par exemple. Elle réduit donc la capacité du modèle à identifier les variables exogènes statistiquement significatives

- **Détection de la**

multicolinéarité peut être présente dans les scénarios suivants :

- Les coefficients de régression présentent des erreurs standard élevées

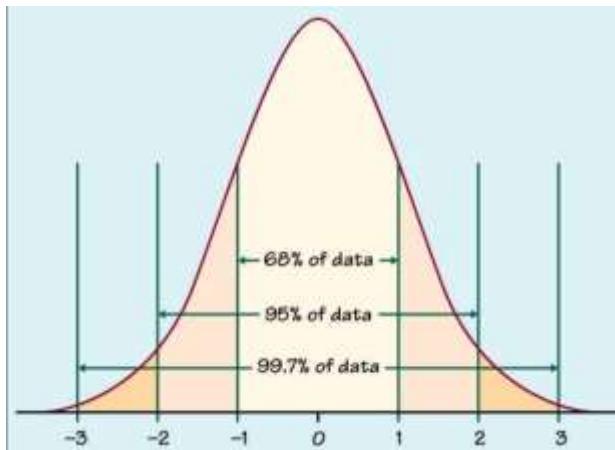
- Le modèle global est statistiquement significatif, mais aucun des coefficients individuels ne l'est.
 - L'inclusion de variables exogènes supplémentaires dans le modèle entraîne des changements importants dans les
 - Les coefficients présentent des signes (négatifs/positifscontraires à ce que l'on pourrait attendre sur la base de la théorie économique.
- **Test de multicolinéarité**
 - Le test de multicolinéarité consiste à calculer le **facteur d'inflation de la variance (VIF)** pour chaque variable exogène. La conclusion sur la multicolinéarité est formulée comme suit
 - Si $1 < \text{VIF} \leq 4$, il n'y a probablement pas de multicolinéarité.
 - Si $4 < \text{VIF} \leq 10$, soupçonner la multicolinéarité
 - Si $\text{VIF} > 10$, il y a une grave multicolinéarité.

5. Hypothèse de normalité : Cette hypothèse concerne la distribution des résidus (différences entre les valeurs observées et prédites) obtenus à la suite d'une estimation par régression. En particulier, le modèle suppose que les résidus de l'estimation suivent une distribution normale (ou gaussienne). L'hypothèse de normalité est généralement représentée par une courbe en forme de cloche, parfaitement symétrique autour de la moyenne (voir **figure 4**)

L'hypothèse de normalité est particulièrement importante pour les tests d'hypothèse et la détermination des intervalles de confiance. La **règle empirique (ou règle 68-95-99.7)** suggère que :

- **Environ 68 % des données se situent à moins d'un écart-type de la moyenne.**
- **Environ 95 % des données se situent à moins de deux écarts types de la moyenne.**
- **Environ 99,7 % des données se situent à moins de trois écarts types de la moyenne.**

Figure 4. Une distribution normale



2.2.1.4 Qualité de l'ajustement : Mesures et

- **Adj- R^2 ou \bar{R}^2** : L'adéquation globale du modèle est mesurée par \bar{R}^2 , qui indique dans quelle mesure le modèle de régression explique la variabilité de la variable dépendante. Plus précisément, \bar{R}^2 aide à comprendre la force de la relation entre les variables dépendantes et indépendantes. Par exemple, une valeur \bar{R}^2 de 0,8 suggère que 80 % de la variance de la variable dépendante est expliquée par la ou les variables indépendantes, ce qui indique une relation forte. En général, lorsque l'on compare différents modèles de régression, une valeur \bar{R}^2 plus élevée indique une meilleure adéquation, ce qui signifie que le modèle explique une plus grande partie de la variance de la variable dépendante.¹
- **Statistique t** : La statistique t dans l'analyse de régression est utilisée pour déterminer si une variable explicative individuelle a une relation statistiquement significative avec la variable dépendante, ce qui résulte de l'évaluation du fait que le coefficient de la variable yjr est significativement différent de zéro. Dans l'analyse de régression, la formule de la statistique t est la suivante :

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i}{SE_{\hat{\beta}_i}}$$

pour $i = 0 \dots n$, où n = nombre de variables exogènes dans le modèle, y compris la constante, et SE est l'erreur standard de $\hat{\beta}_i$.

Pour la vérification de l'hypothèse, et dans le cas d'un test à deux queues, si $|t_{\beta_i}| \geq 1,96$, on peut en déduire que $\hat{\beta}_i$ est statistiquement significatif au **niveau de 5 %**. En outre, si $|t_{\beta_i}| \geq 2,58$, on peut en déduire que $\hat{\beta}_i$ est statistiquement significatif au niveau de **1 %**.

¹ Bien que le R au carré soit une mesure utile, il présente des limites. Il n'indique pas si les variables indépendantes sont liées de manière causale à la variable dépendante et ne tient pas compte de la complexité du modèle. En outre, une valeur élevée de R-carré ne signifie pas nécessairement que le modèle est approprié ; il peut être surajusté par rapport aux données. Elle doit être utilisée avec d'autres mesures lors de la sélection du modèle.

En général (mais pas nécessairement toujours), les variables exogènes sélectionnées pour le modèle final doivent être celles dont les statistiques t sont statistiquement significatives (c'est-à-dire celles pour lesquelles les valeurs p sont inférieures aux niveaux de significativité souhaités).

- **Le critère d'information d'Akaike (AIC) :** L'AIC est un outil de sélection de modèle utilisé pour évaluer la performance d'un modèle par rapport à d'autres modèles alternatifs. Il repose sur l'hypothèse que le modèle le mieux adapté est celui qui explique la plus grande quantité de variations en utilisant le moins de variables indépendantes possibles (**parcimonie**). Lorsque l'on compare deux modèles, par exemple, **celui dont l'AIC est le plus faible est le mieux adapté.**
- **La statistique F :** Dans l'analyse de régression ou l'analyse de la variance (ANOVA), la statistique F détermine si le modèle de régression global est statistiquement significatif

Une **valeur F élevée** suggère que les variables indépendantes expliquent de manière significative la variation de la variable dépendante, tandis qu'une **valeur F faible** indique que les variables indépendantes n'ont peut-être pas d'impact significatif. La **valeur p** associée à la statistique F pour chaque estimation de régression détermine son niveau de significativité.

2.2.1.5 Résultats de l'estimation

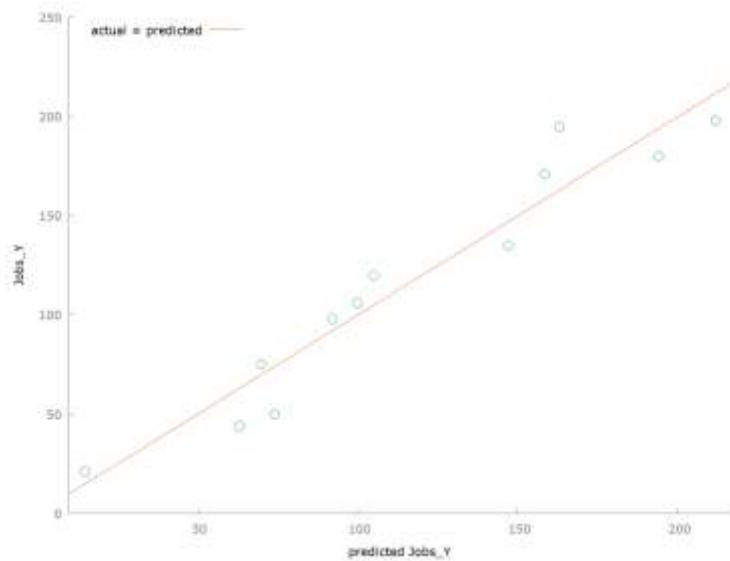
L'équation 1 a été estimée par les MCO, avec les données du tableau 1. Les résultats de l'estimation sont présentés dans le **tableau 2**. Le β_1 (ou $\hat{\beta}_1$) estimé est de **0,012**, qui adopte un signe positif, confirmant ainsi l'analyse visuelle de la figure 1. La significativité du coefficient estimé est mesurée par le **rapport t**, qui pour cet exemple est de **11,24**, avec une **valeur p de <0,0001**, indiquant un niveau élevé de significativité au niveau de 1 %.

Pour cet exemple, le R^2 est de 0,92, ce qui signifie que 92 % de la variance de Jobs-Y est expliquée par G-Exp, ce qui indique une forte relation entre les deux variables. **La figure 5** compare les observations réelles sur Jobs-Y aux valeurs prédictes sur la base du modèle estimé. Il semble qu'il y ait un alignement étroit entre les valeurs observées de la variable dépendante et la droite de régression, ce qui indique à nouveau une forte adéquation du modèle estimé.

Tableau 2. Résultats de l'estimation de la régression des Jobs-Y sur G-Exp

| Model 1: OLS, using observations 1-12 | | | | | |
|---------------------------------------|-------------|------------|--------------------|-------------|--|
| Dependent variable: Jobs_Y | | | | | |
| | Coefficient | Std. Error | t-ratio | p-value | |
| const | -82.38 | 18.34 | -4.490 | 0.0012 *** | |
| G-Exp | 0.012 | 0.001 | 11.24 | <0.0001 *** | |
| Mean dependent var | 116.0833 | | S.D. dependent var | 61.15027 | |
| Sum squared resid | 3014.611 | | S.E. of regression | 17.36264 | |
| R-squared | 0.926710 | | Adjusted R-squared | 0.919382 | |
| F(1, 10) | 126.4452 | | P-value(F) | 5.37e-07 | |
| Log-likelihood | -50.18518 | | Akaike criterion | 104.3704 | |
| Schwarz criterion | 105.3402 | | Hannan-Quinn | 104.0113 | |

Figure 5 : Comparaison des valeurs réelles des Jobs-Y avec les valeurs prédites



2.2.1.6 Test d'hypothèse

Supposons que nous voulions tester l'hypothèse selon laquelle les dépenses publiques (G-Exp) n'ont pas d'impact sur l'emploi des jeunes (Jobs-Y). Quelles sont les étapes à suivre ?

- Soit H_0 l'**hypothèse nulle** selon laquelle GExp n'a pas d'impact sur les Jobs-Y et H_1 l'**hypothèse alternative** selon laquelle GExp a un impact significatif sur les Jobs-Y.

- La vérification de l'hypothèse se fait ensuite de la manière suivante :

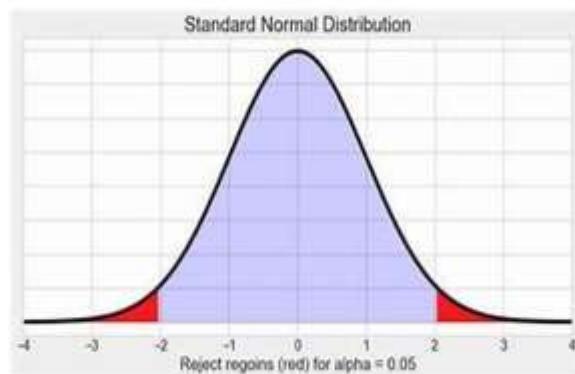
$H_0: \beta_1 = 0$ (Hypothèse nulle)

$H_1: \beta_1 \neq 0$ (Hypothèse alternative)

- Pour tester cette hypothèse, il faut utiliser la **statistique t associée à β_1** , et la comparer à la **valeur seuil (ou critique)** du **niveau de significativité** souhaité (5 % ou 1 % sont les valeurs les plus souvent utilisées. **Voir figure 6**).
 - Pour un niveau de significativité de 5 % (double queue), si $|t_{\beta_1}| \geq 1.96$, on rejette l'hypothèse nulle, sinon on accepte H_0
 - Pour un niveau de significativité de 1 % (2 queues), si $|t_{\beta_1}| \geq 2.58$, on rejette l'hypothèse nulle, sinon on accepte H_0

Figure 6. Distribution normale et test d'hypothèse

Normal distribution and hypothesis testing



- Conclusion : Dans notre exemple, $t_{\beta_1} = 11.24$, donc H_0 est rejeté à la fois au niveau significatif de 5% et de 1%, montrant ainsi que GExp a un impact significatif (et positif) sur les Jobs-Y.

3. Régression multiple

3.1. Le modèle

Une régression multiple suppose l'existence d'une relation linéaire entre une variable dépendante ou endogène et deux ou plusieurs variables indépendantes ou exogènes. Soit Y une variable

dépendante et $X_1, X_2, X_3 \dots X_n$ un ensemble de variables exogènes. Un modèle de régression multiple se présente alors sous la forme suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \dots + \beta_n X_n + \varepsilon \quad (2)$$

où,

β_0 = Constante ou l'ordonnée à l'origine (mesure la valeur de Y lorsque $X_1 = 0$)

$\beta_1 \dots \beta_n$ = Coefficients mesurant le changement de Y (ou réactivité) en réponse à un changement de $X_1, X_2 \dots X_n$

ε = Résidu ou terme d'erreur (mesure la proportion de la variable dépendante) non expliquée par $X_1, X_2 \dots X_n$

3.2 Estimation

Les données suivantes concernent 20 vendeurs et mesurent deux caractéristiques de chaque individu (telles que l'intelligence et l'extraversion), ainsi que les performances de vente hebdomadaires correspondantes (tableau 3).

Supposons que l'on vous confie la tâche d'étudier l'influence de l'**intelligence** et de l'**extraversion** sur les **performances commerciales hebdomadaires**. Sur la base de l'équation (2), le modèle prendra la forme explicite suivante :

$$Perf = \beta_0 + \beta_1 * Int + \beta_2 * Extr + \varepsilon \quad (3)$$

où,

$Perf$ = Performances commerciales hebdomadaires

Int = Intelligence

$Extr$ = Extraversion

En utilisant les données du tableau 3, le modèle (3) peut être estimé avec les MCO, et les résultats sont présentés dans le tableau 4. Le modèle estimé est représenté dans l'équation 4 (**avec les statistiques t entre parenthèses**) :

$$Perf = 993 + 8.22 Int + 49.71 Extr \quad (4)$$

(1.17) (2.53)

$$\bar{R}^2 = 0.27$$

Le tableau 4 montre que la **valeur p de** la variable explicative **Int est** égale à **1,17** et que celle de la variable explicative **Extr** est de **2,53**, avec une **valeur ajustée de R²** égale à **0,27**, ce qui signifie

que les variables explicatives du modèle de l'équation (3) n'expliquent ensemble que 27 % de la variance des performances commerciales hebdomadaires.

Tableau 3. Ventes hebdomadaires et caractéristiques personnelles

| Vendeur | Intelligence | Extraversion | \$ Ventes/semaine |
|---------|--------------|--------------|-------------------|
| 1 | 89 | 21 | 2625 |
| 2 | 93 | 24 | 2700 |
| 3 | 91 | 21 | 3100 |
| 4 | 122 | 23 | 3150 |
| 5 | 115 | 27 | 3175 |
| 6 | 100 | 18 | 3100 |
| 7 | 98 | 19 | 2700 |
| 8 | 105 | 16 | 2475 |
| 9 | 112 | 23 | 3625 |
| 10 | 109 | 28 | 3525 |
| 11 | 130 | 20 | 3225 |
| 12 | 104 | 25 | 3450 |
| 13 | 104 | 20 | 2425 |
| 14 | 111 | 26 | 3025 |
| 15 | 97 | 28 | 3625 |
| 16 | 115 | 29 | 2750 |
| 17 | 113 | 25 | 3150 |
| 18 | 88 | 23 | 2600 |
| 19 | 108 | 19 | 2525 |
| 20 | 101 | 16 | 2650 |

Tableau 4. Résultats de l'estimation - Régression multiple

| Model 2: OLS, using observations 1-20 | | | | |
|---------------------------------------|-------------|------------|---------|-----------------------------|
| Dependent variable: SalesWeek | | | | |
| | Coefficient | Std. Error | t-ratio | p-value |
| const | 993.92 | 788.099 | 1.261 | 0.2243 |
| Intelligence | 8.22 | 7.01256 | 1.172 | 0.2573 |
| Extroversion | 49.71 | 19.6337 | 2.532 | 0.0215 ** |
| Mean dependent var | 2980.000 | | | S.D. dependent var 390.3945 |
| Sum squared resid | 1874584 | | | S.E. of regression 332.0687 |
| R-squared | 0.352643 | | | Adjusted R-squared 0.276484 |
| F(2, 17) | 4.630316 | | | P-value(F) 0.024815 |
| Log-likelihood | -142.8604 | | | Akaike criterion 291.7208 |
| Schwarz criterion | 294.7080 | | | Hannan-Quinn 292.3040 |

3.3. Interprétation des résultats

Les résultats de l'estimation présentés dans cet exemple suggèrent que si l'**intelligence** peut prédire positivement la performance hebdomadaire d'un vendeur, son effet n'est pas statistiquement significatif sur la base de la valeur p correspondante (statistique t = 1,172, avec une valeur p = 0,25). À l'inverse, les résultats indiquent que l'**extraversion** est un facteur prédictif positif et significatif de la performance commerciale hebdomadaire, avec une statistique t = 2,53 et une valeur p = 0,02.

Il convient de noter qu'avec $\bar{R}^2 = 0,27$, les variables explicatives incluses dans le modèle ne représentent que 27 % de la variance des ventes hebdomadaires. Cela indique qu'il est possible

d'inclure d'autres variables qui ne sont pas prises en compte par le modèle actuel et qui pourraient améliorer son pouvoir prédictif.

4. Analyse des séries temporelles

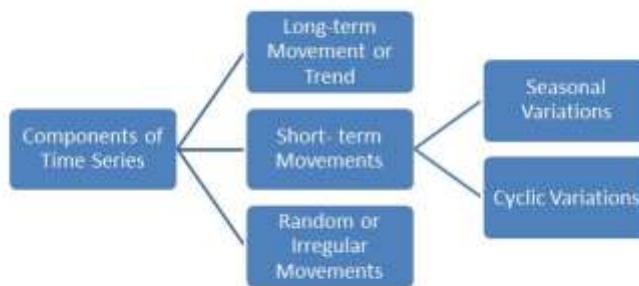
4.1. Définition

Une série chronologique est un ensemble de données ordonnées dans le temps, ce qui implique que la série de données est composée d'observations dont l'ordre est important. En ce sens, les séries temporelles présentent une dépendance temporelle et le fait de changer l'ordre peut modifier la significativité des données.

L'analyse des séries temporelles consiste à prévoir le comportement futur d'une variable sur la base de son évolution historique. Une différence entre l'analyse des séries temporelles et la régression linéaire standard est que les données des séries temporelles ne sont pas nécessairement indépendantes ni identiquement distribuées.

Les composantes d'une série temporelle sont illustrées dans la figure ci-dessous :

Figure 7. Composantes de la série temporelle



4.3. Types de modèles de séries temporelles

4.3.1. Une marche aléatoire

Une **marche aléatoire** dans l'analyse des séries temporelles fait référence à un processus stochastique dans lequel la valeur d'une variable change au fil du temps d'une manière imprévisible et dépend uniquement de sa valeur précédente et d'une perturbation aléatoire. Essentiellement, chaque étape de la série est déterminée par l'étape précédente plus un terme d'erreur aléatoire. Ce concept est souvent utilisé pour modéliser les cours des actions et d'autres

données financières, où les valeurs futures sont supposées être indépendantes des valeurs passées et suivre une trajectoire aléatoire.

Mathématiquement, une marche aléatoire peut être exprimée comme suit :

$$Y_t = Y_{t-1} + \varepsilon_t. \quad (5)$$

où,

Y_t est la valeur de la série au moment (t)

Y_{t-1} est la valeur précédente de la série

ε_t est un terme d'erreur stochastique (non systématique) qui est un bruit blanc (c'est-à-dire que ε_t est indépendamment et identiquement distribué avec une moyenne = 0 et une variance constante σ^2).

Une marche aléatoire est également appelée **processus intégré** (d'un ordre spécifié), c'est-à-dire un processus avec une **racine unitaire** ou présentant une **tendance stochastique**. Il s'agit d'un **processus sans retour à la moyenne** qui peut diverger de la moyenne dans un sens positif ou négatif

Caractéristiques principales

1. Rétention de la mémoire (propriété de Markov)

- La valeur actuelle dépend uniquement de la valeur précédente et d'un choc aléatoire.
- Pas de dépendance directe des valeurs passées au-delà de (Y_{t-1}).

2. Non-stationnarité

- La variance **augmente avec le temps**, ce qui rend le processus non stationnaire.
- La moyenne peut changer de manière imprévisible, ce qui rend les prévisions à long terme difficiles.

3. Accumulation de chocs aléatoires

- Chaque nouvelle valeur résulte de l'addition des valeurs passées et d'une nouvelle composante aléatoire.
- Cela conduit à des **fluctuations non limitées à long terme**.

Applications

- **Marchés financiers** : Les cours des actions suivent souvent une marche aléatoire en raison des influences imprévisibles du marché.
- **Physique** : Utilisé dans la modélisation des processus de diffusion.

- **Macroéconomie** : Décrit l'évolution d'indicateurs clés tels que le PIB lorsque les chocs s'accumulent au fil du temps.

Variations

- **Marche aléatoire avec dérive** : Ajoute une composante de tendance constante (α) pour introduire une croissance ou une décroissance systématique.
- **Marche aléatoire avec tendance déterministe** : Inclut une tendance structurée dépendant du temps (β_t), influençant la trajectoire.

4.3.2. Une marche aléatoire avec une dérive

Une marche aléatoire avec dérive comprend une constante (ou terme de dérive) α et un terme de bruit blanc ε_t . Elle ne revient pas non plus à une moyenne à long terme et a une variance qui dépend du temps. Il est représenté par l'équation 6 comme suit :

$$Y_t = \alpha + Y_{(t-1)} + \varepsilon_t \quad (6)$$

où,

Y_t est la valeur de la série au moment (t)

$Y_{(t-1)}$ est la valeur précédente de la série

α est le terme de dérive, qui représente la **tendance constante** du processus

ε_t est un terme de bruit blanc (choc aléatoire) de moyenne zéro et de variance constante.

Caractéristiques d'une marche aléatoire avec dérive

1. **Présence d'une tendance** : Le terme de dérive (α) fait que la série a une tendance à la hausse ($\alpha > 0$) ou à la baisse ($\alpha < 0$) au fil du temps, plutôt que de fluctuer de manière purement aléatoire.
2. **Non-stationnarité** : La variance augmente avec le temps, rendant la série non stationnaire, ce qui a un impact sur les prévisions et l'inférence statistique.
3. **Accumulation de chocs aléatoires** : Comme chaque valeur dépend de la précédente plus des chocs aléatoires, de petites variations peuvent s'accumuler au fil du temps, entraînant des écarts importants par rapport à la valeur initiale.

Applications

1. Utilisé en **finance, macroéconomie et sciences naturelles**, souvent pour modéliser les cours des actions, les taux de change ou les indicateurs économiques.

4.3.3. Un modèle de tendance déterministe

Un modèle de tendance déterministe est un type de modèle de série temporelle dans lequel la tendance systématique des données est entièrement déterminée par une forme fonctionnelle spécifique (par exemple, linéaire ou exponentielle) sans qu'aucune composante aléatoire n'affecte la tendance. Il suppose que la tendance est prévisible et suit une structure fixe dans le temps. Les modèles de tendance déterministes sont utiles pour les prévisions lorsque la tendance sous-jacente est stable et régie par des facteurs prévisibles.

Le modèle peut être représenté comme suit :

$$Y_t = \alpha + \beta_t + \varepsilon_t \quad (7)$$

où,

Y_t est la valeur de la série au moment (t)

α est le terme de dérive, qui représente la **tendance constante** du processus

β_t est le terme de **tendance déterministe**, qui croît (ou décroît) linéairement avec le temps

ε_t est un terme de bruit blanc (choc aléatoire) de moyenne zéro et de variance constante.

Caractéristiques principales

1. **Tendance prévisible** : La tendance est entièrement spécifiée par la forme fonctionnelle choisie, ce qui la rend hautement **prévisible** dans le temps.
2. **Pas de fluctuations aléatoires dans la tendance** : Tout élément aléatoire dans les données est isolé dans (ε_t) et n'affecte pas la trajectoire systématique.
3. **Stationnarité des résidus** : Le terme d'erreur (ε_t) est généralement stationnaire, ce qui signifie que sa moyenne et sa variance restent constantes.

Applications

- **Modèles de croissance économique** : La croissance du PIB à long terme peut suivre une tendance déterministe, influencée par des facteurs constants tels que la technologie ou l'investissement.
- **Études démographiques** : Les tendances démographiques présentent souvent des schémas déterministes au fil du temps.
- **Science de l'environnement** : Les variables climatiques telles que la température peuvent présenter des tendances déterministes influencées par des processus physiques fixes.

4.3.4. Marche aléatoire avec dérive et tendance déterministe

Une **marche aléatoire avec une tendance déterministe** est un modèle de série temporelle dans lequel la valeur d'une variable évolue en fonction d'**une tendance systématique** et de **fluctuations aléatoires**. Elle s'exprime mathématiquement comme suit :

$$Y_t = \alpha + Y_{t-1} + \beta_t + \varepsilon_t \quad (8)$$

où,

Y_t est la valeur de la série au moment (t)

Y_{t-1}) est la valeur précédente de la série

α est le terme de dérive, qui représente la **tendance constante** du processus

β_t est le terme de **tendance déterministe**, qui croît (ou décroît) linéairement avec le temps

ε_t est un terme de bruit blanc (choc aléatoire) de moyenne zéro et de variance constante.

Caractéristiques principales

1. **Tendance linéaire systématique** : Le terme (β_t) introduit une tendance déterministe qui augmente ou diminue à un taux fixe
2. **Dérive constante** : Le terme (α) ajoute un biais constant à la hausse ou à la baisse au processus, distinct de la tendance temporelle.
3. **Fluctuations aléatoires** : La composante de bruit blanc (ε_t) garantit que la série incorpore toujours une variabilité stochastique, permettant un caractère aléatoire autour du modèle systématique.
4. **Non-stationnarité** : Comme d'autres modèles de marche aléatoire, cette série est non stationnaire parce que sa variance augmente avec le temps et qu'elle n'a pas de moyenne ou de variance fixe.

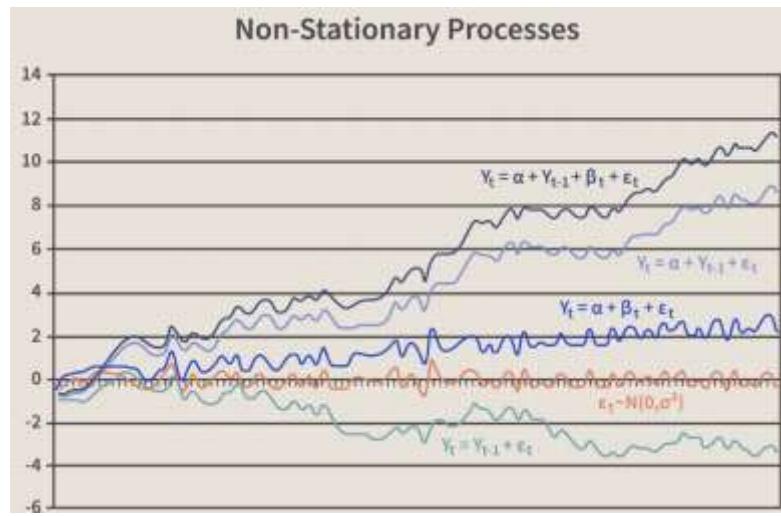
Applications :

Ce modèle est souvent utilisé pour décrire des processus qui suivent une tendance générale tout en présentant des chocs aléatoires :

- **Indicateurs macroéconomiques** (par exemple, PIB avec croissance et fluctuations aléatoires),
- **Prix des actions** ayant une tendance à la croissance à long terme,
- **Modèles de croissance de la population** dans lesquels des événements aléatoires influencent une trajectoire déterministe sous-jacente.

La figure 8 illustre les différents modèles non stationnaires examinés.

Figure 8. Représentation des processus non stationnaires



4.4. Modèles d'autorégression

l'exception du modèle de tendance déterministe, les **équations 5 à 8** sont des versions différentes d'un modèle autorégressif d'ordre 1, également connu sous le nom de **modèle AR(1)**. En général, un **processus autorégressif (AR)** dans les séries temporelles fait référence à un modèle dans lequel la valeur actuelle d'une variable est exprimée comme une combinaison linéaire de ses **valeurs passées** et d'un terme d'erreur aléatoire. Il est couramment utilisé pour modéliser les données de séries temporelles lorsque les observations passées sont supposées influencer les observations futures. L'équation 9 présente une expression générale d'un processus AR d'ordre p :

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \phi_3 Y_{t-3} + \dots + \phi_p Y_{t-p} + \varepsilon_t \quad (9)$$

où,

Y_t est la valeur de la série au moment (t)

$Y_{t-1} \dots Y_{t-p}$ sont les valeurs retardées de la série.

$\phi_1 \dots \phi_p$ sont des coefficients qui mesurent l'influence des valeurs retardées

p représente l'ordre du processus AR, indiquant le nombre de périodes passées utilisées

ε_t terme d'erreur de bruit blanc de moyenne nulle et de variance constante

4.4.1 Estimation des modèles autorégressifs (AR)

4.4.1.1 Importance de la stationnarité

Les séries temporelles peuvent être soit stationnaires, soit non stationnaires. La **stationnarité** des séries temporelles fait référence à une propriété selon laquelle les caractéristiques statistiques de

la série, telles que sa **moyenne**, sa **variance** et sa **structure d'autocorrélation**, restent constantes dans le temps. Une série temporelle stationnaire ne présente pas de tendances à long terme ni de variabilité changeante, ce qui la rend plus facile à analyser et à modéliser.

Pour estimer un modèle autorégressif (AR), il est nécessaire de confirmer que les données de la série temporelle sont stationnaires. **La stationnarité est essentielle** dans l'estimation des séries temporelles car elle garantit une analyse statistique **fiable**, **cohérente** et **significative**. Les points suivants illustrent l'importance de la stationnarité :

1. Propriétés statistiques cohérentes

- Une **série temporelle stationnaire** a une **moyenne**, une **variance** et une **autocorrélation constantes** dans le temps. Cette constance permet aux modèles de saisir les tendances **sans distorsion**.

2. Prévisibilité et précision du modèle

- De nombreux modèles de prévision supposent la **stationnarité** pour que les prévisions futures soient **stables et précises**. Si une série **n'est pas stationnaire**, les prévisions peuvent ne pas être fiables en raison de l'évolution des tendances ou des fluctuations.

3. Validité des hypothèses du modèle

- La plupart des modèles de séries temporelles supposent **un comportement stationnaire** pour une inférence valide. Si une série **n'est pas stationnaire**, les résultats de la régression peuvent être **erronés** et conduire à des conclusions trompeuses.

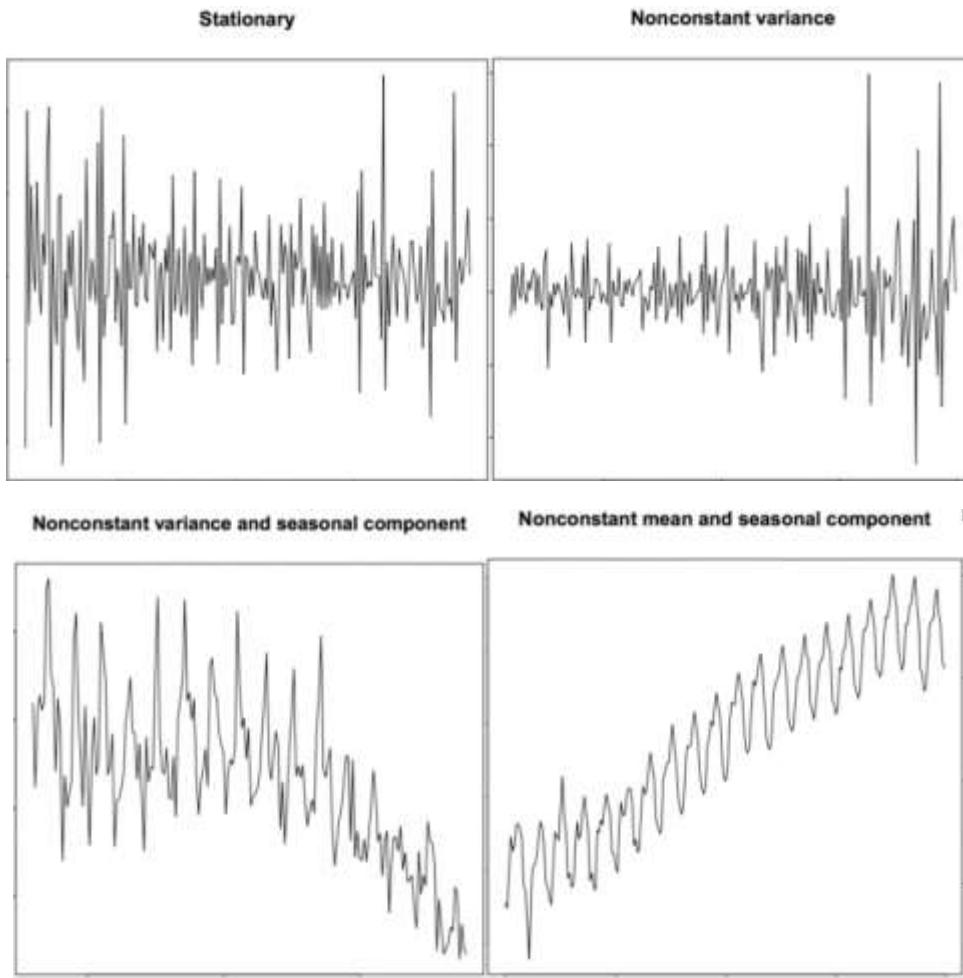
4. Une interprétation plus facile

- Les séries stationnaires sont **plus faciles à analyser** car les relations entre les variables restent **cohérentes**.

4.4.1.2 Détection visuelle de la stationnarité

La **figure 9** une représentation graphique d'une série stationnaire par rapport à une série non stationnaire. On remarque que les séries stationnaires ne présentent pas de tendance visible ni de variabilité changeante. Toutes les autres séries présentent un comportement non constant, que ce soit au niveau de la moyenne ou de la variance, avec ou sans tendance saisonnière, et ne peuvent être estimées sans transformation.

Figure 9. Représentation des séries stationnaires et non stationnaires



4.4.1.3. Tests de stationnarité (ou tests de racine unitaire)

Outre la détection visuelle, il existe plusieurs méthodes pour tester la stationnarité (c'est-à-dire pour tester les racines unitaires dans les séries temporelles), telles que les tests de Phillips-Perron, KPSS ou ADF-GLS. Une méthode couramment utilisée pour tester les racines unitaires est le test de Dickey-Fuller (DF)². Cette méthode est expliquée ci-dessous :

Considérons un modèle **AR(1)** simple de la forme :

$$y_t = \rho y_{t-1} + \varepsilon_t. \quad (10)$$

En soustrayant y_{t-1} des deux côtés, on obtient :

$$y_t - y_{t-1} = \rho y_{t-1} + \varepsilon_t - y_{t-1} \quad (11)$$

En appliquant un opérateur de différence, l'équation (11) peut être ré-exprimée comme suit :

² Comme indiqué, il existe d'autres méthodes de test des racines unitaires, telles que les tests Philips-Perron, KPSS ou ADF-GLS, sur lesquelles je vous encourage à vous renseigner.

$$\Delta y_t = (\rho - 1)y_{t-1} + \varepsilon_t = \delta y_{t-1} + \varepsilon_t, \text{ où } \delta = (\rho - 1) \quad (12)$$

La méthode DF établit le test d'hypothèse suivant :

Hypothèse nulle : $H_0 : \theta = 0$ (équivalent à $\rho = 1$)

Hypothèse alternative : $H_a : \theta < 0$ (équivalent à $\rho < 1$)

Si H_0 ne peut être rejeté, la série a une racine unitaire (c'est-à-dire qu'il s'agit d'une marche aléatoire), elle est donc non stationnaire et ne peut être utilisée pour l'analyse sans transformation.

Si H_0 est rejeté, la série ne présente pas de racine unitaire et peut être utilisée pour une analyse sans transformation puisqu'elle est stationnaire.

4.4.1.4 Transformer des données non stationnaires

Si la série s'avère non stationnaire, des transformations telles que la **différenciation**, le **logarithme** ou la **détrition** peuvent être appliquées pour la rendre stationnaire. Nous nous concentrerons sur la première méthode dans ce cours.

Transformation des données par la différenciation

La différenciation consiste à transformer des séries non stationnaires en séries stationnaires en supprimant les tendances temporelles et la saisonnalité d'une donnée en calculant les différences entre les observations consécutives. Il peut être appliqué aux données plusieurs fois jusqu'à ce que la stationnarité soit obtenue. En général, s'il y a d nombre de racines unitaires dans la série de données, alors la série devra être différenciée d fois pour devenir stationnaire. Ainsi, un processus AR(1) doit être différencié une fois pour être stationnaire, mais un AR(2) doit être différencié deux fois, et un AR(3) doit être différencié trois fois, etc... dans l'ordre pour devenir stationnaire. La différenciation permet de supprimer la dépendance temporelle des données (ou leur dépendance par rapport au temps) et de stabiliser leur moyenne et leur variance.

En utilisant l'exemple d'un modèle de marche aléatoire dans l'équation (5), et en supposant que Y_t^* est la série différenciée, alors :

$$Y_t^* = Y_t - Y_{t-1} = \varepsilon_t \quad (13)$$

Si la différenciation de premier ordre ne produit pas de série stationnaire, il peut être nécessaire de différencier à nouveau les données pour produire une série différenciée de second ordre sous la forme :

$$\begin{aligned} Y_{t-2}^{**} &= Y_t^* - Y_{t-1}^* \\ &= (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) \\ &= Y_t - 2Y_{t-1} + Y_{t-2} \end{aligned} \quad (14)$$

Cette procédure peut être poursuivie jusqu'à ce que la série devienne stationnaire.

4.5. Modélisation et prévision des séries temporelles

Une fois la stationnarité établie, l'analyse porte sur la modélisation de la série. Les modèles ARIMA sont largement utilisés dans les séries temporelles, ARIMA signifiant **Autoregression Integrated Moving Averages** models (modèles de **moyennes** mobiles intégrées à l'**autorégression**). Cette classe de modèles peut être décomposée en 3 composantes :

4.5.1. Modèles autorégressifs (AR)

Les modèles autorégressifs ou AR sont ceux dans lesquels la valeur actuelle de la série temporelle peut être obtenue en utilisant les valeurs précédentes de la même série temporelle → la valeur actuelle est une moyenne pondérée de ses valeurs passées

La forme générale d'un modèle AR(p) est représentée comme suit :

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \dots + \phi_p y_{t-p} + \varepsilon_t \quad (16)$$

où toutes les composantes sont définies comme dans l'équation (9). Il convient toutefois de noter que l'équation (16) est une forme plus générale de l'équation (9) en raison de la présence d'une droite de tendance constante α .

Le nombre de retards significatifs de la variable qui sont conservés dans le modèle est appelé "ordre" du modèle AR. Par conséquent, un modèle AR(1) est un modèle autorégressif d'ordre 1 ; un modèle AR(2) est un modèle autorégressif d'ordre 2, etc...

Comme nous l'avons vu, un modèle AR(p) est généralement connu comme un modèle autorégressif d'ordre p, où p représente le nombre de retards inclus dans le modèle.

4.5.2. Modèles de moyenne mobile (MA)

Un modèle de moyenne mobile ou MA est un modèle dans lequel la valeur actuelle de la série est définie comme une combinaison linéaire des erreurs passées. Il s'agit d'une méthode de régression utilisée pour lisser les fluctuations et mettre en évidence les tendances. En supposant que les erreurs sont distribuées de manière indépendante selon la loi normale, un modèle MA peut s'écrire comme suit :

$$y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \dots + \theta_q \varepsilon_{t-q} \quad (17)$$

Le nombre de retards inclus dans le modèle MA est considéré comme l'"ordre" du modèle. Par conséquent, une MA (1) est appelée moyenne mobile d'ordre 1 ; une MA (2) est appelée moyenne mobile d'ordre 2, ...

En général, un **modèle MA (q)** est connu comme un modèle de moyenne mobile d'ordre q.

4.5.3. Série de données intégrées

Une série temporelle intégrée est un type de série temporelle qui présente une **non-stationnarité** mais qui peut être transformée en une série **stationnaire** en la différenciant un certain nombre de fois. Le nombre de différences nécessaires pour atteindre la stationnarité est appelé **l'ordre d'intégration**. Les séries temporelles intégrées contiennent souvent une **racine unitaire**, ce qui signifie qu'elles ont une **tendance persistante** qu'il faut différencier pour l'éliminer. Le nombre de différences nécessaires pour atteindre la stationnarité est appelé **l'ordre d'intégration**.

En général, une série est dite intégrée d'ordre d , ou $I(d)$, si nécessite un nombre d d'écart d'ordre d pour devenir stationnaire, où $d \geq 0$.

Plus précisément, nous avons :

- Un $I(0)$ indique une série stationnaire sans racines unitaires ni tendance.
- Un $I(1)$ est une série avec une racine unitaire qui doit être différenciée une fois pour devenir stationnaire.
- Un $I(2)$ est une série avec deux racines unitaires qui doit être différenciée deux fois pour devenir stationnaire.
- ...

Récapitulation

En combinant les équations (16) et (17) et en tenant compte de l'ordre d'intégration, un modèle de moyenne mobile autorégressive ARMA(p, q) peut généralement être exprimé comme suit :

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \phi_3 y_{t-3} + \dots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \theta_3 \varepsilon_{t-3} + \dots + \theta_q \varepsilon_{t-q} \quad (18)$$

Si, en outre, y_t est intégré d'ordre d , l'équation (18) devient un modèle ARIMA(p, d, q), où $p, d, q \geq 0$.

Configurations possibles

$$\text{➤ Si } y_t \text{ est un ARMA (1,0), alors } y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t \quad (19)$$

$$\text{➤ Si } y_t \text{ est un ARMA (0,1), alors } y_t = \alpha + \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad (20)$$

$$\text{➤ Si } y_t \text{ est un ARMA (1,1), alors } y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad (21)$$

Si, en outre, y_t est intégré d'ordre 1, l'équation (21) prend la forme d'un modèle ARIMA(1,1,1).

4.6. Étude de cas : PIB réel du Nigeria (RGDP) sur la période 1970-2017

4.6.1. Visualisation et test des racines unitaires

Cette section présente une étude de cas comportement du PIB réel du Nigeria (RGDP) sur la période 1970-2017. La **figure 10** représente le logarithme du PIB en niveaux et en différence première. La visualisation des données suggère que la série semble être intégrée en niveau, mais qu'elle semble stationnaire en différence, bien qu'avec peut-être une tendance.

Une méthode de Dickey-Fuller augmentée (ADF) a été appliquée pour tester la présence de racines unitaires dans la série. Les résultats sont présentés dans le **tableau 5**. Lorsque la méthode est appliquée à la série RGDP sous forme de niveau, la *valeur p= 0,98* sans tendance incluse et 0,80 avec tendance, ce qui indique que l'hypothèse nulle d'une racine unitaire ne peut être rejetée.

Le **tableau 6** montre les résultats d'un test de racine unitaire sur la première différence du logarithme du RGDP pour le Nigeria en utilisant la méthode ADF. Ici, contrairement aux résultats obtenus pour le cas du logarithme du PIB en niveaux, la *valeur p* est de **0,00039** lorsqu'une tendance n'est pas incluse, et de **0,00197** lorsqu'une tendance est incluse. Ces résultats indiquent que l'hypothèse nulle de racine unitaire est rejetée, ce qui implique que les séries sont stationnaires en différence première, avec une tendance et peuvent donc être utilisées pour l'analyse de régression.

Figure 10. Graphique du logarithme du PIB et de sa première différence pour le Nigeria

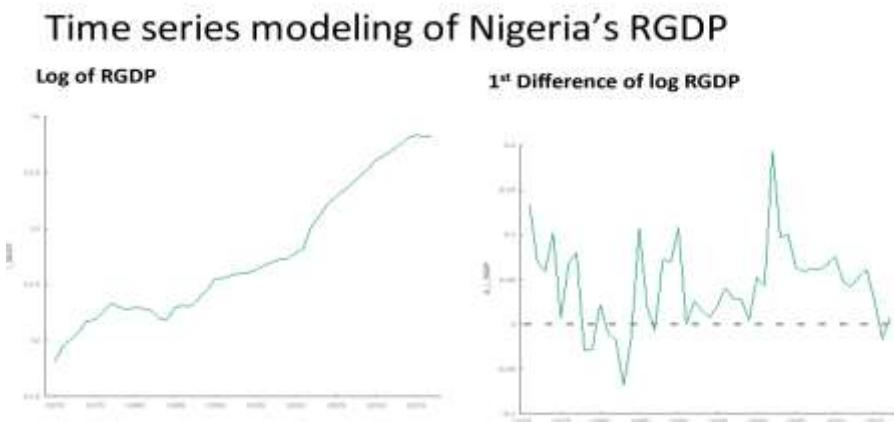
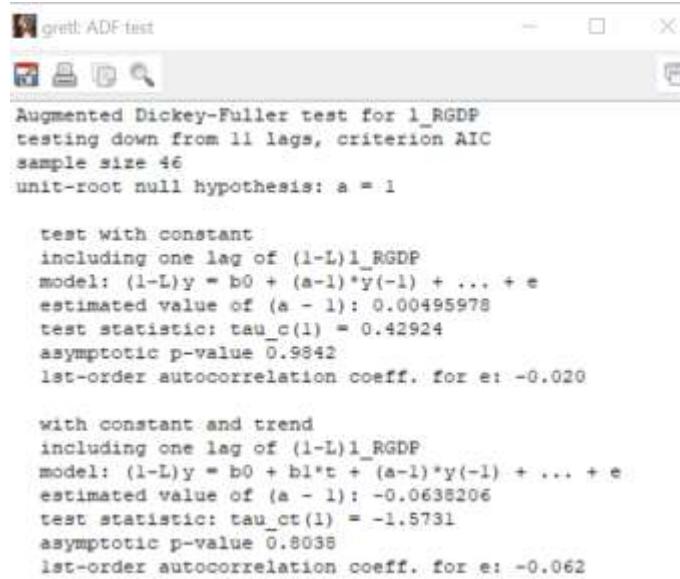


Tableau 5. Test de racine unitaire sur le logarithme du RGDP (en niveaux) pour le Nigeria³

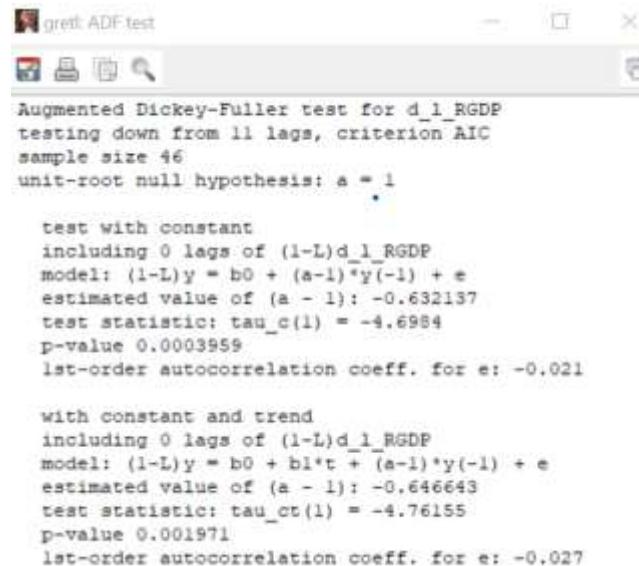
³ Les estimations et les tests d'hypothèses ont été effectués à l'aide de Gretl, un logiciel statistique libre. D'autres progiciels de ce type incluent eViews, RATS, SPSS, etc...


 gretl: ADF-test
 Augmented Dickey-Fuller test for l_RGDP
 testing down from 11 lags, criterion AIC
 sample size 46
 unit-root null hypothesis: $\alpha = 1$

 test with constant
 including one lag of $(1-L)l_RGDP$
 model: $(1-L)y = b_0 + (\alpha-1)*y(-1) + \dots + e$
 estimated value of $(\alpha - 1)$: 0.00495978
 test statistic: $\tau_{c}(1) = 0.42924$
 asymptotic p-value 0.9842
 1st-order autocorrelation coeff. for e : -0.020

 with constant and trend
 including one lag of $(1-L)l_RGDP$
 model: $(1-L)y = b_0 + b_1*t + (\alpha-1)*y(-1) + \dots + e$
 estimated value of $(\alpha - 1)$: -0.0638206
 test statistic: $\tau_{ct}(1) = -1.5731$
 asymptotic p-value 0.8038
 1st-order autocorrelation coeff. for e : -0.062

Tableau 6. Test de racine unitaire sur la première différence du logarithme du RGDP


 gretl: ADF-test
 Augmented Dickey-Fuller test for d_l_RGDP
 testing down from 11 lags, criterion AIC
 sample size 46
 unit-root null hypothesis: $\alpha = 1$

 test with constant
 including 0 lags of $(1-L)d_l_RGDP$
 model: $(1-L)y = b_0 + (\alpha-1)*y(-1) + e$
 estimated value of $(\alpha - 1)$: -0.632137
 test statistic: $\tau_{c}(1) = -4.6984$
 p-value 0.0003959
 1st-order autocorrelation coeff. for e : -0.021

 with constant and trend
 including 0 lags of $(1-L)d_l_RGDP$
 model: $(1-L)y = b_0 + b_1*t + (\alpha-1)*y(-1) + e$
 estimated value of $(\alpha - 1)$: -0.646643
 test statistic: $\tau_{ct}(1) = -4.76155$
 p-value 0.001971
 1st-order autocorrelation coeff. for e : -0.027

4.6.2. Modélisation du (log) RGDP du Nigeria en première différence comme ARMA(1,1)

Ayant établi que la série RGDP est stationnaire en différence première, son comportement peut alors être étudié. Le tableau 7 montre les résultats de l'étude du RGDP en différence première comme ARMA(1,1) de la forme $y_t = \alpha + \phi_1 y_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1}$ comme dans l'équation (21).

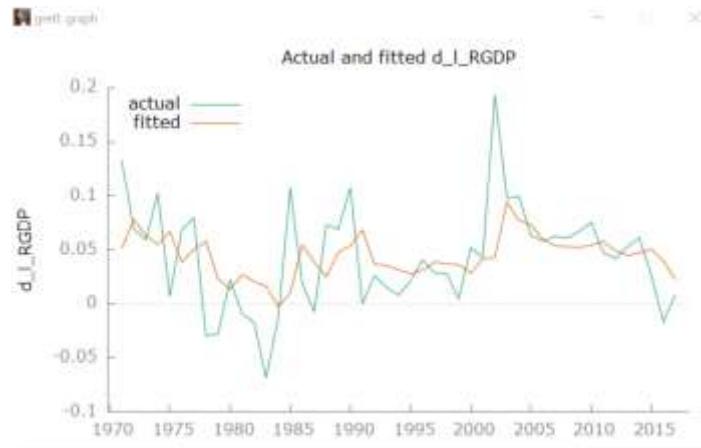
Le coefficient **AR** estimé $\hat{\phi}_1 = 0,646$, avec une statistique z= 2,20, et une valeur p = 0,0275, ce qui est statistiquement significatif au niveau de 5 %. Toutefois, le coefficient **MA** $\hat{\theta}_1 = -0,30$, avec une statistique z de -0,85 et une valeur p de 0,3952, n'est pas statistiquement significatif. Les résultats globaux indiquent que sur la période 1970-2017⁴ , le RGDP du Nigeria se comporte comme un modèle **ARMA(1,0)** ou simplement **AR(1)**. Le modèle estimé est représenté graphiquement par rapport aux données réelles dans la **figure 11**.

Tableau 7. Test du RGDP du Nigeria en tant que modèle ARMA

| grmt model 22 | | | | | |
|---|------------|---------------------|-----------|--------|-----|
| File Edit Tests Save Graphs Analysis LaTeX | | | | | |
| Function evaluations: 39 | | | | | |
| Evaluations of gradient: 13 | | | | | |
| Model 22: ARMA, using observations 1971-2017 (T = 47) | | | | | |
| Estimated using AS 197 (exact ML) | | | | | |
| Dependent variable: d_1_RGDP | | | | | |
| Standard errors based on Hessian | | | | | |
| coefficient | std. error | z | p-value | | |
| const | 0.0436934 | 0.0120928 | 3.613 | 0.0003 | *** |
| phi_1 | 0.646624 | 0.293310 | 2.205 | 0.0275 | ** |
| theta_1 | -0.309718 | 0.364257 | -0.8503 | 0.3952 | |
| Mean dependent var | 0.042778 | S.D. dependent var | 0.047885 | | |
| Mean of innovations | -0.001341 | S.D. of innovations | 0.043606 | | |
| R-squared | 0.153591 | Adjusted R-squared | 0.134782 | | |
| Log-likelihood | 80.44280 | Akaike criterion | -152.8856 | | |
| Schwarz criterion | -145.4850 | Hannan-Quinn | -150.1007 | | |

Figure 11. Graphique du PIB différentiel réel par rapport au PIB différentiel ajusté au 1er trimestre

⁴ Les résultats estimés commencent en 1971 car une année a été utilisée pour calculer le décalage du PIBR.



5. Modèles vectoriels autorégressifs (VAR)

5.1. Contexte

La méthodologie de l'autorégression vectorielle (VAR), introduite en 1980 par le professeur Christopher Sims, est une extension de l'analyse des séries temporelles. Elle fournit un cadre pour l'analyse des données économiques et financières en décrivant les données, les prévisions, l'inférence structurelle et l'analyse politique.

Les modèles VAR sont utilisés dans l'analyse économique pour étudier les relations dynamiques entre plusieurs variables temporelles, telles que le PIB, l'inflation, les taux d'intérêt et les taux de change. Les modèles VAR sont largement utilisés pour prédire les valeurs futures des indicateurs économiques sur la base des données passées, et peuvent aider les décideurs politiques à évaluer l'impact des politiques monétaires et fiscales sur l'économie.

5.2 Description du modèle

Stock et Watson (2001) décrivent un VAR comme "**un modèle linéaire à n équations et n variables dans lequel chaque variable est à son tour expliquée par ses propres valeurs retardées, plus les valeurs actuelles et passées des $n-1$ variables restantes**". L'hypothèse clé de ce type de modèle est que les variables incluses peuvent s'influencer mutuellement dans le temps.

Le nombre de retards des variables endogènes incluses dans le modèle est connu comme **l'ordre du VAR**. Le nombre optimal de retards doit être déterminé par des tests statistiques tels que l'**AIC**, etc... et doit être le même pour toutes les variables endogènes. Ainsi, un $\text{VAR}(p)$ comprend p retards des variables endogènes.

Prenons l'exemple d'un VAR à 2 équations et 2 variables ($y_{1,t}, y_{2,t}$). Soit $p=1$ indiquant que cet exemple ne comprendra qu'un seul retard. Le modèle peut alors être exprimé sous forme matricielle comme suit :

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix} \quad (22)$$

Le modèle (22) peut être articulé de manière équivalente sous la forme d'un système de deux équations comme suit :

$$\begin{aligned} y_{1,t} &= c_1 + a_{1,1}y_{1,t-1} + a_{1,2}y_{2,t-1} + \varepsilon_{1,t} \\ y_{2,t} &= c_2 + a_{2,1}y_{1,t-1} + a_{2,2}y_{2,t-1} + \varepsilon_{2,t} \end{aligned} \quad (23)$$

Dans les deux cas, l'observation au temps t de chaque variable dépend de **ses propres valeurs retardées** ainsi que des **valeurs retardées de l'autre variable** du VAR.

5.3 Estimation d'un modèle VAR

Supposons que, dans le cadre d'une étude de cas, vous soyez chargé d'étudier les comportements des séries **RGDP** et **Consommation** (sous forme logarithmique), et la manière dont elles peuvent s'influencer mutuellement pour le cas du Nigéria sur la période 1970-2017. Dans cet exemple, $y_{1,t} = RGDP$ et $y_{2,t} = Consommation$. Après avoir testé et résolu les problèmes potentiels de racines unitaires dans les deux séries, l'estimation VAR a été effectuée.

Les tableaux 8 et 9 présentent les résultats des estimations VAR. Les deux variables endogènes sont le *RGDP en logarithme différentiel* et la *consommation en logarithme différentiel*

Tableau 8. Résultats VAR pour l'équation RGDP

Système VAR, ordre de retard 1
Estimations MCO, observations 1972-2017 ($T = 46$)

Equation 1 : RGDP

| | Coefficient | Erreur std. | Rapport t | Valeur p | |
|----------------------------|------------------|-----------------------|--------------|---------------|-----|
| Const. | 0.0242042 | 0.00865786 | 2.796 | 0.0077 | *** |
| ld_RGDP_1 | 0.325822 | 0.138712 | 2.349 | 0.0235 | ** |
| Id_Cons_1 | 0.0601755 | 0.0514098 | 1.171 | 0.2482 | |
| Moyenne var dépendante | 0.040814 | S.D. var. dépendante | 0.046462 | | |
| Somme des résidus au carré | 0.080469 | S.E. de la régression | 0.043259 | | |
| R au carré | 0.171622 | R-carré ajusté | 0.133092 | | |
| F(2, 43) | 4.454323 | Valeur P(F) | 0.017455 | | |
| Rho | -0.020566 | Durbin-Watson | 2.040129 | | |

Tableau 9. Résultats du VAR pour l'équation de consommation

Equation 2 : Consommation

| | <i>Coefficient</i> | <i>Erreur std.</i> | <i>Rapport t</i> | <i>Valeur p</i> | |
|----------------------------|--------------------|-----------------------|------------------|-----------------|-----|
| Const. | 0.0100136 | 0.0231700 | 0.4322 | 0.6678 | |
| ld_RGDP_1 | 1.01819 | 0.371216 | 2.743 | 0.0088 | *** |
| ld_Cons_1 | -0.414135 | 0.137582 | -3.010 | 0.0044 | *** |
| Moyenne var dépendante | 0.037637 | S.D. var. dépendante | | 0.129369 | |
| Somme des résidus au carré | 0.576311 | S.E. de la régression | | 0.115770 | |
| R au carré | 0.234778 | R-carré ajusté | | 0.199186 | |
| F(2, 43) | 6.596423 | Valeur P(F) | | 0.003173 | |
| Rho | -0.018770 | Durbin-Watson | | 2.036548 | |

Interprétation : Dans le cas du Nigeria, le décalage du RGDP est un déterminant significatif et positif de la consommation. Cependant, l'inverse n'est pas vrai. La consommation décalée ne produit pas d'impact significatif sur le RGDP.

5.4 Analyse des résidus et fonctions de réponse impulsionale

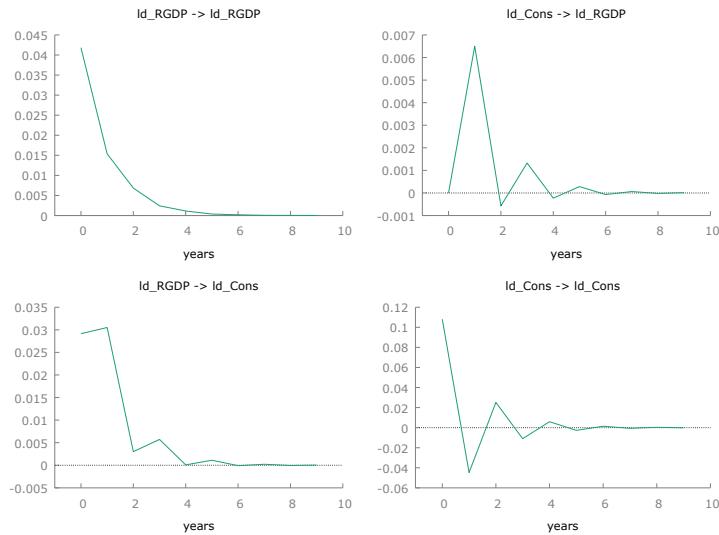
Les résidus de l'estimation du VAR sont représentés dans la **figure 12**. Ils ne montrent pas de variations importantes entre les années 1960 et la fin des années 1980. Toutefois, au-delà de cette période, les variations importantes des résidus appellent des recherches plus approfondies sur la situation économique, y compris les prix du pétrole brut, et la politique du gouvernement au cours de ces années

De même, les fonctions de réponse aux impulsions (IRF) sont représentées à la **figure 13**. Les réponses aux impulsions (IR) sont utilisées pour obtenir une meilleure intuition du **comportement dynamique** du modèle. Elles suivent la réaction d'une variable de réponse à un choc ponctuel dans une variable d'impulsion, indiquant le nombre de périodes nécessaires pour que la réponse initiale revienne à zéro.

Figure 12. Analyse des résidus du VAR



Figure 13. Fonctions de réponse impulsionale



5.5 Performance globale de l'approche VAR

Les modèles VAR sont devenus des outils macroéconométriques efficaces et flexibles pour l'analyse des séries temporelles multivariées, en ce qui concerne les données et les prévisions. Cependant, lorsqu'il s'agit d'inférence structurelle et d'application politique, **l'approche VAR seule n'est pas suffisante**. Dans un modèle VAR, toutes les variables sont dépendantes les unes des autres, de sorte que les valeurs individuelles des paramètres ne fournissent que des informations limitées sur la réaction du système à un choc.

Le "problème d'identification" récurrent en économétrie (c'est-à-dire la causalité par rapport à la corrélation) n'est pas résolu par l'approche VAR, car il nécessite une analyse plus complexe des données, y compris l'utilisation de la théorie économique et d'autres types appropriés de connaissances et d'expertise.

Références

1. **Analyse de régression : Un guide intuitif pour l'utilisation et l'interprétation des modèles linéaires** - Jim Frost (2020) - Publication indépendante
2. **Régression appliquée : Une introduction** - Colin Lewis-Beck (2015) - SAGE Publications
3. **Analyse de régression par l'exemple** - Samprit Chatterjee & Ali S. Hadi (2006) - Wiley
4. **Introduction à l'analyse de régression linéaire** - Douglas C. Montgomery, Elizabeth A. Peck & G. Geoffrey Vining (2012) - Wiley
5. **Modèles linéaires généralisés** - John Fox (2016) - SAGE Publications
6. **Stratégies de modélisation de la régression** - Frank E. Harrell Jr (2015) - Springer
7. **Prévision : Principles and Practice** - Rob J. Hyndman & George Athanasopoulos (2013) - OTexts
8. **Analyse des séries temporelles : Forecasting and Control** - George E. P. Box & Gwilym M. Jenkins (2015) - Wiley
9. **L'analyse des séries temporelles : An Introduction** - Chris Chatfield (2016) - Chapman & Hall/CRC
10. **Analyse des séries temporelles** - James Douglas Hamilton (1994) - Princeton University Press
11. **Analyse appliquée des séries temporelles** - Wayne A. Woodward, Henry L. Gray & Alan C. Elliott (2017) - CRC Press
12. **Analyse multivariée des séries temporelles : Avec R et applications financières** - Ruey S. Tsay (2014) - Wiley
13. **Analyse vectorielle autorégressive structurelle** - Lutz Kilian & Helmut Lütkepohl (2017) - Cambridge University Press
14. **Econométrie appliquée des séries temporelles** - Helmut Lütkepohl & Markus Krätzig (2004) - Cambridge University Press
15. **Vector Autoregressive Models for Multivariate Time Series** - Todd E. Clark & Michael W. McCracken (2020) - Wiley