United Nations
Economic Commission
for Africa

The African Institute for Economic Development and Planning (IDEP)

Modélisation macroéconomique pour le développement durable

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

Professeur Sylvain BOKO

2025

Objectifs du module

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.

Explorer les différents types de régression - Examiner la régression linéaire simple et multiple, les séries chronologiques et le VAR.

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.

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.

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.

Résultats de l'apprentissage

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.

4.1. Introduction à l'analyse de régression

Représentation du modèle

- Définition du modèle de régression linéaire
  - Évalue la relation entre les variables dépendantes et indépendantes
- Analyse de régression univariée
  - Comprend une variable dépendante et une variable indépendante
- Analyse de régression multiple
  - Comprend une variable dépendante et plusieurs variables indépendantes
- Représentation des équations

$$\boldsymbol{Y} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1 X_1} + \boldsymbol{\varepsilon} \qquad (1)$$

  - Y = Variable dépendante
  - X1 = Variable indépendante
  - β0 = Constante
  - β1 = Pente de la droite
  - ε = Résidu ou terme d'erreur (mesure la proportion de la variable dépendante inexpliquée par la variable exogène)

Importance et types de données

Types de données

Données d'observation
- Recueillis à partir de scénarios réels
- Les enquêtes, les dossiers administratifs et les indicateurs économiques en sont des exemples.

Données expérimentales
- Obtenues à partir d'expériences contrôlées
- Les variables sont manipulées pour observer leurs effets

Données de panel
- Suivi des mêmes entités dans le temps
- Fournit des informations sur les changements et les tendances

Données transversales
- Instantané de différentes entités à un moment donné

Données de séries temporelles

Importance des données dans l'analyse de régression

- Les données constituent l'épine dorsale de l'analyse de régression, permettant 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 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.
- 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.

Tableau 1 : Dépenses mensuelles du gouvernement pour le programme de formation professionnelle

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

- *Imaginons que le gouvernement mette en œuvre un programme de formation professionnelle dans le cadre d'une politique de relance.*
- *Le tableau 1 présente les dépenses publiques mensuelles consacrées au programme de formation professionnelle et le nombre correspondant d'emplois créés en 2020.*
- *À l'aide de l'équation (1), estimez la relation entre les emplois créés pour les jeunes et les dépenses publiques.*
- *Sur la base des résultats, analysez la force de la relation*

Visualisation des données et des lignes de régression

Importance de la visualisation des données dans l'analyse de régression

- Aide les analystes à explorer les relations entre les variables
- Facilite la compréhension et la communication des données

Diagrammes de dispersion pour l'évaluation des corrélations

- Évaluation visuelle des corrélations entre les variables
- Détermine si les relations sont linéaires ou non linéaires

Mise en évidence des schémas et des tendances

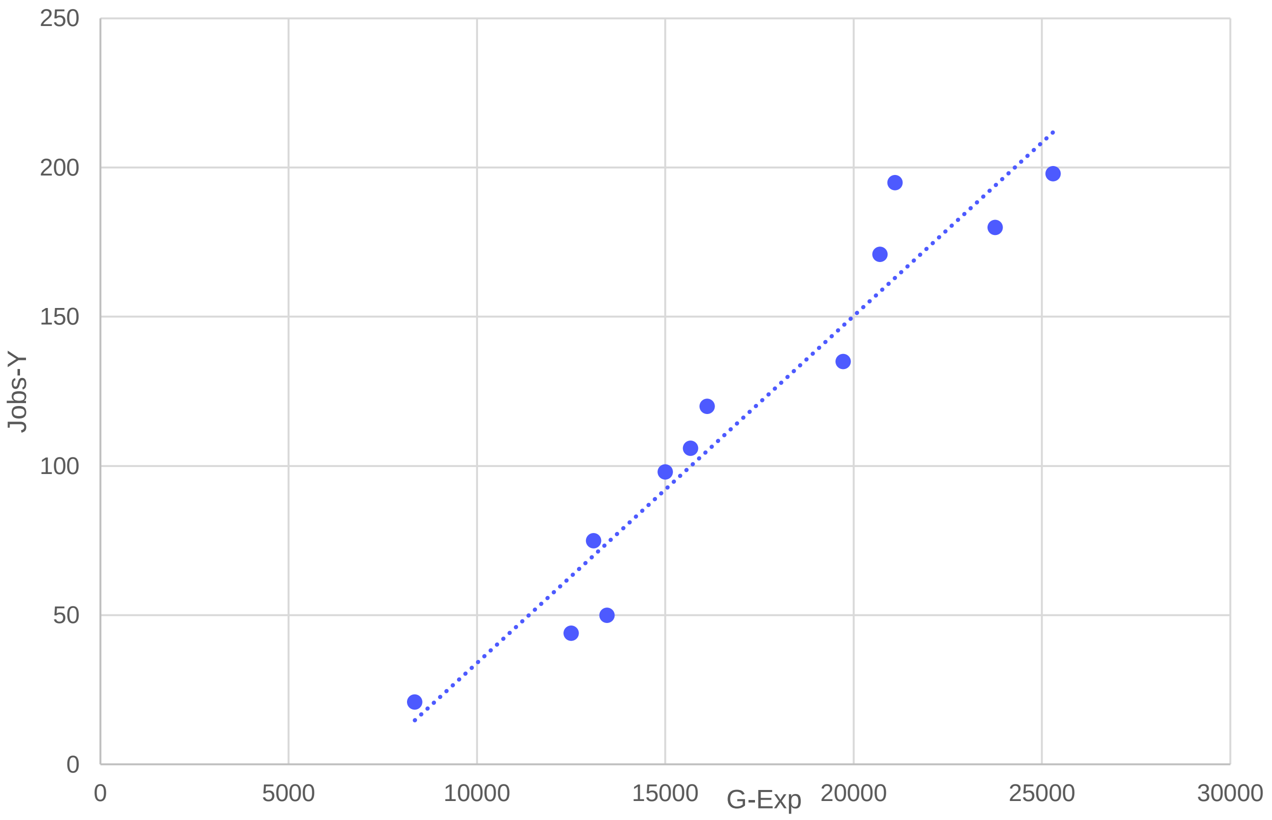
- Révèle des modèles, des groupes et des tendances sous-jacents
- Fournit des informations qui ne sont pas évidentes à partir des données brutes

Figure 1 : Représentation graphique des données

- Axe horizontal : variable explicative (G-Exp)
- Axe vertical : variable dépendante (Jobs-Y)

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

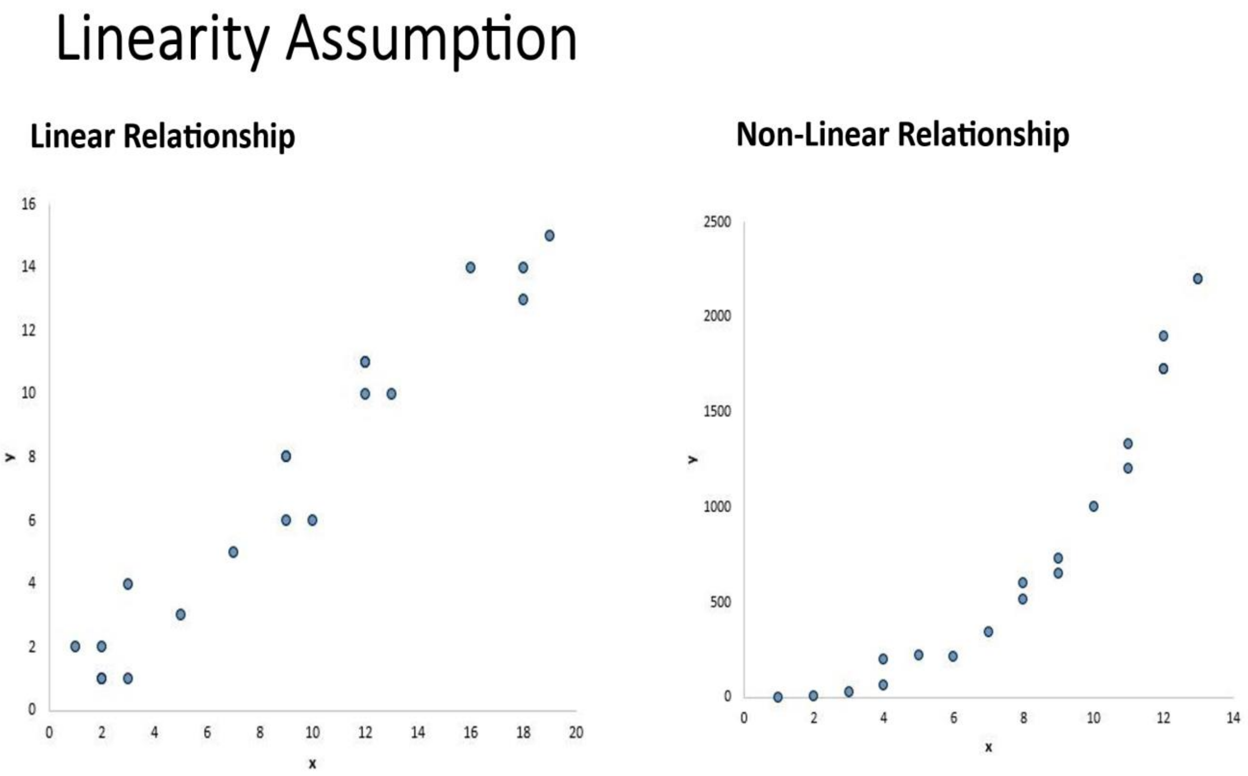
Figure 1. Représentation visuelle de Jobs-Y par rapport à G-Exp



Procédure d'estimation

- **Modèle linéaire simple**
  - L'équation (1) représente un modèle linéaire simple
- **Méthode des moindres carrés ordinaires (MCO)**
  - La méthode des moindres carrés ordinaires (MCO) est une technique d'analyse de régression de base utilisée pour estimer les coefficients d'un modèle linéaire
- **Objectif des MCO**
  - Trouver la droite qui minimise la somme des différences quadratiques entre les valeurs observées et les valeurs prédites.
- **Utilisation courante**
  - Efficace pour identifier les relations entre les variables dépendantes et indépendantes
- **Importance des hypothèses**
  - La régression par les moindres carrés ordinaires (MCO) repose sur une série d'hypothèses pour garantir la validité de ses résultats.

Hypothèses clés



Linéarité

- Le modèle suppose que la relation entre les variables indépendantes (X) et la variable dépendante (Y) est linéaire.
- La violation peut indiquer l'absence de relation ou une relation courbée.
- Utiliser des méthodes de transformation des données si l'on soupçonne une non-linéarité. Exemples : transformation logarithmique, racine carrée ou réciproque.

Hypothèses clés

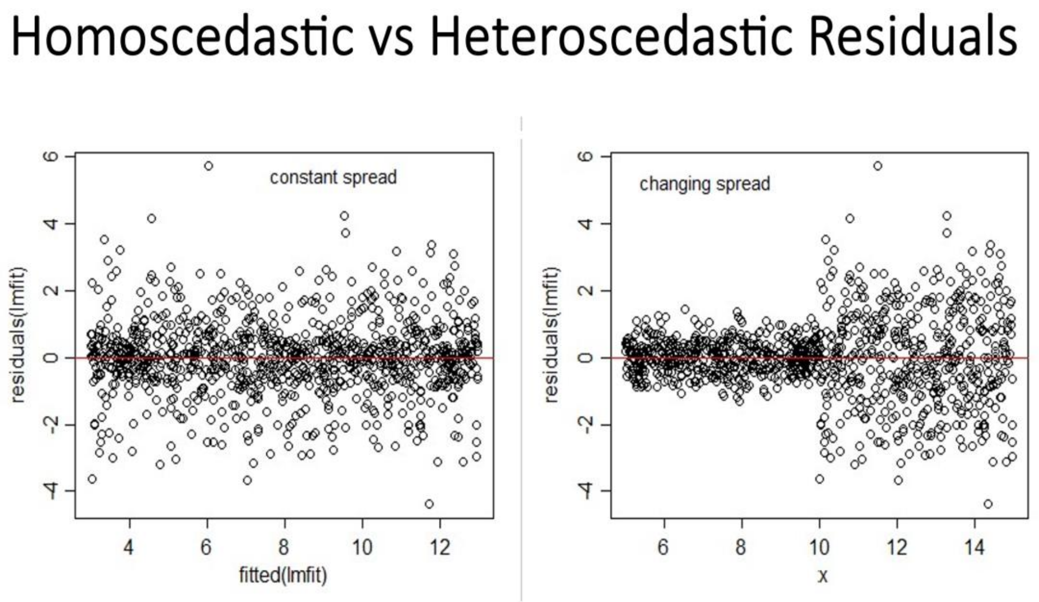
Distribution des termes d'erreur

- **Indépendance des observations**
  - Pas de corrélation entre les résidus (erreurs) des observations
  - Les termes d'erreur sont indépendants les uns des autres
- **Termes d'erreur identiquement distribués**
  - Les termes d'erreur suivent la même distribution de probabilité
  - Moyenne et variance identiques pour tous les termes d'erreur
- **Importance dans les modèles de régression**
  - Garantit des prédictions de modèle non biaisées et fiables
  - Simplifie l'analyse mathématique
  - Facilite l'application de techniques statistiques telles que les tests d'hypothèse

Hypothèses clés

Homoscédasticité

- **Définition de l'homoscédasticité**
  - Hypothèse d'une variance constante des résidus pour tous les niveaux de variables indépendantes
  - Assure des estimations non biaisées et efficaces dans l'analyse de régression
- **Impact de l'homoscédasticité**
  - Prévisions stables pour différentes valeurs des variables indépendantes
  - Les violations conduisent à l'hétéroscédasticité, ce qui entraîne des résultats de régression peu fiables.
  - Conclusions erronées dans les tests d'hypothèses
- **Test d'hétéroscédasticité**
  - Utilisation de la méthode Breusch-Pagan
  - Hypothèse nulle : les termes d'erreur ont une variance constante
  - Hypothèse alternative : les termes d'erreur ont une variance non constante
- Les **solutions à l'hétéroscédasticité** comprennent la transformation ou la redéfinition de la variable dépendante, ou l'utilisation d'une régression pondérée.



Hypothèses clés

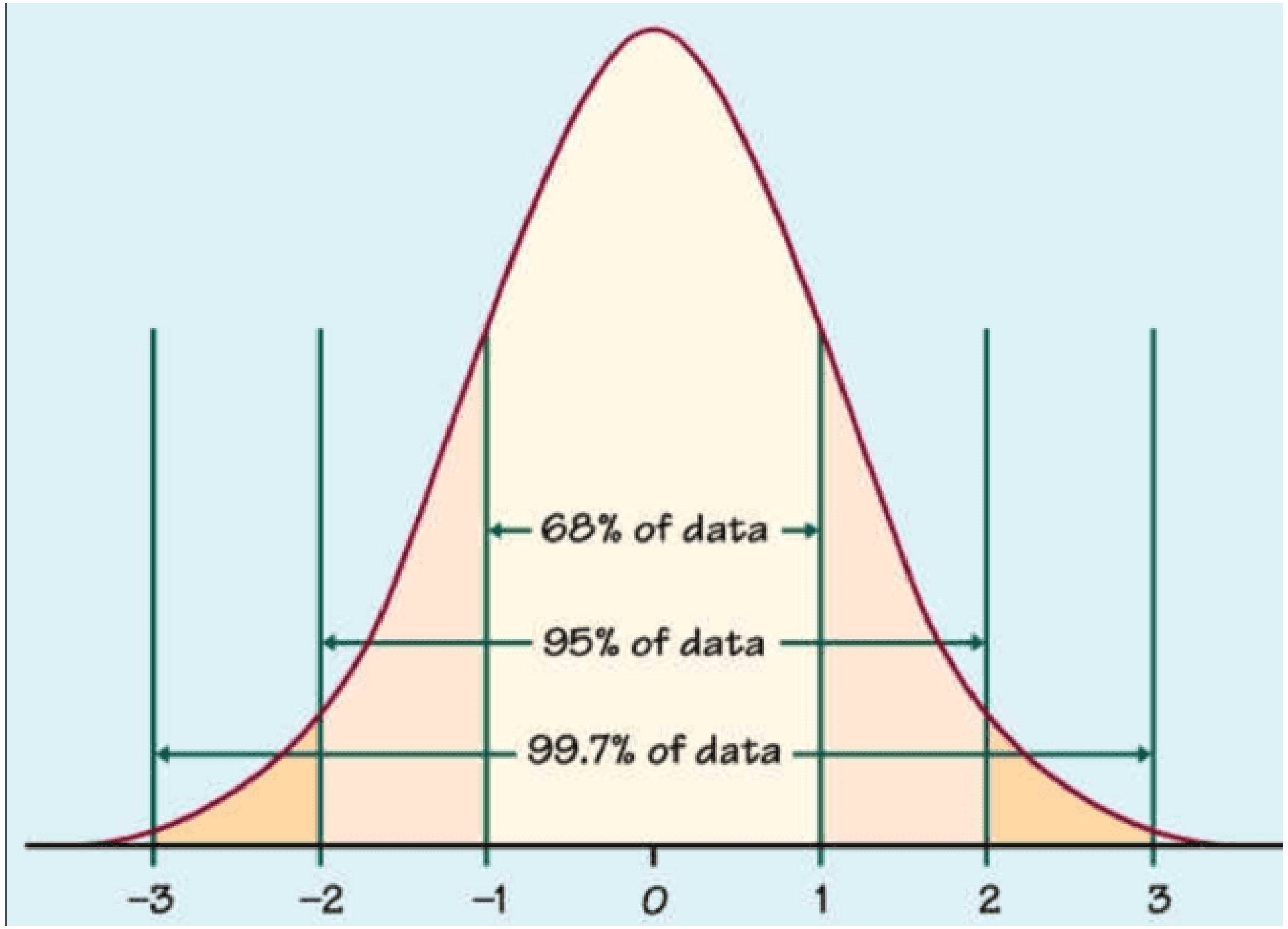
Multicolinéarité

- **Définition de la multicolinéarité**
  - Apparaît dans les modèles de régression multiple
  - Implique une forte corrélation entre les variables indépendantes
- **Impact sur les coefficients**
  - Distorsion de l'estimation des coefficients
  - De petites modifications des données entraînent d'importantes fluctuations
  - Conduit à des coefficients instables
- **Réduction de l'interprétabilité**
  - Difficile de déterminer l'influence des variables individuelles
  - Affaiblit la significativité statistique
- **Inflation des erreurs types**
  - Provoque des erreurs standard gonflées
  - Affaiblir le pouvoir du modèle

Hypothèses clés

Distribution normale

- **Hypothèse de normalité**
  - Concerne la distribution des résidus
  - Les résidus sont les différences entre les valeurs observées et les valeurs prédites.
- **Caractéristiques de la distribution normale**
  - Courbe en forme de cloche
  - Symétrique autour de la moyenne
- **Importance des tests d'hypothèses**
  - Détermine les intervalles de confiance
  - Essentiel pour une vérification précise des hypothèses
- **Règle empirique (règle 68-95-99.7)**
  - 68% des données dans un écart-type
  - 95% des données dans deux écarts-types
  - 99,7 % des données dans trois écarts-types



Résultats de l'estimation

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

Qualité de l'ajustement : Mesures et métriques

R-carré ajusté

- **Ajustement global du modèle**
  - Mesurée par la valeur du R-carré ajusté
  - Explique la variabilité de la variable dépendante
- **Comprendre la force des relations**
  - Le R-carré ajusté indique la force entre les variables dépendantes et indépendantes.
  - Exemple : Un R-carré ajusté de 0,8 indique que 80 % de la variance est expliquée.
- **Comparaison de différents modèles de régression**
  - Une valeur plus élevée du R-carré ajusté indique une meilleure adéquation.
  - Le modèle explique une plus grande partie de la variance de la variable dépendante

Qualité de l'ajustement : Mesures et métriques :

t-statistique

- **Définition de la statistique t**
  - Utilisé pour déterminer la significativité statistique des variables explicatives
  - Détermine si le coefficient de la variable est significativement différent de zéro
- **Formule pour la statistique t**

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

  - S'applique à chaque variable exogène du modèle et à la constante
- **Test d'hypothèse**
  - Pour le test à deux queues, |t-statistique| ≥ 1,96 indique une significativité au seuil de 5 %.
  - |t-statistique| ≥ 2,58 indique une significativité au niveau de 1 %.
- **Niveaux de significativité**
  - Les variables exogènes sélectionnées doivent généralement (mais pas toujours) être celles dont les t-statistiques sont statistiquement significatives.
  - Les valeurs p doivent être inférieures aux seuils de significativité choisis.

Qualité de l'ajustement : Mesures et paramètres

Critère d'information d'Akaike (AIC)

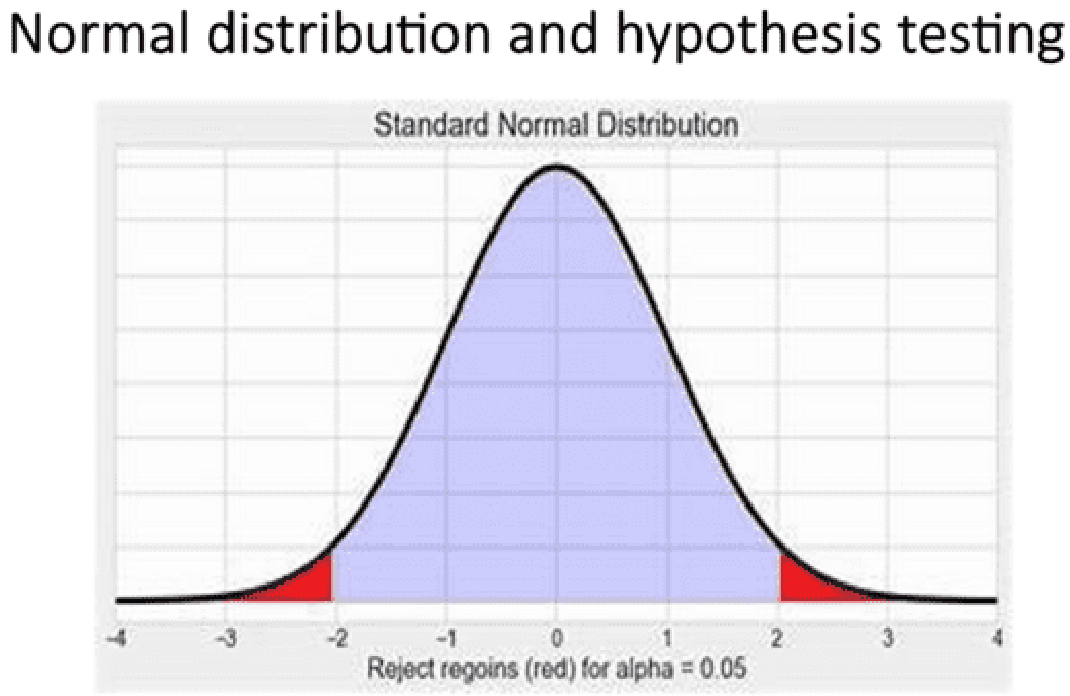
- **Définition de l'AIC**
  - Un outil de sélection de modèles
  - Évalue la performance du modèle
- **Hypothèse de l'AIC**
  - Le modèle le mieux adapté explique la plus grande variation
  - Utilise le moins de variables indépendantes possible (parcimonie)
- **Comparaison des modèles avec l'AIC**
  - Un AIC plus faible indique un modèle mieux adapté

Qualité de l'ajustement : Mesures et métriques

F-Statistique

- **Rôle de la statistique F ou (ANOVA) dans l'analyse de régression**
  - Détermine la significativité statistique du modèle de régression global
- **Valeur F élevée Implications**
  - Les variables indépendantes expliquent de manière significative la variation de la variable dépendante
- **Faible valeur F Implications**
  - Les variables indépendantes peuvent ne pas avoir d'impact significatif
- **Détermination du niveau de significativité**
  - La valeur p associée à la statistique F détermine le niveau de significativité.

Test d'hypothèse



- Mise en place de l'hypothèse
  - Hypothèse nulle ( ) : GExp n'a pas d'impact sur Jobs-Y (**$\boldsymbol{\beta_1}$ = 0)**
  - Hypothèse alternative ( ) : GExp a un impact significatif sur Jobs-Y (**$\boldsymbol{\beta_1}$ ≠ 0)**
- Procédure de test
  - Utiliser la statistique t associée à $\boldsymbol{\beta_1}$
  - Comparer la statistique t à la valeur critique pour le niveau de significativité souhaité
- Niveaux de significativité
  - Niveau de significativité de 5 % : rejeter si $|t_{\beta_1}|$ ≥ 1,96, sinon accepter
  - Niveau de significativité de 1 % : rejeter si $|t_{\beta_1}|$ ≥ 2,58, sinon accepter
- Dans notre exemple, **= 11,24**, donc **est rejeté aux niveaux de significativité de 5% et 1%.**
- Par conséquent, GExp a un impact significatif (et positif) sur Jobs-Y.

Modèle de régression multiple

Définition de la régression multiple

- Relation linéaire entre les variables dépendantes et indépendantes

Composantes du modèle

- Variable dépendante (Y)
- Variables indépendantes (X1, X2, X3 ... Xn)

Équation de régression multiple

- **$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn + \varepsilon$**

Explication des termes

- $\beta 0$: Constante ou intercept
- $\beta 1$ ... βn : Coefficients mesurant la variation de Y
- ε : Résidu ou terme d'erreur

Tableau :
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

Estimation

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

- **Les performances commerciales sont-elles déterminées par l'intelligence ou l'extraversion d'une personne ?**
- **Estimation du modèle**
  - $\boldsymbol{Perf = \beta_0 + \beta_1 * Int + \beta_2 * Extr + \varepsilon}$ (3)
  - L'équation (3) est estimée à l'aide de la méthode des moindres carrés ordinaires (MCO)
- **Résultats présentés dans le tableau 4**
- **L'intelligence comme prédicteur**
  - Prédiction positive des performances hebdomadaires
  - Effet non statistiquement significatif (statistique t = 1,172, valeur p = 0,25)
- **L'extraversion comme facteur prédictif**
  - Prédicteur positif et significatif de la performance des ventes hebdomadaires
  - Statistique t = 2,53, valeur p = 0,02
- **Pouvoir explicatif du modèle**
  - Les variables explicatives expliquent 27% de la variance des ventes hebdomadaires
  - Inclusion potentielle d'autres variables pour améliorer le pouvoir prédictif

4.2. Analyse des séries temporelles

Plan

- Introduction à l'analyse des séries temporelles
- Objectifs de l'analyse des séries temporelles
- Types de modèles de séries temporelles
- Modèles d'autorégression
- Modélisation et prévision des séries temporelles
- Étude de cas : PIB réel du Nigeria (RGDP) sur la période 1970-2017

Définition des séries temporelles

Définition des séries temporelles

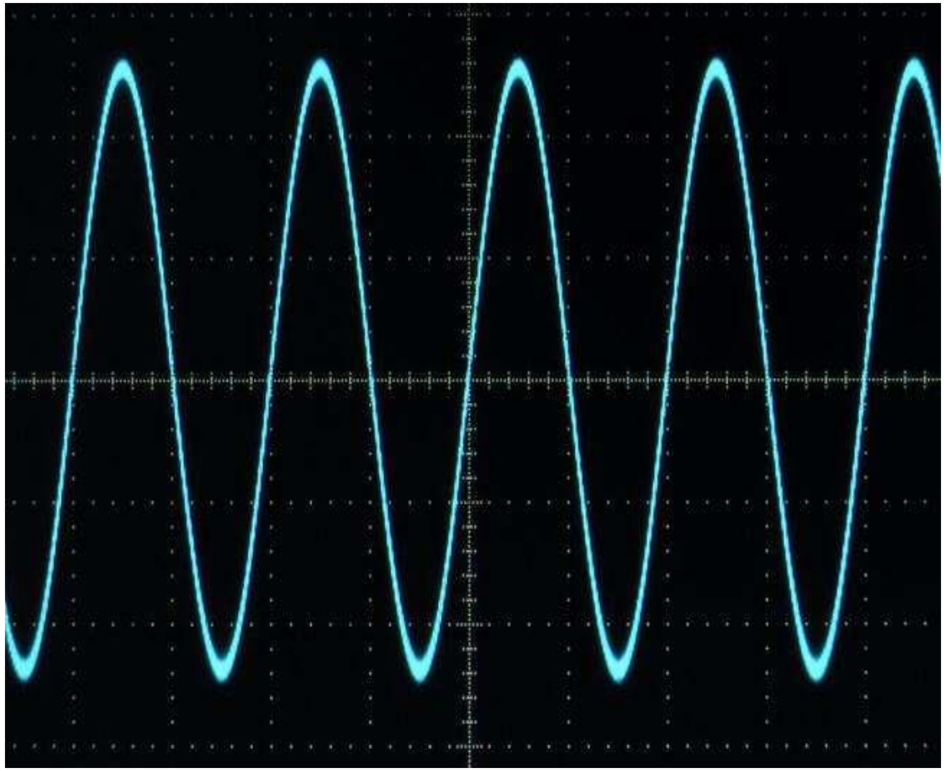
Un ensemble de données ordonnées dans le temps

Dépendance temporelle

Affichage de la dépendance temporelle

Le changement d'ordre peut modifier la significativité des données

Importance de l'analyse des séries temporelles



- **Analyse des séries temporelles**
  - Prévision des comportements futurs sur la base de modèles historiques
- **Différent de la régression linéaire standard**
  - Les données des séries temporelles ne sont pas nécessairement indépendantes
  - Les résidus des séries temporelles ne sont pas nécessairement identiquement distribués

Marche aléatoire

- **Définition de la marche aléatoire**
  - Processus stochastique avec des changements imprévisibles dans le temps
  - Dépend de la valeur précédente et de la perturbation aléatoire
- **Application aux données financières**
  - Utilisé pour modéliser les prix des actions
  - Valeurs futures supposées indépendantes des valeurs passées
- **Expression mathématique**

$$Y_t = Y_{t-1} + \varepsilon_t$$

  - Y_t : valeur au moment t
  - Y_{t-1} : valeur précédente
  - ε_t : terme d'erreur stochastique
- 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 se caractérise par un processus sans retour à la moyenne et qui peut s'en écarter dans un sens positif ou négatif.

Marche aléatoire

- **Rétention de la mémoire**
  - La valeur actuelle ne dépend que de la valeur précédente et d'un choc aléatoire.
  - Pas de dépendance directe à l'égard des valeurs passées au-delà de (Yt-1)
- **Non-stationnarité**
  - La variance augmente avec le temps
  - La moyenne peut changer de façon imprévisible
  - Les prévisions à long terme deviennent difficiles
- **Accumulation de chocs aléatoires**
  - Chaque nouvelle valeur résulte de l'addition des valeurs passées et d'une nouvelle composante aléatoire
  - Conduit à des fluctuations non limitées à long terme

Variations de la marche aléatoire

- **Marche aléatoire avec dérive**
  - Ajoute une composante de tendance constante (α)
  - Introduction d'une croissance ou d'un déclin systématique
- **Marche aléatoire avec tendance déterministe**
  - Inclut une tendance structurée dépendant du temps (βt)
  - Influence la trajectoire

Marche aléatoire avec dérive

- **Définition et composantes**
  - Inclut un terme de dérive constant α
  - Inclut un terme de bruit blanc εt
- **Caractéristiques**
  - Ne revient pas à une moyenne à long terme
  - A une variance dépendant du temps
- **Représentation des équations**

$$Yt = \alpha + Yt\text{-}1 + \varepsilon t$$

  - Yt est la valeur de la série au moment t
  - Yt-1 est la valeur précédente de la série
  - α représente la tendance constante
  - εt est un choc aléatoire de moyenne nulle et de variance constante

Caractéristiques de la marche aléatoire avec dérive

- **Présence d'une tendance**
  - Le terme de dérive (α) entraîne une tendance à la hausse ($\alpha > 0$) ou à la baisse ($\alpha < 0$) de la série au fil du temps.
- **Non-stationnarité**
  - La variance augmente avec le temps, ce qui rend la série non stationnaire
  - Impacts sur les prévisions et l'inférence statistique
- **Accumulation de chocs aléatoires**
  - Chaque valeur dépend de la précédente plus des chocs aléatoires
  - De petites variations peuvent s'accumuler au fil du temps et conduire à des écarts importants par rapport à la valeur initiale.

Modèle déterministe de tendance

- **Définition du modèle déterministe de tendance**
  - Tendance systématique déterminée par une forme fonctionnelle spécifique
  - Aucune composante aléatoire n'affecte la tendance
- **Hypothèses**
  - La tendance est prévisible
  - suit une structure fixe dans le temps
- **Utilité**
  - Utile pour les prévisions
  - La tendance sous-jacente est stable et prévisible
- Représentation du modèle

$$Y_t = \alpha + \beta t + \varepsilon_t$$

Marche aléatoire avec dérive et tendance déterministe

- **Définition de la marche aléatoire avec tendance déterministe**
  - Modèle de série temporelle combinant tendance systématique et fluctuations aléatoires
- **Expression mathématique**

$$Y_t = \alpha + Y_{t-1} + \beta t + \varepsilon_t$$

- **Composantes du modèle**
  - Y_t : Valeur de la série au moment t
  - Y_{t-1} : Valeur précédente de la série
  - α : Terme de dérive représentant une tendance constante
  - βt : Terme de tendance déterministe croissant ou décroissant linéairement avec le temps
  - ε_t : terme de bruit blanc de moyenne nulle et de variance constante

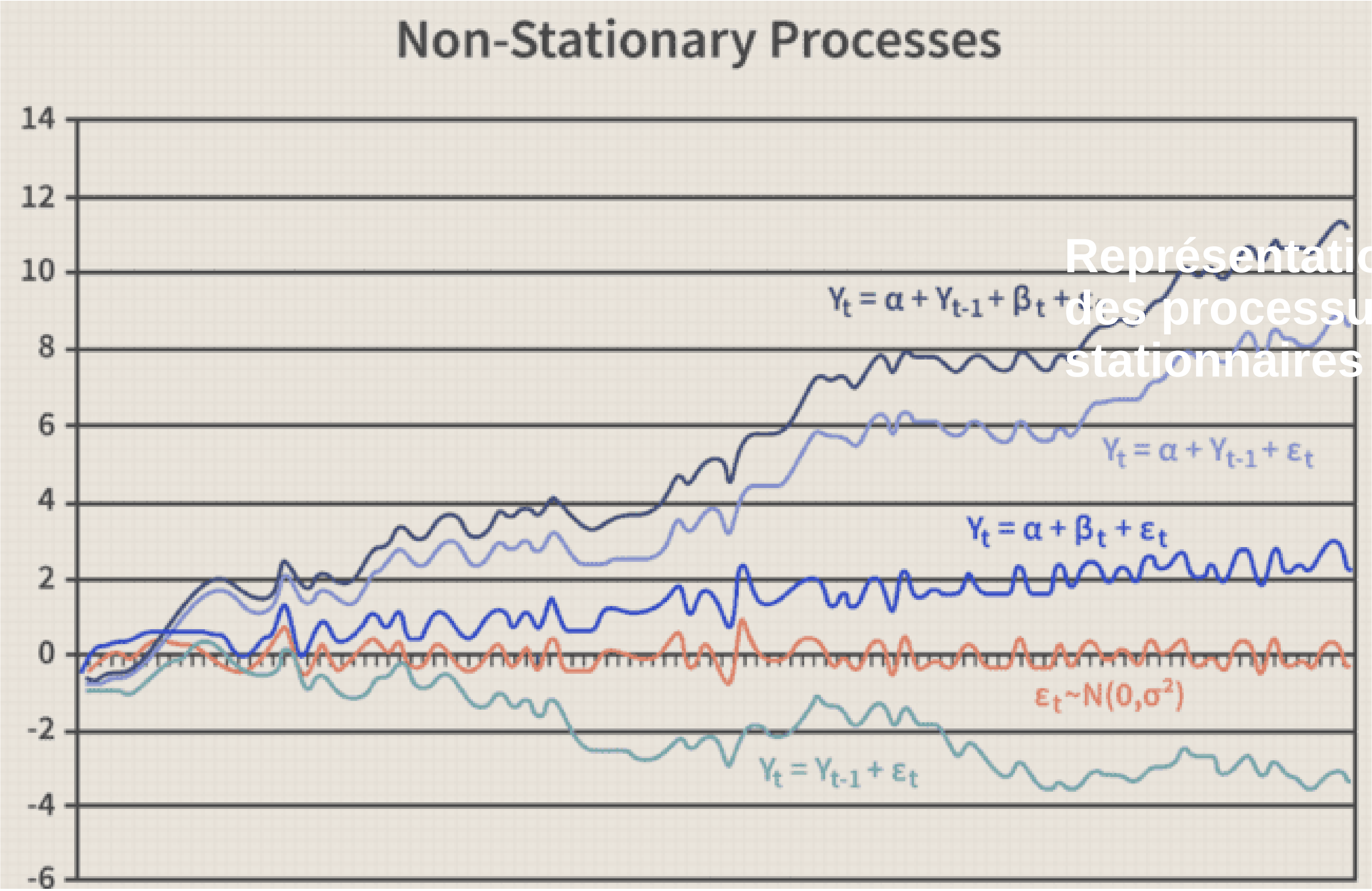
Caractéristiques de la marche aléatoire avec dérive et tendance déterministe

- **Tendance linéaire systématique**
  - Introduit une tendance déterministe
  - Augmentation ou diminution à taux fixe
- **Dérive constante**
  - Ajoute un biais régulier à la hausse ou à la baisse
  - Séparée de la tendance temporelle
- **Fluctuations aléatoires**
  - Composante du bruit blanc
  - Intégration de la variabilité stochastique
  - Permet un caractère aléatoire autour du modèle systématique
- **Non-stationnarité**
  - La variance augmente avec le temps

APPLICATIONS DE LA MARCHE ALÉATOIRE AVEC DÉRIVE ET TENDANCE DÉTERMINISTE

- **Indicateurs macroéconomiques**
  - PIB avec croissance et fluctuations aléatoires
- **Prix des actions**
  - Tendance à la croissance à long terme
- **Modèles de croissance démographique**
  - Des événements aléatoires influencent une trajectoire déterministe

Représentatio
des processu
stationnaires



Introduction aux modèles d'autorégression

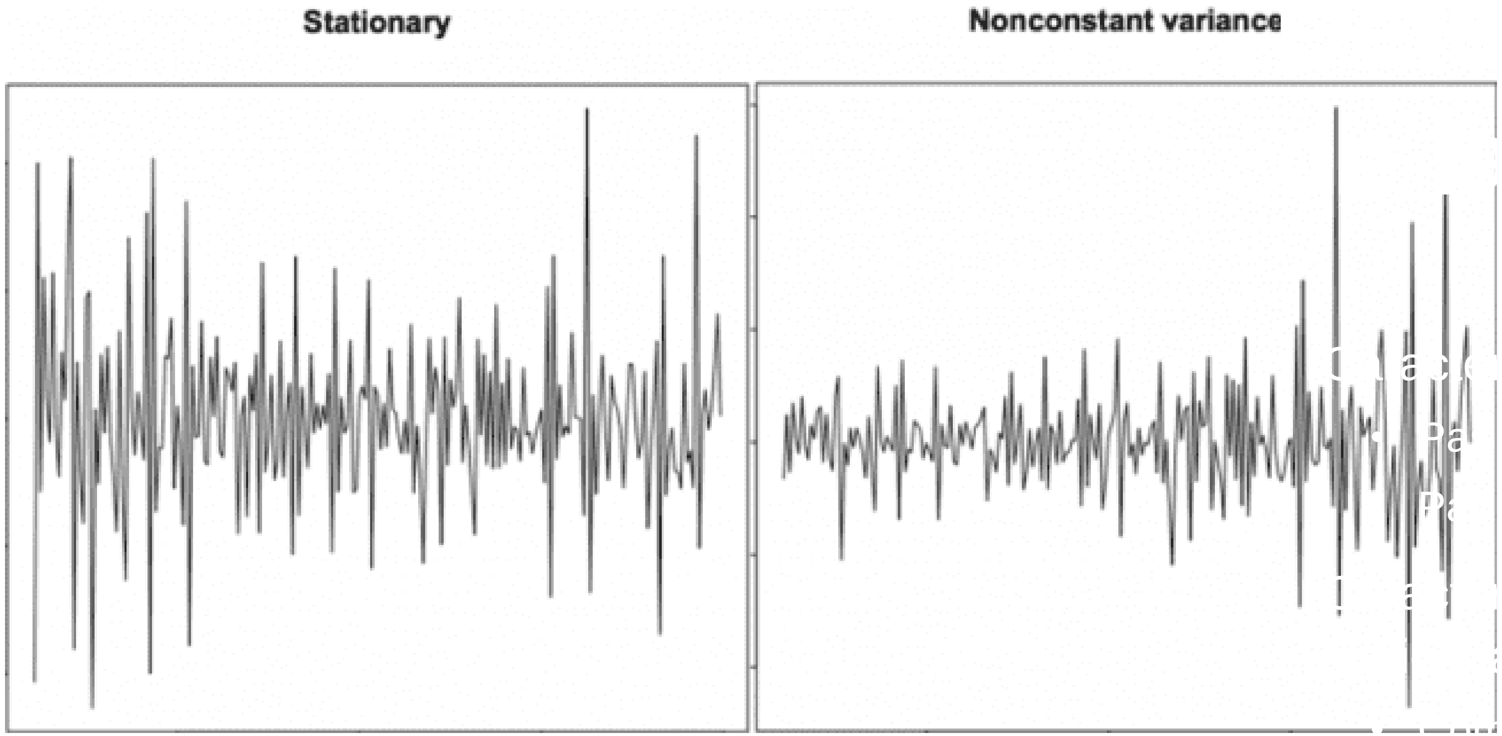
- **Processus d'autorégression (AR)**
  - Valeur actuelle exprimée comme une combinaison linéaire des valeurs passées
  - Inclut un terme d'erreur aléatoire
- **Modèles AR(1)**
  - Les équations 5 à 8 sont des versions différentes de AR(1)
  - Utilisé pour modéliser des données de séries temporelles
- Expression générale du processus de RA
  - Équation 9 :

$$Yt = \phi 1yt\text{-}1 + \phi 2yt\text{-}2 + \phi 3yt\text{-}3 + ... + \phi pyt\text{-}p + \varepsilon t$$

  - Yt est la valeur au moment t
  - Yt-1...Yt-p sont des valeurs retardées.
  - ϕ1...ϕp sont des coefficients
  - p représente l'ordre du processus AR

Importance de la stationnarité

- **Propriétés statistiques cohérentes**
  - Les séries stationnaires ont une moyenne, une variance et une autocorrélation constantes.
  - Permet aux modèles de capturer des motifs sans distorsion
- **Prévisibilité et précision du modèle**
  - Les modèles de prévision supposent la stationnarité pour des prévisions stables et précises.
  - Les séries non stationnaires conduisent à des prévisions peu fiables
- **Validité des hypothèses du modèle**
  - La plupart des modèles supposent un comportement stationnaire pour une inférence valide.
  - Les séries non stationnaires peuvent donner lieu à des résultats de régression erronés.
- **Une interprétation plus facile**
  - Les relations entre les variables restent cohérentes dans les séries stationnaires


Stationary
Nonconstant variance

Détection de la stationnarité : Tests paramétriques

- **Le test de Dickey-Fuller**

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

$$y_t = \rho y_{t-1} + \varepsilon_t \quad (6)$$

Note : à partir de l'équation (6), en soustrayant y_{t-1} des deux côtés, on obtient :

$$y_t - y_{t-1} = \rho y_{t-1} + \varepsilon_t - y_{t-1} \quad (7)$$

En appliquant un opérateur de différence, l'équation (7) peut donc être réécrite comme suit :

$$\Delta y_t = (\rho-1)y_{t-1} + \varepsilon_t = \delta y_{t-1} + \varepsilon_t \text{ , où } \delta = (\rho-1) \quad (8)$$

La méthode DF établit le test d'hypothèse suivant :

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

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

Si H_0 ne peut être rejetée, 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 une analyse sans transformation.

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

Remarquez qu'il s'agit d'un test unilatéral à gauche.

Détection de la stationnarité : Tests paramétriques

- <u>**Test de Dickey-Fuller augmenté (ADF)**</u>

➢ En raison de la possibilité de corrélation sérielle dans l'estimation de l'équation (8), le test de Dickey-Fuller augmenté (ADF) a été développé et est appliqué à un processus AR généralisé d'ordre supérieur de la forme :

$$\Delta y_t = \alpha + \beta_t + \gamma y_{t-1} + \delta_1 \Delta y_{t-1} + \delta_2 \Delta y_{t-2} + \ldots + \delta_{p-1} \Delta y_{t-p+1} + \varepsilon_t \quad \mathbf{(11)}$$

▪ Le test de racine unitaire est effectué sur γ comme suit :

Hypothèse nulle : H_0 : $\gamma = 0$

Hypothèse alternative : H_a : $\gamma < 0$

➢ **Si H_0 : $\gamma = 0$ ne peut être rejetée, cela indiquerait la présence d'une racine unitaire, et donc de données non stationnaires.**

➢ **Si H_0 : $\gamma = 0$ est rejetée, il n'y a pas de racine unitaire et les données sont stationnaires.**

Remarque : il s'agit également d'un test unilatéral vers la gauche, de sorte que seuls les nombres négatifs sont pris en compte. Les statistiques de test calculées doivent donc être plus négatives que la valeur critique correspondante (compte tenu de la taille de l'échantillon, du niveau de significativité souhaité et de l'inclusion ou non de la dérive et/ou de la tendance déterministe).

LA TRANSFORMATION DE DONNÉES NON STATIONNAIRES

DIFFERENCIER

- La différenciation consiste à éliminer les tendances temporelles et les caractéristiques saisonnières d'une donnée en calculant les différences entre les observations consécutives. Il peut être appliqué aux données plusieurs fois jusqu'à l'obtention de la stationnarité.
- 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... 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.

Classe de modèles ARIMA (Autoregressive Integrated Moving Averages)

Une fois la stationnarité établie, l'analyste se tourne vers la modélisation de la série. Les modèles ARIMA sont largement utilisés dans les séries temporelles, ARIMA signifiant Autoregressive Integrated Moving **Averages** models (modèles de **moyennes** mobiles intégrées autorégressives).

Cette catégorie de modèles peut être décomposée en trois éléments :

- Modèles **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 :
  - $y_t = \alpha + \phi\, y + \phi\, y + \phi_{1t-1\ 2t-2\ 3t-3} + y\ldots + \phi\, y_{pt-p} + \varepsilon_t$

où ε_t est supposé être un bruit blanc.

- Le nombre de retards significatifs de la variable qui sont conservés dans le modèle est appelé "ordre" du modèle AR. Ainsi, un modèle AR(1) est un modèle autorégressif d'ordre 1 ; et AR(2) est un modèle autorégressif d'ordre 2, etc...
- En général, un modèle AR(p) est connu comme un modèle autorégressif d'ordre p, où p représente le nombre de retards inclus dans le modèle
- Les modèles AR sont typiquement appliqués aux séries économiques, aux prix des actions, aux températures, etc...

Classe de modèles ARIMA (Autoregressive Integrated Moving Averages) (Suite)

2. Modèles 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 atténuer les fluctuations et mettre en évidence les tendances.
- En supposant que les erreurs sont distribuées de manière indépendante avec la distribution normale, un modèle MA peut être écrit comme suit :

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

- Le nombre de retards inclus dans le modèle MA est considéré comme l'"ordre" du modèle. Ainsi, 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 appelé modèle de moyenne mobile d'ordre q.

Classe de modèles ARIMA (Autoregressive Integrated Moving Averages) (Suite)

3. I : Intégré

- Une série temporelle est dite "intégrée" lorsqu'il est nécessaire d'appliquer une différenciation à la série afin de la rendre stationnaire et apte à l'estimation. Une fois encore, l'"ordre" d'intégration est important.
- En général, on dit qu'une série est intégrée d'ordre d, ou I(*d)*, si elle nécessite un nombre *d*'écarts d'ordre *d* pour devenir stationnaire lorsque $d \geq 0$.
- 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.
  - ...

Classe de modèles ARIMA (Autoregressive Integrated Moving Averages) (Suite)

4. Classe de modèles ARIMA

En ajoutant les équations (1) et (2) ci-dessus, nous pouvons maintenant représenter un modèle **ARMA(p, q)** (moyenne mobile autorégressive) général comme suit :

- $\boldsymbol{y_t = \alpha + \phi\, y + \phi\, y + \phi_{1t\text{-}1\ 2t\text{-}2\ 3t\text{-}3} + y\ldots + \phi\, y_{pt\text{-}p} + \varepsilon_t + \theta_1\, \varepsilon_{t\text{-}1} + \theta_2\, \varepsilon_{t\text{-}2} + \theta\, \varepsilon_{3t\text{-}3} + \ldots + \theta_q\, \varepsilon_{t\text{-}q}}$
- Si, en outre, y_t est intégré d'ordre d, le modèle ci-dessus devient un modèle **ARIMA(p, d, q)**, où p, d, q ≥ 0

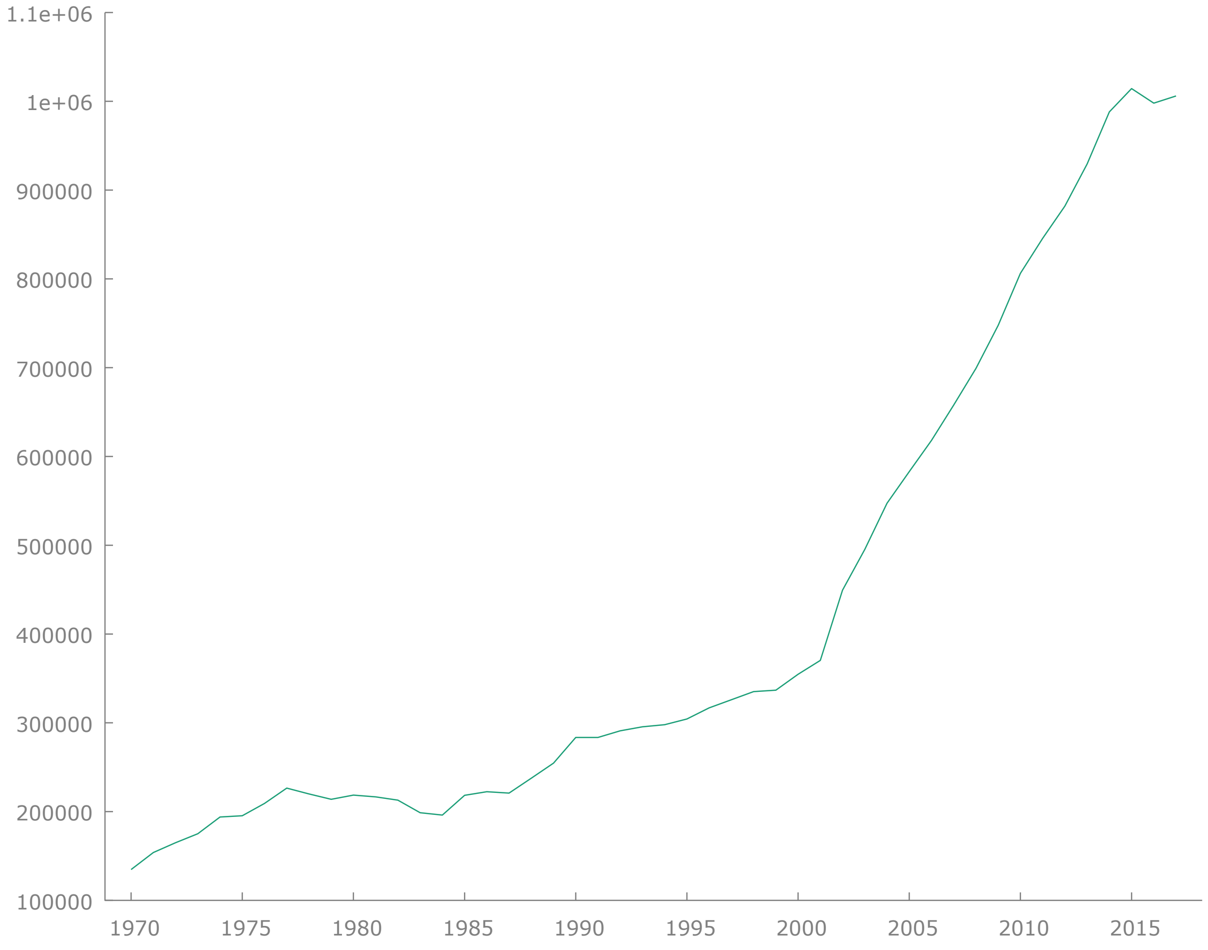
Exemples :

- Si $\boldsymbol{y_t}$ est un ARMA (1,0), alors $\boldsymbol{y_t = \alpha + \phi\, y_{1t\text{-}1} + \varepsilon_t}$
- Si $\boldsymbol{y_t}$ est un ARMA (0,1), alors $\boldsymbol{y_t = \alpha + \varepsilon_t + \theta_1\, \varepsilon_{t\text{-}1}}$
- Si $\boldsymbol{y_t}$ est un ARMA (1,1), alors $\boldsymbol{y_t = \alpha + \phi\, y_{1t\text{-}1} + \varepsilon_t + \theta_1\, \varepsilon_{t\text{-}1}}$
- Si, en outre, $\boldsymbol{y_t}$ est intégré d'ordre 1, le modèle ci-dessus devient un modèle ARIMA(1,1,1).

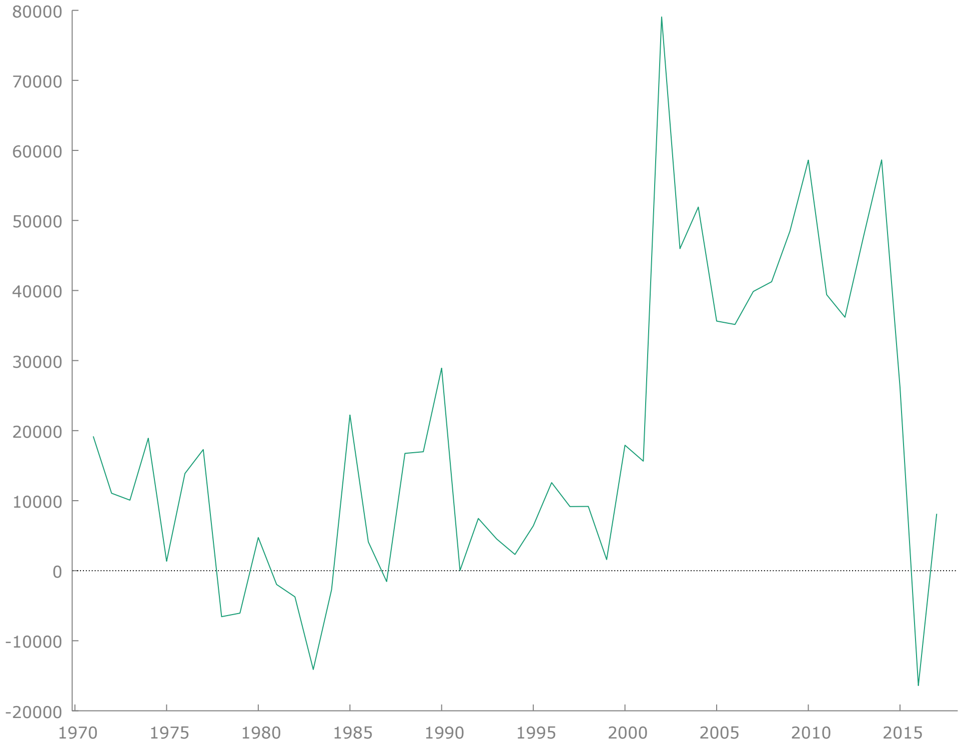

Examen du comportement du PIB
réel (RGDP) pour le cas du Nigéria
sur la période 1970-2017

Nigeria : RGDP

Nigeria : RGDP



Nigeria : d(RGDP)



Tests ADF RGDP du Nigeria :

Nigeria_RGDP

Test de Dickey-Fuller augmenté pour le RGDP

à partir de 9 retards, critère AIC

taille de l'échantillon 46

hypothèse nulle de racine unitaire : a = 1

test avec constante

y compris un décalage de (1-L)RGDP

modèle : (1-L)y = b0 + (a-1)*y(-1) + ... + e

valeur estimée de (a - 1) : 0.0117133

statistique de test : tau_c(1) = 1,02882

Valeur p asymptotique 0,997

Coefficient d' autocorrélation du 1er ordre pour e : -0,023

avec constante et tendance

y compris un décalage de (1-L)RGDP

modèle : (1-L)y = b0 + b1*t + (a-1)*y(-1) + ... + e

valeur estimée de (a - 1) : -0.0240578

statistique de test : tau_ct(1) = -1.15933

Valeur p asymptotique 0,9174

Coefficient d' autocorrélation du 1er ordre pour e : -0,047

d(RGDP)

Test de Dickey-Fuller augmenté pour d_RGDP

à partir de 9 retards, critère AIC

taille de l'échantillon 46

hypothèse nulle de racine unitaire : a = 1

test avec constante

y compris 0 retard de (1-L)d_RGDP

modèle : (1-L)y = b0 + (a-1)*y(-1) + e

valeur estimée de (a - 1) : -0.337011

statistique de test : tau_c(1) = -2.97198

Valeur p 0,04512

Coefficient d' autocorrélation du 1er ordre pour e : -0,087

avec constante et tendance

y compris 0 retard de (1-L)d_RGDP

modèle : (1-L)y = b0 + b1*t + (a-1)*y(-1) + e

valeur estimée de (a - 1) : -0.490275

statistique de test : tau_ct(1) = -3.64385

Valeur p 0,03689

Coefficient d' autocorrélation du 1er ordre pour e : -0,004

Test du RGDP du Nigeria en tant que modèle ARMA

gretl: model 22

File Edit Tests Save Graphs Analysis LaTeX

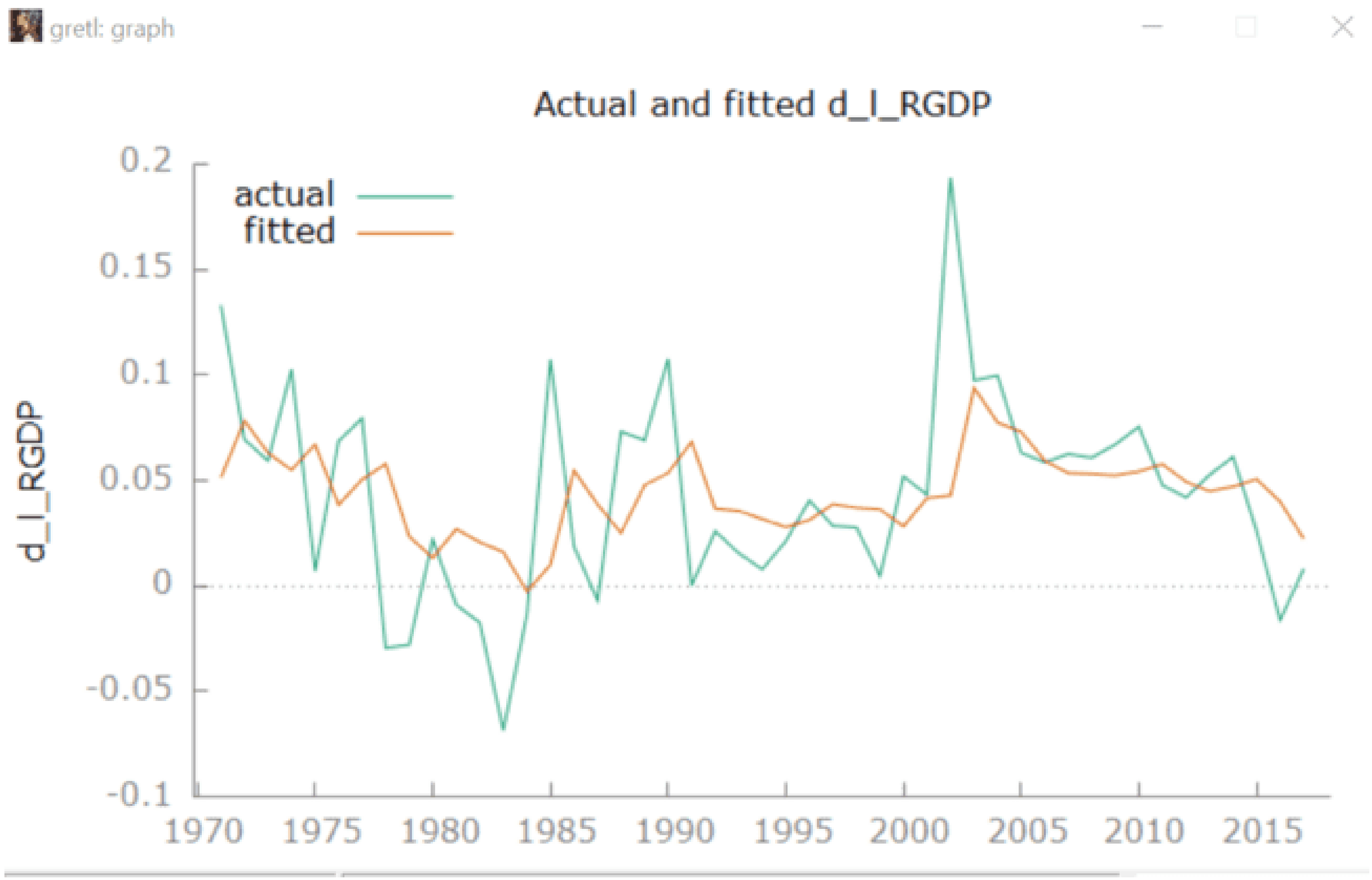
```
Function evaluations: 39
Evaluations of gradient: 13

Model 22: ARMA, using observations 1971-2017 (T = 47)
Estimated using AS 197 (exact ML)
Dependent variable: d_l_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
```

Graphique des données réelles vs données ajustées du RGDP différéncié d’ordre 1 (Nigeria)



4.3. Modèles vectoriels autorégressifs

Modèle vectoriel autorégressif (VAR)

1. Contexte

• Un modèle vectoriel autorégressif (VAR) décrit la relation dynamique entre **k** variables endogènes dans le temps. 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 est que les variables incluses dans un VAR s'influencent 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 VAR(*p*) comprend *p* retards des variables endogènes.

- 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 VAR(*p*) comprend *p* retards des variables endogènes.

- L'approche VAR a été introduite en 1980 par le professeur Christopher Sims comme un nouveau cadre d'analyse des données économiques et financières, avec la promesse de fournir "une approche cohérente et crédible de la description des données, de la prévision, de l'inférence structurelle et de l'analyse politique".

Modèle vectoriel autorégressif (VAR)

- <u>Exemple :</u> Un modèle VAR(p) où k=2 et p=1 est un VAR(1) à 2 variables et peut être exprimé sous forme matricielle comme suit :
- $\begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} + \begin{bmatrix} a_{1,1} & a_{1,2} \\ a_{2,1} & a_{2,2} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{bmatrix}$
- Ce qui est équivalent au système de 2 équations suivant :
- $y_{1,t} = c_1 + a_{1,1}y_{1,t-1} + a_{1,2}y_{2,t-1} + \varepsilon_{1,t}$
  $y_{2,t} = c_2 + a_{2,1}y_{1,t-1} + a_{2,2}y_{2,t-1} + \varepsilon_{2,t}$
- Par conséquent, l'observation actuelle (temps *t*) de chaque variable dépend de ses propres valeurs retardées ainsi que des valeurs retardées de l'autre variable dans le VAR.

2. Performance de l'approche VAR

- Les modèles VAR ont été des outils macroéconométriques efficaces et flexibles pour l'analyse des séries temporelles multivariées, en termes de description des données et de prévision. Cependant, lorsqu'il s'agit d'inférence structurelle et d'application politique, l'approche VAR seule n'est pas suffisante.
- 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 elle nécessite une analyse plus complexe des données, y compris l'utilisation de la théorie économétrique et de la connaissance institutionnelle.

Étude de cas du Nigeria, période 1970-2017

Résultats de l'estimation de la modélisation VAR du RGDP et de la consommation.

VAR system, lag order 1
OLS estimates, observations 1972-2017 (T = 46)

Equation 1: RGDP

	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
Const	0.024	0.008	2.796	0.0077	***
ld_RGDP_1	**0.325**	**0.138**	**2.349**	**0.0235**	**
ld_Cons_1	**0.060**	**0.051**	**1.171**	**0.2482**	

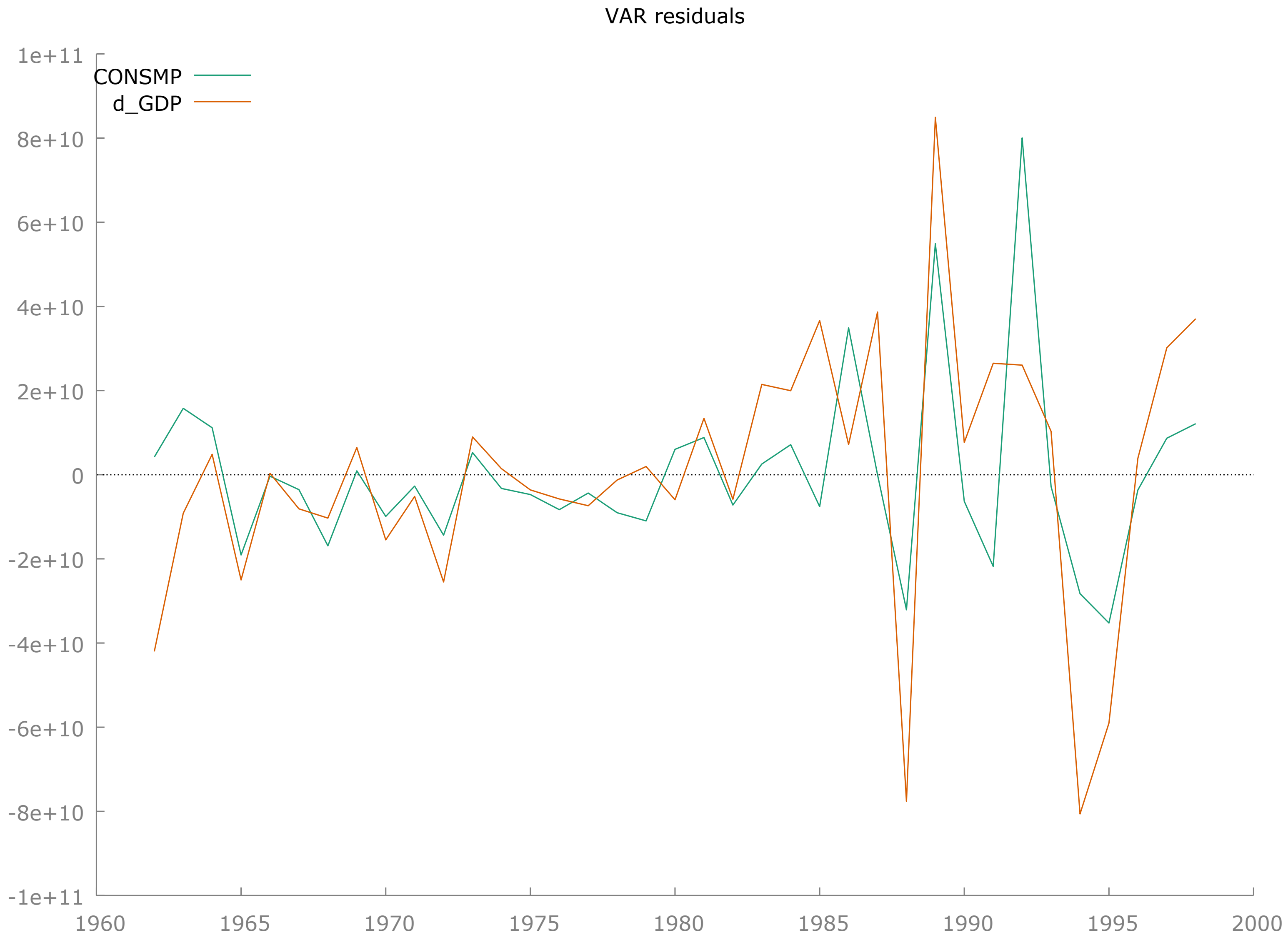
Mean dependent var	0.040814	S.D. dependent var	0.046462
Sum squared resid	0.080469	S.E. of regression	0.043259
R-squared	0.171622	Adjusted R-squared	0.133092
F(2, 43)	4.454323	P-value(F)	0.017455
Rho	−0.020566	Durbin-Watson	2.040129

Equation 2: Consumption

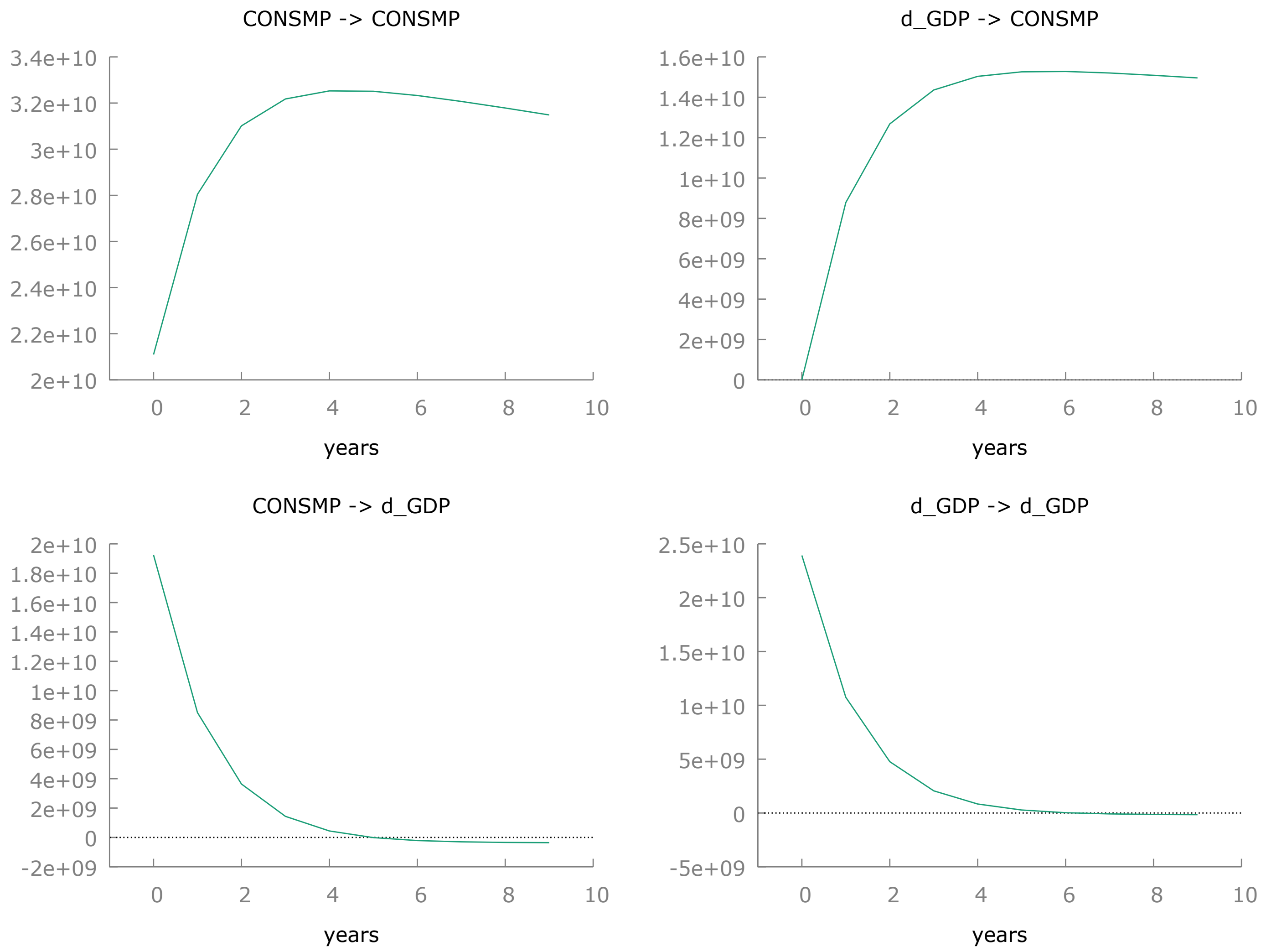
	Coefficient	*Std. Error*	*t-ratio*	*p-value*	
Const	0.010	0.023	0.432	0.6678	
ld_RGDP_1	**1.018**	**0.371**	**2.743**	**0.0088**	***
ld_Cons_1	**−0.414**	**0.137**	**−3.010**	**0.0044**	***

Mean dependent var	0.037637	S.D. dependent var	0.129369
Sum squared resid	0.576311	S.E. of regression	0.115770
R-squared	0.234778	Adjusted R-squared	0.199186
F(2, 43)	6.596423	P-value(F)	0.003173
Rho	−0.018770	Durbin-Watson	2.036548

RESIDUS DU VAR



FONCTIONS DE RÉPONSE IMPULSIONNELLE



MERCI!