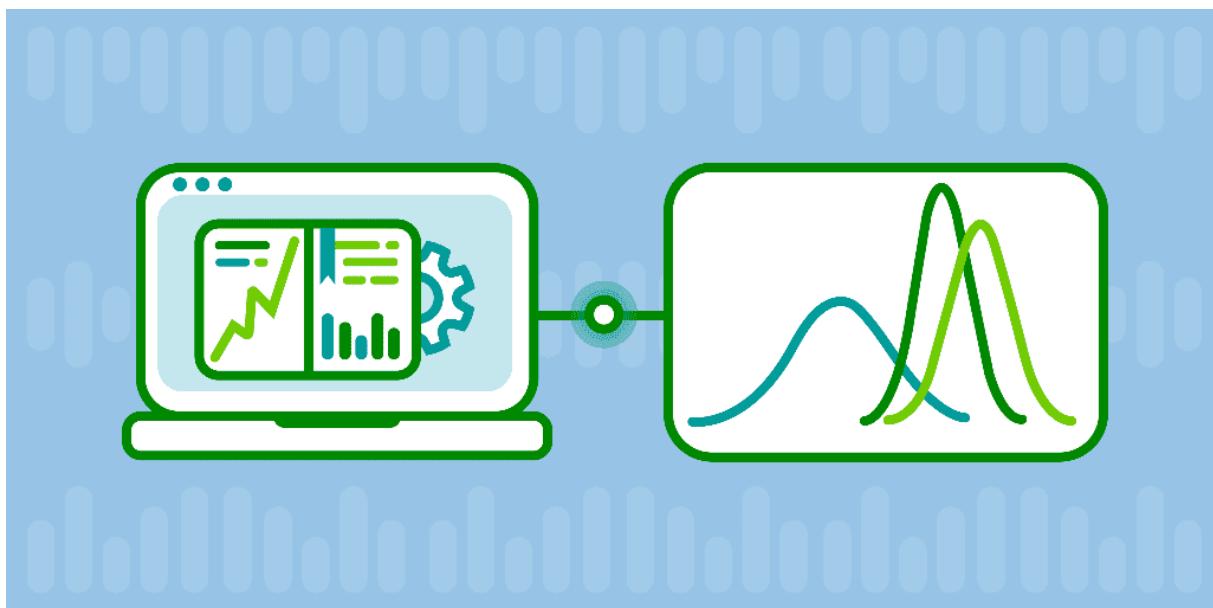


Économétrie bayésienne avec Python et Intelligence Artificielle : Théories, Données et Applications



Freudel AZOKLY

Table des matières

Introduction générale	3
Problématique et Hypothèses de Recherche	5
Hypothèses de Recherche	5
Partie 1 – Fondements théoriques	7
1.1 – Rappels et limites de l'économétrie classique.....	7
1.2 – Introduction à l'économétrie bayésienne.....	8
1.3 – Méthodes d'estimation bayésiennes.....	9
1.4 – Liens avec l'intelligence artificielle	11
1.5 – Bibliographie critique approfondie.....	12
A. Théorie bayésienne générale : Fondations et débats méthodologiques	12
B. Exemples pratiques et applications concrètes de l'économétrie bayésienne	14
C. Méthodes computationnelles avec Python : Des algorithmes au déploiement pratique.....	15
D. Applications économiques et articles récents : Concrétisations et innovations	17
E. Articles récents sur l'IA et Python en économétrie bayésienne : Synergies et perspectives	18
F. Ressources numériques complémentaires : un levier pour une mise en œuvre rigoureuse	20
Partie 2 – Données, variables et préparation	22
2.1 – Présentation des données : Source, nature et structure des données	22
2.2 – Variables	23
2.2.1 – Variable dépendante	23
2.2.2 – Variables explicatives principales	24
2.2.3 – Variables dérivées.....	24
2.2.4 – Justification économétrique	24
2.3 – Analyse descriptive	25
2.4-Méthodologie.....	33
2.4.1. Approche générale	33
2.4.2. Environnements et outils utilisés	33
2.4.3. Démarche économétrique	34
2.4. Implémentation des modèles.....	34
Partie 3 – Modélisation, estimation et évaluation	35
3.1 – Spécification du modèle bayésien	35
3.2 – Implémentation sous Python	37
3.3 – Analyse et interprétations des Résultats	41
1. Régression linéaire classique (MCO)	41
2. Modélisation bayésienne.....	45
3.4 Validation croisée et performance prédictive en MCO	51

3.5 Validation bayésienne.....	51
3.6 Synthèse	52
3.7 Statut des hypothèses	53
3.8– Extension IA / Machine Learning.....	53
Annexes	57
Bibliographie.....	68

Introduction générale

L'économétrie apparaît comme étant un pilier central de l'analyse économique moderne. Elle permet de tester des hypothèses théoriques, d'estimer des relations quantitatives entre variables,

et de guider les politiques publiques ou les décisions privées à partir de données empiriques. Les méthodes économétriques traditionnelles, notamment la méthode des moindres carrés ordinaires (MCO) ou le maximum de vraisemblance (MLE), ont longtemps dominé la pratique. Ces approches supposent généralement des hypothèses fortes (indépendance, normalité, homoscédasticité) et se focalisent sur des estimations ponctuelles des paramètres. Cependant, elles peuvent être limitées dans des contextes incertains, lorsque l'échantillon est petit, ou lorsqu'il est nécessaire d'intégrer des informations préexistantes. C'est dans ce cadre que l'économétrie bayésienne apparaît comme une alternative pertinente et de plus en plus explorée.

L'économétrie bayésienne repose sur le théorème de Bayes, qui permet de mettre à jour des connaissances initiales (les **a priori**) en fonction des données observées, pour obtenir une distribution **a posteriori** des paramètres du modèle. Contrairement aux méthodes classiques, cette approche considère les paramètres comme des variables aléatoires et non comme des constantes fixes. Elle permet ainsi de mieux représenter l'incertitude et d'estimer, par exemple, la probabilité qu'un paramètre soit supérieur à un certain seuil, plutôt que de se limiter à un simple intervalle de confiance. Ce cadre devient particulièrement utile dans les situations où les données sont limitées, bruyantes, ou lorsqu'on souhaite combiner différentes sources d'information.

En parallèle, l'intelligence artificielle (IA) et, plus spécifiquement, l'apprentissage automatique (machine learning), ont profondément transformé les méthodes d'analyse de données. L'IA permet de modéliser des relations complexes, non linéaires, et d'apprendre automatiquement des structures cachées dans les données massives. Les outils d'IA s'appuient souvent sur des techniques bayésiennes, notamment dans les modèles probabilistes ou les méthodes d'inférence approximative (comme l'échantillonnage MCMC ou les méthodes variationnelles). Ainsi, l'économétrie bayésienne et l'intelligence artificielle ne sont pas des domaines disjoints, mais peuvent être vus comme complémentaires, partageant un socle méthodologique commun et des objectifs similaires : tirer le meilleur parti des données pour informer la décision.

Dans ce contexte, ce mémoire propose d'explorer les fondements théoriques et les applications pratiques de l'économétrie bayésienne, en s'appuyant sur le langage Python, largement utilisé aujourd'hui dans les communautés scientifiques et techniques. Trois grands objectifs structurent cette recherche. Le premier est de présenter de façon claire les concepts de base de l'économétrie bayésienne et de les comparer aux approches classiques, en mettant l'accent sur la manière dont elle gère l'incertitude. Le deuxième objectif est de construire un modèle bayésien appliqué à l'analyse du **PIB par habitant**, en mobilisant des variables macroéconomiques telles que les dépenses publiques, le niveau d'éducation ou encore le taux d'emploi. Enfin, le troisième objectif est de comparer les résultats obtenus par l'approche bayésienne avec ceux issus des méthodes économétriques classiques. Cette comparaison permettra de discuter les apports concrets des distributions *a posteriori*, des intervalles de crédibilité, et plus généralement de la philosophie bayésienne pour l'analyse économique.

Ce mémoire est structuré en trois parties. La première partie présente les fondements théoriques de l'économétrie bayésienne ainsi que ses différences avec les approches fréquentistes. La deuxième partie développe un modèle empirique pour étudier les déterminants du PIB par habitant à partir d'un jeu de données macroéconomiques. Enfin, la troisième partie compare les

résultats des différentes approches, discute leur interprétation, et ouvre sur des perspectives liées à l'intelligence artificielle et à l'usage des méthodes bayésiennes dans l'économie numérique.

Problématique et Hypothèses de Recherche

Dans un contexte économique mondial marqué par la recherche d'une croissance inclusive et durable, la compréhension des facteurs qui influencent le PIB par habitant est un enjeu crucial pour orienter les politiques publiques. Si les méthodes économétriques classiques, comme les moindres carrés ordinaires (MCO) ou le maximum de vraisemblance (MLE), ont permis d'identifier certains déterminants (éducation, dépenses publiques, emploi), elles reposent sur des hypothèses souvent restrictives et ne rendent pas toujours compte de l'incertitude entourant les estimations.

Or, les décideurs ont besoin d'indicateurs qui expriment non seulement une estimation centrale, mais aussi une mesure claire du degré de confiance qu'on peut lui accorder. L'économétrie bayésienne, en intégrant l'incertitude via des distributions *a priori* et *a posteriori*, propose une approche plus souple et potentiellement plus informative. Pourtant, cette méthode reste encore peu utilisée dans l'analyse macroéconomique du PIB par habitant.

Dès lors, une question centrale se pose : **l'économétrie bayésienne permet-elle une meilleure compréhension et prévision du PIB par habitant, par rapport aux méthodes classiques ?** Cette interrogation soulève plusieurs enjeux : l'identification rigoureuse des variables macroéconomiques pertinentes, la construction d'un modèle bayésien interprétable, et l'évaluation de la plus-value des résultats bayésiens en termes d'aide à la décision.

Hypothèses de Recherche

L'analyse empirique repose sur trois hypothèses principales, formulées à partir des données disponibles et de la littérature économique.

Hypothèse 1 :

Les dépenses publiques en éducation et en santé ont un effet positif et significatif sur le PIB par habitant.

Cette hypothèse s'inscrit dans la continuité des théories de la croissance endogène, qui attribuent un rôle central au capital humain dans le développement économique. Les variables edu_spend (dépenses d'éducation en % du PIB) et health_spend (dépenses de santé en % du PIB), ainsi que leur somme edu_health, seront mobilisées pour estimer leur impact sur log_gdp (logarithme du PIB par habitant). L'objectif est de déterminer si ces investissements publics sont systématiquement associés à une augmentation du niveau de vie, et de quantifier la force et la fiabilité de cet effet dans une perspective probabiliste.

Hypothèse 2 :

L'approche bayésienne permet une meilleure quantification de l'incertitude entourant les estimations que les méthodes économétriques classiques.

Au-delà de la seule estimation des coefficients, cette hypothèse porte sur la capacité de l'économétrie bayésienne à fournir une représentation plus complète de l'incertitude. Contrairement aux intervalles de confiance fréquents, les intervalles de crédibilité bayésiens permettent d'exprimer la plausibilité des effets estimés en intégrant les données et les a priori. L'analyse comparera ainsi les résultats bayésiens avec ceux obtenus par moindres carrés ordinaires (MCO), en examinant la largeur des intervalles, la forme des distributions a posteriori et la probabilité que certains effets dépassent un seuil donné.

Hypothèse 3 :

Un modèle bayésien intégrant les principales variables explicatives offre de meilleures performances prédictives du PIB par habitant qu'un modèle par OLS.

Cette hypothèse vise à évaluer l'apport concret de l'approche bayésienne en matière de prévision économique. En utilisant des critères de performance tels que le RMSE (erreur quadratique moyenne), la log-vraisemblance prédictive ou les scores de couverture, il sera possible de comparer la précision des modèles bayésien et classique dans la prédiction du PIB par habitant. L'enjeu est d'identifier si l'approche bayésienne, grâce à ses mécanismes d'inférence, offre une robustesse accrue face à l'hétérogénéité des pays et à la qualité variable des données.

Partie 1 – Fondements théoriques

1.1 – Rappels et limites de l'économétrie classique

L'économétrie classique constitue le socle de la modélisation empirique en économie depuis plusieurs décennies. Elle repose principalement sur deux méthodes d'estimation : la méthode des moindres carrés ordinaires (MCO) pour les modèles linéaires, et le maximum de vraisemblance (MLE) pour des modèles plus généraux, notamment lorsque les erreurs ne suivent pas une loi normale ou que la structure du modèle est non linéaire. Ces deux approches partagent des fondements fréquentistes : l'inférence repose uniquement sur les données observées, sans incorporation explicite d'informations a priori. Les estimateurs obtenus via MCO sont réputés être les meilleurs dans la classe des estimateurs linéaires sans biais (théorème de Gauss-Markov), à condition que certaines hypothèses soient respectées c'est à dire la linéarité de la relation entre les variables, l'absence d'autocorrélation, l'homoscédasticité des erreurs, la normalité des résidus pour les inférences, et l'indépendance des observations. Ces hypothèses permettent de construire des outils statistiques familiers : tests de significativité (t de Student), tests globaux (F de Fisher), intervalles de confiance, ou encore coefficients de détermination (R^2).

Cependant, l'application de ces méthodes en pratique révèle plusieurs limites structurelles. Premièrement, la petite taille d'échantillon, fréquente dans les études économiques spécifiques (secteurs, régions, données longitudinales courtes), remet en question les propriétés asymptotiques des estimateurs. Dans ces cas, les intervalles de confiance et les tests statistiques sont moins fiables, car basés sur une approximation qui suppose un grand nombre d'observations. Deuxièmement, la multicolinéarité, c'est-à-dire la forte corrélation entre variables explicatives, complique l'interprétation des coefficients estimés et augmente leur variance. Cela rend les estimations instables : une petite variation dans les données peut entraîner un changement important dans les résultats, réduisant ainsi leur robustesse.

Troisièmement, la présence de données manquantes, d'erreurs de mesure ou de valeurs aberrantes est courante dans les bases de données économiques, mais difficile à gérer dans le cadre fréquentiste sans recourir à des techniques ad hoc. Ces anomalies peuvent biaiser les résultats ou affaiblir leur précision, notamment lorsque les mécanismes de données manquantes sont ignorés ou mal modélisés. Quatrièmement, la structure rigide du modèle linéaire peut être inadaptée lorsque les relations entre les variables sont en réalité non linéaires ou hétérogènes selon les sous-groupes observés (effets différenciés par sexe, âge, niveau de revenu, etc.). Ainsi, les modèles classiques ont du mal à intégrer cette complexité sans multiplier les hypothèses ou recourir à des segmentations souvent arbitraires.

Enfin, l'un des reproches majeurs adressés à l'économétrie classique réside dans sa gestion limitée de l'incertitude. En effet, les estimations sont souvent accompagnées d'un intervalle de confiance, mais celui-ci ne reflète qu'une incertitude conditionnelle sur les données observées, sans intégrer d'information externe ni permettre une interprétation probabiliste directe. Par exemple, un intervalle de confiance à 95 % ne signifie pas que le paramètre a 95 % de chances d'être dans cet intervalle, mais plutôt que si l'on répétait l'échantillonnage un grand nombre de

fois, 95 % des intervalles contiendraient le vrai paramètre. Cette nuance, souvent mal comprise, peut limiter l'interprétation et l'usage des résultats dans les prises de décision concrètes.

Ces différentes faiblesses soulignent justement la nécessité de méthodes économétriques alternatives capables de mieux intégrer l'incertitude, de traiter les problèmes de petite taille d'échantillon, et d'exploiter toutes les sources d'information disponibles. Dans ce contexte, l'approche bayésienne apparaît donc comme une réponse pertinente et prometteuse, car elle permet de construire des modèles plus souples, d'inclure des connaissances préalables de manière formalisée, et de fournir des résultats interprétables directement en termes de probabilités.

1.2 – Introduction à l'économétrie bayésienne

L'économétrie bayésienne propose une alternative conceptuelle et méthodologique à l'économétrie classique, fondée sur une autre vision de la probabilité. Alors que l'approche fréquentiste considère les paramètres comme des constantes fixes à estimer à partir des données, l'approche bayésienne traite ces paramètres comme des variables aléatoires, dont on cherche à estimer une distribution de probabilité. Le cœur de cette démarche repose sur le **théorème de Bayes**, qui exprime comment actualiser nos connaissances sur un paramètre à partir de nouvelles observations. Formellement, la distribution a posteriori des paramètres est obtenue en combinant la distribution a priori (qui représente l'information disponible avant les données) avec la vraisemblance (l'information apportée par les données), selon la formule :

$$\text{Posterior} \propto \text{Vraisemblance} \times \text{Prior}$$

Cette perspective permet de replacer l'analyse économétrique dans un cadre véritablement probabiliste, en mettant au centre de la réflexion l'incertitude et sa modélisation. Contrairement aux intervalles de confiance fréquentistes (souvent mal interprétés), l'approche bayésienne fournit des intervalles de crédibilité, c'est-à-dire des intervalles dans lesquels le paramètre a une probabilité donnée de se trouver, compte tenu des données. Cette interprétation directe des résultats facilite leur communication, notamment auprès de décideurs ou de non-spécialistes.

Par ailleurs, l'un des atouts majeurs de l'économétrie bayésienne réside dans sa grande flexibilité. Elle permet d'intégrer des informations a priori provenant de la littérature, d'experts ou d'études antérieures. Cette possibilité est particulièrement utile en économie, où les données sont souvent rares, coûteuses ou entachées d'incertitudes structurelles. Les modèles bayésiens peuvent ainsi être conçus pour refléter les connaissances préexistantes sur un phénomène, et s'adapter à des échantillons de petite taille sans nécessiter de simplifications excessives.

De plus, l'approche bayésienne est particulièrement adaptée aux situations économétriques complexes, notamment en cas de multicolinéarité, de non-normalité des erreurs, de données manquantes ou de structures hiérarchiques. Par exemple, dans un modèle avec effets fixes

régionaux ou temporels, le bayésianisme permet de construire des modèles hiérarchiques (ou multiniveaux), dans lesquels les effets spécifiques à chaque groupe peuvent être modélisés comme issus d'une distribution commune. Ce type de modélisation améliore l'estimation en partageant l'information entre groupes, tout en préservant la spécificité de chaque entité.

Un autre point fort réside dans l'approche intégrée et cohérente du traitement de l'incertitude : toutes les sources d'incertitude (sur les données, les paramètres, la structure du modèle) sont prises en compte dans une même démarche probabiliste. Cela permet d'éviter certains paradoxes ou biais d'inférence propres à l'approche fréquentiste, notamment en cas de tests multiples, de sélection de modèles, ou de données fortement bruitées.

Pour finir, les progrès récents en informatique et en intelligence artificielle ont rendu l'utilisation de l'économétrie bayésienne plus accessible, notamment grâce aux méthodes d'estimation numérique comme les chaînes de Markov par Monte Carlo (MCMC), et aux bibliothèques Python modernes (PyMC, NumPyro, Stan via CmdStanPy). Ces outils permettent de simuler efficacement les distributions *a posteriori* même dans des modèles complexes, rendant ainsi le bayésianisme applicable dans la pratique à grande échelle.

En somme, l'économétrie bayésienne constitue une avancée méthodologique majeure pour l'analyse économique. En intégrant la notion d'incertitude dès la construction du modèle, en offrant une grande flexibilité structurelle, et en bénéficiant des apports récents de l'informatique et de l'IA, elle apparaît comme un cadre particulièrement adapté pour l'étude approfondie des déterminants macroéconomiques du PIB par habitant. Ce potentiel sera mis à l'épreuve dans la suite de ce mémoire à travers un exemple concret de modélisation et d'estimation bayésienne.

1.3 – Méthodes d'estimation bayésiennes

L'estimation en économétrie bayésienne repose sur la caractérisation de la distribution *a posteriori* des paramètres du modèle, qui combine l'information issue des données (vraisemblance) avec les connaissances *a priori* sous forme d'une distribution *a priori*. Cependant, cette distribution *a posteriori* est souvent difficile, voire impossible à obtenir sous une forme analytique fermée, surtout pour des modèles complexes ou non linéaires. Pour contourner cette difficulté, la communauté statistique et économétrique s'appuie sur des méthodes numériques dites de simulation, dont les plus utilisées sont les méthodes de Monte Carlo par chaînes de Markov (MCMC).

Les méthodes **MCMC** permettent de générer un grand nombre d'échantillons dépendants, appelés « chaîne de Markov », qui convergent en distribution vers la distribution cible, c'est-à-dire la distribution *a posteriori* des paramètres. Ces échantillons simulés sont ensuite utilisés pour approximer les caractéristiques de la distribution, telles que les moments (moyenne, variance), les quantiles, ou pour construire des intervalles de crédibilité bayésiens. L'intérêt de MCMC est qu'elle offre une solution universelle pour estimer des modèles bayésiens très flexibles, même lorsque les calculs analytiques sont intraitables.

Parmi les algorithmes MCMC classiques, on distingue le **Gibbs Sampling**, qui est une méthode itérative où chaque paramètre est tiré séquentiellement à partir de sa distribution conditionnelle

complète, c'est-à-dire la distribution du paramètre sachant toutes les autres variables et les données. Cette méthode est particulièrement efficace lorsque ces distributions conditionnelles sont simples et connues explicitement. Par exemple, dans un modèle linéaire bayésien avec des lois a priori conjuguées, le Gibbs Sampling permet d'échantillonner successivement les coefficients et la variance de l'erreur. Toutefois, sa mise en œuvre peut être limitée si les distributions conditionnelles ne sont pas facilement accessibles.

L'**algorithme Metropolis-Hastings** est une généralisation qui fonctionne même quand les distributions conditionnelles ne sont pas simples à échantillonner directement. Ce procédé propose un échantillon candidat selon une distribution proposée, puis décide de l'accepter ou de le rejeter selon un critère probabiliste, garantissant que la chaîne de Markov converge vers la distribution cible. Bien que très flexible, cet algorithme nécessite un bon choix de la distribution de proposition pour assurer une convergence rapide, sans quoi la chaîne peut se « coincer » dans certaines régions de l'espace des paramètres, ce qui diminue l'efficacité de l'estimation.

Dans les modèles économétriques modernes, particulièrement ceux à grande dimension ou non linéaires, les méthodes classiques MCMC peuvent être coûteuses en temps de calcul et produire des chaînes fortement corrélées. Pour répondre à ces limites, des méthodes plus récentes ont été développées, comme le **Hamiltonian Monte Carlo (HMC)**. Cette technique s'appuie sur des principes issus de la mécanique hamiltonienne : on introduit des variables auxiliaires (appelées « momenta ») et on simule un système dynamique qui explore efficacement l'espace des paramètres en suivant des trajectoires continues. Cette exploration permet de proposer des mouvements « guidés » vers des zones à haute probabilité, réduisant ainsi l'auto-corrélation des échantillons et accélérant la convergence.

L'algorithme **No-U-Turn Sampler (NUTS)** est une amélioration adaptative du HMC qui évite la nécessité de définir a priori la longueur des trajectoires hamiltoniennes, un paramètre délicat à régler. NUTS détecte automatiquement quand la trajectoire commence à revenir en arrière (« no U-turn »), ce qui garantit un échantillonnage efficace et évite le surcoût computationnel. Cette méthode est aujourd'hui largement utilisée dans les logiciels d'inférence bayésienne, tels que **Stan** ou **PyMC3/4**, qui facilitent grandement l'application des modèles bayésiens complexes dans divers domaines, y compris l'économétrie.

Au-delà des aspects techniques, l'adoption de ces méthodes numériques avancées ouvre la voie à des analyses plus riches : elles permettent d'incorporer des modèles hiérarchiques complexes, de gérer des données manquantes ou des structures temporelles dynamiques, et surtout d'obtenir une quantification rigoureuse de l'incertitude. Cette flexibilité accrue est essentielle pour répondre aux défis posés par les données économiques réelles, souvent incomplètes, bruitées, ou non conformes aux hypothèses classiques.

Ainsi, maîtriser ces différentes techniques d'estimation bayésienne est un prérequis indispensable pour tirer pleinement parti de la puissance de l'économétrie bayésienne, en particulier dans l'analyse macroéconomique que ce mémoire propose. Les chapitres suivants mettront en œuvre ces méthodes pour modéliser et interpréter les déterminants du PIB par habitant, en exploitant la richesse de l'approche bayésienne.

1.4 – Liens avec l'intelligence artificielle

L'intelligence artificielle (IA) et l'économétrie bayésienne sont aujourd'hui de plus en plus interconnectées, ouvrant de nouvelles perspectives pour la modélisation économique et l'analyse de données complexes. L'essor des données massives (big data), des capacités de calcul accrues, et des algorithmes sophistiqués ont permis à l'IA d'apporter des solutions innovantes aux défis traditionnels de l'économétrie.

D'abord, l'IA excelle dans le **traitement et l'analyse de volumes importants de données hétérogènes**, souvent non structurées, telles que des séries temporelles, des images, du texte ou des données comportementales. Cette capacité est particulièrement utile en économie, où les sources de données sont multiples (indicateurs macroéconomiques, données microéconomiques, données alternatives comme celles issues des réseaux sociaux ou de la géolocalisation). Les modèles de deep learning, notamment, permettent d'extraire automatiquement des représentations complexes, ce qui améliore la qualité et la pertinence des variables explicatives intégrées dans les modèles économétriques.

L'**automatisation des processus de modélisation** est un autre avantage majeur apporté par l'IA. Des algorithmes d'apprentissage automatique permettent de sélectionner automatiquement les variables pertinentes, d'identifier des relations non linéaires, voire d'adapter dynamiquement la structure du modèle en fonction des données. Cette automatisation est essentielle pour gérer la complexité croissante des données économiques et pour éviter les erreurs humaines dans la spécification des modèles. Couplée à l'économétrie bayésienne, cette automatisation donne naissance à des méthodes d'**inférence approximative** qui permettent de résoudre des problèmes d'estimation statistique difficiles, notamment lorsque le calcul exact des distributions a posteriori est prohibitif.

Parmi ces méthodes, la **variation bayésienne** se distingue par son efficacité à approcher la distribution a posteriori en optimisant une fonction de divergence entre la distribution approchée et la distribution vraie. Cette approche est particulièrement utile pour entraîner des modèles complexes issus de l'IA, comme les réseaux de neurones profonds, tout en conservant la rigueur probabiliste de l'approche bayésienne.

Le champ du **deep learning bayésien** illustre parfaitement la convergence entre IA et économétrie bayésienne. Par exemple, **Bayes by Backprop** est une méthode innovante qui intègre la propagation en arrière (backpropagation) dans un cadre bayésien. Plutôt que d'estimer un unique ensemble de poids pour un réseau de neurones, cette méthode estime une distribution de probabilité sur ces poids, ce qui permet d'incorporer explicitement l'incertitude des paramètres dans les prédictions. Ceci est crucial en économie, où une mauvaise estimation peut conduire à des décisions coûteuses.

Une autre technique phare est le **Dropout**, d'abord conçu pour prévenir le surapprentissage en désactivant aléatoirement des neurones pendant l'entraînement. Des travaux récents ont montré que le Dropout peut être interprété comme une approximation bayésienne, fournissant une

mesure de l'incertitude sur les prédictions du réseau. Cette approche simple mais puissante permet ainsi de bénéficier d'une estimation probabiliste sans coûts computationnels excessifs.

Au-delà de ces techniques, l'IA apporte aussi des **outils pour l'inférence approximative**, tels que les réseaux de neurones variationnels (Variational Autoencoders, VAE) et les méthodes de Monte Carlo variationnelle, qui facilitent l'estimation des modèles bayésiens dans des contextes où les modèles sont trop complexes pour des méthodes classiques.

Enfin, cette synergie entre IA et économétrie bayésienne ouvre la voie à des applications concrètes dans divers domaines économiques : prévision macroéconomique, analyse des comportements consommateurs, évaluation des risques financiers, ou encore analyse des politiques publiques. En combinant la puissance d'apprentissage des modèles d'IA avec la rigueur probabiliste de l'approche bayésienne, il devient possible d'obtenir des modèles à la fois performants, robustes, et capables de quantifier précisément l'incertitude, offrant ainsi des outils plus fiables pour la prise de décision économique.

1.5 – Bibliographie critique approfondie

La bibliographie critique constitue une étape clé pour appuyer la rigueur scientifique de ce document. Elle vise non seulement à lister des références pertinentes mais surtout à articuler un dialogue critique entre les sources afin d'éclairer le choix méthodologique, la pertinence théorique et la valeur ajoutée pratique des outils bayésiens dans la modélisation macroéconomique. Cette vision bibliographique, structuré en trois axes complémentaires, permet de définir clairement la place et l'apport de l'économétrie bayésienne dans le contexte spécifique de l'analyse du PIB par habitant.

A. Théorie bayésienne générale : Fondations et débats méthodologiques

L'économétrie bayésienne repose sur des bases conceptuelles et mathématiques solides, et pour bien en saisir les enjeux, il est essentiel de se référer aux ouvrages fondamentaux de **Gary Koop (2003)** et **Tony Lancaster (2004)**, qui constituent aujourd'hui le socle théorique incontournable de cette discipline.

Gary Koop propose une analyse exhaustive et formalisée des fondements mathématiques de l'économétrie bayésienne. Il accorde une attention particulière à la spécification des distributions a priori, véritables piliers de la démarche bayésienne. En effet, ces distributions permettent d'intégrer dans l'estimation non seulement les données observées, mais aussi toute connaissance externe ou information préalable, ce qui représente un avantage considérable par rapport aux méthodes fréquentistes. Koop développe également des méthodes d'estimation adaptées à différents types de modèles, des modèles linéaires classiques aux modèles plus complexes tels que les VAR ou les modèles hiérarchiques, qui sont particulièrement pertinents dans le cadre macroéconomique. Son travail met en lumière les avantages des méthodes bayésiennes, notamment la formalisation explicite de l'incertitude à travers la distribution a posteriori, ce qui permet une inférence probabiliste complète et souvent plus réaliste dans un

contexte économique marqué par le bruit et la complexité des données. Toutefois, la rigueur et la densité de son style exigent une solide formation en statistique avancée et en mathématiques, ce qui fait de son ouvrage une référence exigeante mais précieuse pour bâtir une base théorique robuste. De plus, il souligne la nécessité d'une grande prudence dans la spécification des priors, car un mauvais choix peut biaiser les résultats.

En revanche, Tony Lancaster adopte une approche plus progressive et pédagogique, visant à rendre l'économétrie bayésienne accessible à un public plus large, notamment aux économistes peu familiers avec les statistiques avancées. Son ouvrage établit clairement les liens entre l'économétrie classique et bayésienne, ce qui facilite la compréhension des apports spécifiques de cette dernière, notamment dans les situations où les données sont limitées, bruitées ou présentent des dépendances complexes, telles que celles que l'on rencontre fréquemment dans les séries temporelles macroéconomiques comme le PIB par habitant. Lancaster met en avant la stabilité et la fiabilité accrues des estimations bayésiennes dans ces contextes, ainsi que la richesse des intervalles de crédibilité qui fournissent une information plus complète que les intervalles de confiance classiques. Sa démarche pédagogique en fait une ressource précieuse pour justifier, dès la première partie d'un mémoire, l'adoption d'une approche bayésienne.

Les travaux de Koop et Lancaster convergent vers une idée centrale : l'économétrie bayésienne enrichit considérablement la modélisation macroéconomique en proposant un cadre cohérent et flexible pour traiter l'incertitude et la complexité des données, particulièrement dans des environnements dynamiques et multidimensionnels. Cela est particulièrement pertinent pour l'étude du PIB par habitant, où la variabilité et les chocs exogènes rendent souvent insuffisantes les méthodes classiques. Cependant, ils insistent également sur les défis méthodologiques inhérents à cette approche. En particulier, la spécification des distributions a priori requiert une justification rigoureuse et des analyses de sensibilité afin d'éviter les biais. De plus, la validation des modèles bayésiens doit s'appuyer sur des diagnostics adaptés, tels que les tests de convergence des algorithmes MCMC, et l'évaluation comparative des modèles via des critères d'information bayésiens comme le DIC ou le WAIC. Enfin, les exigences computationnelles peuvent être élevées, nécessitant une maîtrise des outils numériques et des algorithmes sophistiqués.

En somme, ces deux ouvrages fournissent un cadre à la fois théorique et pédagogique qui permet de comprendre les forces de l'économétrie bayésienne tout en soulignant la nécessité d'une démarche rigoureuse et critique. Ces éléments constituent la base indispensable à la première partie de ton mémoire, où tu expliqueras en détail la pertinence et les enjeux méthodologiques de l'approche bayésienne appliquée à l'analyse du PIB par habitant.

Avec plaisir ! Voici un développement de la suite, mettant en avant des exemples pratiques et des applications concrètes de l'économétrie bayésienne, en particulier appliquée à la modélisation macroéconomique telle que le PIB par habitant.

B. Exemples pratiques et applications concrètes de l'économétrie bayésienne

L'économétrie bayésienne, grâce à sa flexibilité et à sa capacité à formaliser l'incertitude, trouve aujourd'hui de nombreuses applications concrètes en macroéconomie, notamment dans la modélisation du PIB par habitant. Plusieurs études récentes illustrent comment cette approche enrichit l'analyse économique en permettant une meilleure prise en compte des contraintes des données réelles et en fournissant des résultats plus robustes et interprétables.

1. Modélisation des séries temporelles macroéconomiques avec modèles VAR bayésiens

Les modèles vectoriels autorégressifs (VAR) sont largement utilisés pour étudier les dynamiques économiques multivariées, telles que l'évolution conjointe du PIB, de l'inflation et du chômage. Dans un contexte fréquentiste classique, les VAR sont parfois fragiles en raison du grand nombre de paramètres à estimer par rapport à la taille des échantillons disponibles, ce qui conduit à des problèmes de surparamétrisation et d'instabilité des estimations.

L'approche bayésienne permet d'introduire des distributions a priori qui régularisent les estimations, réduisant ainsi la variance des paramètres tout en maintenant un faible biais. Par exemple, la spécification d'une prior conjugée de type Minnesota (encourageant la stationnarité et la décroissance des coefficients lointains) s'est révélée efficace dans l'estimation des VAR bayésiens appliqués au PIB par habitant et à d'autres variables macroéconomiques. Cette méthode améliore notamment la qualité des prévisions économiques et la robustesse des analyses de chocs structurels, comme le montrent les travaux de Giordani et Kohn (2014).

2. Modèles hiérarchiques bayésiens pour la comparaison des performances économiques régionales

Une autre application importante concerne la modélisation hiérarchique, où les données sont organisées en groupes imbriqués, par exemple des observations du PIB par habitant à plusieurs niveaux territoriaux (pays, régions, départements). L'économétrie bayésienne hiérarchique permet d'estimer simultanément les effets globaux et spécifiques à chaque groupe, en empruntant de la force statistique entre les unités et en réduisant ainsi l'impact des petits échantillons régionaux.

Cette approche est particulièrement pertinente pour évaluer les disparités économiques régionales, identifier les déterminants locaux de la croissance, ou encore mesurer les effets des politiques publiques territoriales. En intégrant des priors flexibles et en utilisant des algorithmes MCMC, les chercheurs peuvent obtenir des distributions a posteriori précises sur les paramètres régionaux, ce qui facilite l'analyse des incertitudes propres à chaque région. Par exemple, Rossi et al. (2005) ont démontré la puissance des modèles bayésiens dans des contextes économiques variés, dont l'analyse microéconomique et macroéconomique.

3. Intégration de variables explicatives nombreuses et bruitées via des modèles bayésiens avec sélection de variables

En macroéconomie, les bases de données sont souvent riches en variables explicatives, dont beaucoup peuvent être colinéaires ou peu informatives. L'économétrie bayésienne, via des

techniques de sélection de variables et de régularisation bayésiennes (comme le Lasso bayésien ou les modèles spike-and-slab), permet de gérer efficacement ces situations en identifiant automatiquement les variables les plus pertinentes tout en contrôlant le surapprentissage.

Ces méthodes facilitent l'incorporation de nombreuses données économiques, sociales et environnementales pour expliquer le PIB par habitant, tout en offrant des mesures probabilistes des contributions respectives des variables. L'implémentation avec des outils comme PyMC en Python facilite ces analyses complexes grâce à des algorithmes efficaces et une interface flexible.

4. Prévisions économiques et prise en compte des incertitudes

L'un des apports majeurs de l'économétrie bayésienne est sa capacité à produire des prévisions économiques qui intègrent pleinement l'incertitude des paramètres et des données. Contrairement aux méthodes classiques qui fournissent souvent des intervalles de confiance asymptotiques parfois trompeurs, les intervalles de crédibilité bayésiens donnent une meilleure appréciation des risques associés aux prévisions.

Par exemple, dans l'estimation de la croissance future du PIB par habitant, les modèles bayésiens peuvent intégrer des scénarios alternatifs via des priors différents, des chocs exogènes ou des changements structurels, offrant ainsi aux décideurs une vision plus complète des possibles évolutions économiques et de leur probabilité.

Ces exemples démontrent clairement que l'économétrie bayésienne n'est pas seulement une alternative théorique aux méthodes classiques, mais un outil puissant pour l'analyse empirique du PIB par habitant et d'autres variables macroéconomiques. Grâce à ses propriétés uniques, elle permet d'adresser des problématiques complexes telles que les petits échantillons, les nombreuses variables explicatives, la hiérarchie des données ou l'incertitude des prévisions avec une rigueur méthodologique et une souplesse inégalées.

C. Méthodes computationnelles avec Python : Des algorithmes au déploiement pratique

L'avènement de l'économétrie bayésienne s'est trouvé propulsé ces dernières années par le développement d'outils open source performants et accessibles, particulièrement dans le langage Python. Cette évolution a permis non seulement de démocratiser les méthodes bayésiennes, longtemps réservées à des environnements spécialisés ou des chercheurs disposant de ressources informatiques conséquentes, mais aussi de les intégrer efficacement dans des projets de recherche appliquée. Python, avec ses bibliothèques dédiées, offre aujourd'hui une plateforme complète pour modéliser, estimer, valider, et visualiser des modèles économétriques complexes. Deux ouvrages majeurs illustrent cette dynamique : *Bayesian Analysis with Python* d'Osvaldo Martin et *Doing Bayesian Data Analysis* de John Kruschke.

Dans son ouvrage, Osvaldo Martin présente un manuel particulièrement pragmatique, centré sur la bibliothèque PyMC, qui s'est imposée comme un standard de l'estimation bayésienne via Monte Carlo par chaînes de Markov (MCMC). PyMC permet d'implémenter des modèles

bayésiens allant des plus simples aux plus sophistiqués, incluant des modèles hiérarchiques, non linéaires, ou dynamiques. Par exemple, dans le cadre de la modélisation du PIB par habitant, PyMC facilite la construction d'un modèle VAR bayésien intégrant plusieurs variables macroéconomiques interdépendantes, où les distributions a priori peuvent refléter des connaissances économiques ou des résultats antérieurs. Les algorithmes avancés comme Hamiltonian Monte Carlo (HMC) ou No-U-Turn Sampler (NUTS) permettent d'explorer efficacement des espaces de paramètres multidimensionnels, garantissant une meilleure convergence et une réduction du temps de calcul comparé aux méthodes classiques Gibbs Sampling ou Metropolis-Hastings. Cette rapidité et cette précision sont indispensables pour manipuler des bases de données volumineuses ou des modèles complexes, ce qui est courant dans l'analyse macroéconomique.

L'approche de Martin ne se limite pas à la simple mise en œuvre : il insiste également sur l'importance des diagnostics de convergence (par exemple, la statistique R-hat, les tracés de chaînes, ou l'analyse des autocorrélations) et sur les techniques d'optimisation pour améliorer la qualité des inférences. Ces aspects sont essentiels pour garantir la fiabilité des résultats et la robustesse des conclusions tirées du modèle. De plus, l'accent mis sur la reproductibilité des analyses, à travers une organisation modulaire et documentée des scripts Python, s'inscrit dans une démarche scientifique rigoureuse qui valorise la transparence et la collaboration.

En parallèle, l'ouvrage de John Kruschke, bien que s'appuyant principalement sur R, offre une pédagogie exemplaire pour aborder l'économétrie bayésienne, notamment pour les non-spécialistes. Il vulgarise les notions techniques complexes et propose une multitude d'exemples illustratifs, allant de modèles linéaires simples à des structures hiérarchiques, en passant par des tests d'hypothèses bayésiens. Son approche très visuelle et interactive facilite la compréhension de phénomènes tels que la distribution a posteriori, l'importance des priors, et le comportement des chaînes MCMC. Ces visualisations, facilement transposables dans Python grâce à des outils comme ArviZ, permettent d'enrichir la communication des résultats dans ton mémoire, rendant les conclusions accessibles à un public plus large, y compris aux décideurs économiques qui peuvent ne pas être familiers avec les techniques statistiques avancées.

Concrètement, la combinaison de ces deux références permet d'aborder tous les aspects clés du déploiement pratique de l'économétrie bayésienne en Python, de la définition du modèle à l'interprétation des résultats. Par exemple, pour un modèle bayésien appliqué au PIB par habitant, on peut partir d'une spécification simple, modélisant le PIB en fonction de variables explicatives comme l'investissement, la consommation, et les exportations. Grâce à PyMC, on construit ensuite une chaîne d'échantillonnage qui permet de générer une distribution a posteriori des coefficients, tout en contrôlant la qualité de la convergence. On peut alors produire des intervalles crédibles pour les estimations, visualiser les relations entre variables, et simuler des scénarios économiques alternatifs en modifiant les priors ou en intégrant des effets hiérarchiques (par exemple, différenciant les régions géographiques ou les périodes temporelles).

Enfin, cette maîtrise des méthodes computationnelles ouvre aussi la porte à des intégrations avancées avec des outils d'intelligence artificielle et d'apprentissage automatique, où les modèles bayésiens peuvent être enrichis ou utilisés comme composantes probabilistes dans des

architectures plus larges. Cette dimension pratique et évolutive confère à ton mémoire une double valeur scientifique et technique, en phase avec les tendances actuelles de la recherche en économétrie et en data science.

D. Applications économiques et articles récents : Concrétisations et innovations

L'application des méthodes bayésiennes dans l'analyse économique, en particulier pour des indicateurs macroéconomiques comme le PIB par habitant, s'est considérablement développée ces dernières années. Cette évolution repose sur la reconnaissance de la capacité des approches bayésiennes à gérer des données complexes, souvent imparfaites, et à intégrer de manière formelle des connaissances a priori issues de la théorie économique ou d'observations historiques.

Dans ce contexte, l'ouvrage collectif de Giordani, Kohn et Strachan (2014) constitue une référence majeure. Les auteurs y explorent une large palette de modèles bayésiens dynamiques et hiérarchiques, spécialement adaptés à la macroéconomie. Par exemple, ils présentent des modèles à vecteurs autorégressifs bayésiens (BVAR) qui permettent d'améliorer la précision des prévisions du PIB en intégrant des distributions a priori sur les paramètres, souvent basées sur des études économiques antérieures ou des données historiques longues. Ces modèles hiérarchiques sont particulièrement utiles pour modéliser la variabilité régionale ou sectorielle du PIB par habitant, en tenant compte à la fois des caractéristiques communes à l'économie nationale et des spécificités locales.

Une application pratique illustrée dans ce travail est la prévision des cycles économiques : les méthodes bayésiennes permettent de mieux modéliser l'incertitude liée aux chocs exogènes, tels que les variations des prix du pétrole ou les changements dans la politique fiscale, qui influencent fortement la croissance économique. Ces modèles, grâce à leur nature probabiliste, fournissent non seulement une estimation ponctuelle du PIB mais aussi une distribution complète des valeurs possibles, permettant aux décideurs politiques d'évaluer les risques et les probabilités associées à différents scénarios économiques. Par exemple, dans la gestion de la politique monétaire, disposer d'une distribution a posteriori précise du PIB futur aide à calibrer les taux d'intérêt avec plus de confiance (Giordani, Kohn, & Strachan, 2014).

Par ailleurs, l'étude de Rossi, Allenby et McCulloch (2005) illustre la puissance des méthodes bayésiennes dans des domaines apparemment éloignés, comme le marketing, mais dont les techniques sont transférables à l'économie. Leur travail démontre que l'approche bayésienne gère efficacement les données limitées, hétérogènes et bruitées, typiques de nombreuses situations empiriques. Par exemple, dans le cadre d'études sur le comportement des consommateurs, les auteurs utilisent des modèles bayésiens pour capturer les préférences non observables et la variabilité individuelle, avec une grande robustesse face aux échantillons réduits.

Cette robustesse est également un atout essentiel en macroéconomie, où les données sur le PIB par habitant peuvent souffrir d'erreurs de mesure, de révisions statistiques ou de périodes manquantes. L'approche bayésienne permet ainsi d'intégrer des informations supplémentaires,

comme des données provenant d'enquêtes ou des prévisions de tiers, via les distributions a priori, pour compenser ces lacunes. Par exemple, dans le cadre de l'évaluation des politiques publiques, un modèle bayésien pourrait intégrer les résultats d'enquêtes de terrain avec les données officielles pour obtenir une estimation plus fiable de l'impact sur le PIB régional (Rossi, Allenby, & McCulloch, 2005).

De plus, ces études soulignent l'intérêt des méthodes bayésiennes pour modéliser la non-linéarité et les interactions complexes entre variables économiques. Dans le cas du PIB, des relations non linéaires existent souvent entre la croissance économique, les investissements publics, la consommation, et les facteurs externes comme les crises financières ou les fluctuations du commerce international. Les modèles bayésiens permettent de capturer ces effets d'une manière plus flexible qu'avec des méthodes fréquentistes classiques, ouvrant la voie à une meilleure compréhension des mécanismes sous-jacents à la croissance économique.

Enfin, ces références mettent en avant les défis méthodologiques et les bonnes pratiques, notamment la nécessité d'une spécification rigoureuse des distributions a priori et l'importance des diagnostics de convergence des chaînes MCMC pour assurer la validité des estimations. Ces points sont cruciaux pour garantir la robustesse et la crédibilité des résultats, et méritent une attention particulière dans ton mémoire, où la modélisation du PIB par habitant peut impliquer des structures complexes et des données souvent difficiles à exploiter.

E. Articles récents sur l'IA et Python en économétrie bayésienne : Synergies et perspectives

L'intégration croissante de l'intelligence artificielle (IA) avec l'économétrie bayésienne ouvre de nouvelles perspectives méthodologiques et applicatives pour la modélisation économique, en particulier dans un contexte de données massives, dynamiques et hétérogènes. Ce rapprochement entre deux champs historiquement distincts repose sur une volonté commune : traiter rigoureusement l'incertitude tout en exploitant la puissance des calculs modernes et des architectures d'apprentissage automatique.

Les méthodes **Empirical Bayes**, longtemps considérées comme un pont entre fréquentisme et Bayésianisme, occupent une place centrale dans cette convergence. Les travaux pionniers de **Casella (1985)** ont jeté les bases d'une inférence à la fois robuste et empirique, capable d'estimer les distributions a priori à partir des données observées. Cette approche a été considérablement enrichie par les contributions d'**Efron (2010)**, qui a montré que l'Empirical Bayes peut s'adapter à des cadres de haute dimension, ce qui le rend particulièrement pertinent pour l'analyse macroéconomique contemporaine où le nombre de covariables (démographiques, institutionnelles, environnementales) augmente fortement. Dans le cas de la modélisation du PIB par habitant, ces méthodes permettent d'inclure des informations historiques ou contextuelles (ex. cycles économiques, chocs exogènes) tout en ajustant dynamiquement les priors aux nouvelles données.

La synergie entre économétrie bayésienne et **machine learning** est particulièrement féconde lorsqu'il s'agit de capter des effets non linéaires, des interactions complexes ou des hétérogénéités structurelles dans les données. L'article de **Hahn, Carvalho et Lee (2019)** met en lumière l'émergence de modèles bayésiens structurels capables d'expliquer des comportements hétérogènes dans les données économiques, notamment via des distributions postérieures hiérarchiques ou des arbres décisionnels bayésiens. Cette approche permet d'aller au-delà des modèles linéaires classiques en proposant des représentations flexibles, utiles pour des variables comme le PIB, souvent influencées par des régimes économiques distincts (croissance, récession, stagnation).

L'apport des outils **Python** à cette révolution méthodologique est fondamental. **PyMC3** (Salvatier, Wiecki & Fonnesbeck, 2016) a démocratisé l'accès à des techniques d'estimation complexes (telles que Hamiltonian Monte Carlo ou le No-U-Turn Sampler), permettant d'estimer des modèles bayésiens même en présence d'un grand nombre de paramètres ou de structures latentes. La visualisation des résultats est facilitée par la bibliothèque **ArviZ** (Kumar et al., 2021), qui propose des outils performants pour analyser les chaînes MCMC, comparer les modèles via des critères bayésiens (WAIC, LOO-CV) et interpréter les résultats de manière graphique. Ces outils, désormais standards en économétrie bayésienne computationnelle, permettent une exploration efficace et transparente des incertitudes modélisées.

Sur le plan théorique, **Murphy (2012)** et **Pearl (1988)** rappellent l'importance du **raisonnement probabiliste** comme fondement de l'intelligence artificielle. L'approche bayésienne permet de représenter formellement l'incertitude, d'apprendre de nouvelles connaissances à partir de données et de prendre des décisions rationnelles fondées sur des distributions de croyances. Ces principes sont directement applicables en économie, notamment dans la modélisation des comportements d'agents, les politiques publiques adaptatives, ou l'évaluation de scénarios macroéconomiques incertains.

Enfin, les contraintes computationnelles liées à l'inférence bayésienne ont stimulé l'émergence d'approches alternatives aux méthodes MCMC. **Blei, Kucukelbir et McAuliffe (2017)** ont synthétisé les avancées en **inférence variationnelle**, qui offrent une approximation rapide des distributions postérieures en remplaçant l'échantillonnage stochastique par une optimisation déterministe. Bien que légèrement moins précise, cette approche permet de traiter des bases de données économiques volumineuses avec des modèles complexes en un temps raisonnable, ce qui constitue un atout précieux dans un contexte d'analyse empirique à grande échelle.

Ce corpus théorique et appliqué, à l'intersection de l'économétrie, de l'IA et de l'analyse bayésienne, constitue le socle de la démarche adoptée dans ce mémoire. Il justifie l'usage d'outils comme **PyMC** et **ArviZ** pour implémenter des modèles économétriques souples, interprétables et calibrés sur données réelles, tout en assurant une rigueur inférentielle adaptée aux incertitudes inhérentes aux phénomènes économiques étudiés.

F. Ressources numériques complémentaires : un levier pour une mise en œuvre rigoureuse

Dans un contexte où la complexité des modèles statistiques croît avec celle des données économiques, le recours à des ressources numériques actualisées, maintenues par des communautés actives, constitue un levier essentiel pour la rigueur, la reproductibilité et l'innovation méthodologique en économétrie bayésienne. Ce type de ressources offre un triple avantage : une mise en œuvre pratique facilitée, un accès aux meilleures pratiques de modélisation, et un renforcement de la transparence scientifique.

PyMC : un socle moderne pour la modélisation bayésienne

La documentation officielle de PyMC (<https://www.pymc.io/>) est aujourd’hui l’un des piliers de la modélisation probabiliste en Python. Ce site rassemble des informations exhaustives sur l’architecture du langage, les distributions disponibles, les algorithmes d’échantillonnage de type MCMC (notamment le No-U-Turn Sampler – NUTS), ainsi que sur les techniques d’inférence variationnelle utilisées pour accélérer la convergence. Elle inclut également de nombreux **tutoriels interactifs** et des cas d’étude adaptés à des situations réelles, facilitant l’apprentissage progressif et l’autonomie dans la construction de modèles économétriques complexes.

L’utilisation de cette documentation dans le cadre du mémoire ne se limite pas à une fonction d’apprentissage ; elle garantit également une **mise en œuvre rigoureusement alignée avec les standards de la recherche scientifique contemporaine**, en évitant les erreurs d’implémentation ou les approximations méthodologiques.

ArviZ : l’outil fondamental pour l’évaluation et la comparaison des modèles

Complément naturel de PyMC, ArviZ (Kumar et al., 2021) s’est imposé comme l’outil de référence pour l'**analyse et la visualisation des résultats bayésiens**. Il permet, grâce à des fonctions comme `az.plot_trace()` ou `az.plot_posterior()`, d’évaluer de manière visuelle et statistique la convergence des chaînes de Markov, de détecter d’éventuels problèmes d’autocorrélation, et de résumer les distributions postérieures des paramètres. Des métriques avancées, telles que la statistique **R-hat**, le **nombre effectif d’échantillons**, ou encore les critères d’information **WAIC** et **LOO-CV**, permettent une **évaluation rigoureuse de la qualité des inférences et la comparaison objective de plusieurs modèles**.

Dans le cadre de ce mémoire, l’usage d’ArviZ permettra d’éviter des conclusions hâtives fondées sur des modèles instables ou mal identifiés, en renforçant la fiabilité des résultats et leur pertinence économique.

Les notebooks de McElreath : une ressource pédagogique et appliquée

Enfin, les **notebooks interactifs de Richard McElreath**, issus de son ouvrage *Statistical Rethinking*, constituent un **instrument d’apprentissage pratique et progressif**. Leur adaptation à Python via PyMC (projet “Rethinking with PyMC”) permet de retrouver, dans un environnement code-first, les concepts théoriques essentiels de l’inférence bayésienne (modèles hiérarchiques, shrinkage, effets aléatoires, etc.). Chaque notebook est structuré de

manière à présenter un problème, proposer une modélisation, l'estimer, puis interpréter les résultats – une **logique parfaitement adaptée au travail empirique du mémoire**.

Ces notebooks permettent de **tester des idées, d'expérimenter différents modèles, et de simuler des données** pour mieux comprendre leur comportement. Ce type de ressource offre donc une **intégration idéale entre la théorie, la pratique, et la pédagogie**, rendant l'apprentissage autonome plus robuste, et facilitant la transposition à des données macroéconomiques réelles.

L'usage articulé de **PyMC pour la modélisation, ArviZ pour l'analyse diagnostique, et des notebooks pédagogiques pour l'apprentissage appliqué** permet une mise en œuvre transparente, reproductible et scientifiquement solide de l'économétrie bayésienne.

Synthèse critique approfondie

L'ensemble de la bibliographie mobilisée révèle une **convergence forte entre fondements théoriques solides, innovations techniques récentes, et applications empiriques économiquement significatives**. L'économétrie bayésienne, bien loin de constituer une alternative marginale aux approches fréquentistes classiques, se présente aujourd'hui comme une **méthode centrale pour traiter les incertitudes structurelles** inhérentes à l'analyse des phénomènes économiques contemporains, en particulier lorsqu'il s'agit de modéliser des indicateurs aussi volatils et multidimensionnels que le **PIB par habitant**.

À la faveur des avancées informatiques, cette approche connaît une diffusion plus large grâce à des **outils open source puissants comme PyMC, ArviZ ou NumPyro**, qui facilitent à la fois l'estimation, la visualisation et le diagnostic des modèles. Ces outils permettent une montée en complexité sans perte de lisibilité, tout en s'ancrant dans un écosystème Python hautement interopérable. En cela, l'**économétrie bayésienne devient progressivement accessible à un public non spécialiste**, ce qui en accroît la portée dans les sphères de la recherche appliquée, de la décision publique et de la prospective économique.

Toutefois, **plusieurs limites structurelles et méthodologiques sont mises en évidence dans la littérature :**

- **La spécification des distributions a priori** demeure un point épineux : entre subjectivité potentielle, contraintes computationnelles et arbitrage entre informativité et flexibilité, le choix des priors influence fortement les résultats, en particulier sur des échantillons de petite taille (Efron, 2010 ; Casella, 1985).
- **Le diagnostic de convergence et la validation des modèles** restent techniquement exigeants, en dépit des apports d'ArviZ. Les chaînes MCMC peuvent diverger, mal explorer l'espace postérieur ou converger vers des modes locaux. Des indicateurs comme R-hat ou WAIC nécessitent une **interprétation experte**, qui suppose un bagage statistique conséquent (Hoffman & Gelman, 2014).
- **La montée en charge computationnelle** pose problème dès que les modèles incluent des effets hiérarchiques profonds, des interactions non linéaires ou des distributions

complexes. L'inférence variationnelle (Blei et al., 2017) propose des alternatives rapides mais parfois biaisées, impliquant un **compromis entre précision et rapidité**.

- Enfin, la **barrière technique et cognitive à l'entrée** est encore importante : manipuler correctement PyMC ou Stan nécessite une double compétence en programmation et en statistique bayésienne, souvent absente dans les formations économiques traditionnelles.

Ces tensions, loin d'invalider l'approche, **soulignent la nécessité d'un positionnement épistémologique clair**, mais aussi d'une **démarche méthodologique progressive, transparente et rigoureuse**. Pour ce mémoire, elles nourriront non seulement les parties théoriques et techniques, mais aussi la discussion critique autour de la robustesse des résultats, de la validité externe des modèles, et de l'interprétabilité économique des inférences produites.

En somme, la littérature analysée légitime pleinement l'usage de l'économétrie bayésienne assistée par Python dans le traitement des données économiques incertaines, à condition de **rester conscient des exigences méthodologiques** que cette rigueur implique. Cette approche offre une voie féconde pour renouveler les pratiques économétriques et proposer des lectures plus nuancées et probabilistes de la réalité économique.

Partie 2 – Données, variables et préparation

2.1 – Présentation des données : Source, nature et structure des données

Les données mobilisées dans ce mémoire proviennent principalement de la base de données en libre accès **World Development Indicators (WDI)**, mise à disposition par la **Banque mondiale**. Cette source est internationalement reconnue pour la qualité, la cohérence et la couverture étendue de ses indicateurs macroéconomiques, sociaux et environnementaux. Elle constitue une référence dans la littérature économétrique appliquée (Easterly & Levine, 2001 ; Barro, 2000), en particulier lorsqu'il s'agit de modéliser les déterminants du développement économique.

Les données utilisées sont de nature quantitative et longitudinale (panel data), organisées par pays et par année. La fréquence est **annuelle**, ce qui est adapté pour capter les dynamiques économiques structurelles plutôt que les chocs conjoncturels de court terme. La période retenue s'étend de **2000 à 2023**, soit une fenêtre temporelle de **24 années**, permettant de couvrir à la fois des phases de croissance soutenue, de crise mondiale (2008-2009), et de reprise post-COVID.

Le volume total de données dépend du sous-échantillon retenu (en fonction de la disponibilité des indicateurs pour chaque pays), mais inclut 30 pays de différentes zones géographiques et niveaux de développement, assurant une diversité contextuelle utile à l'analyse.

Justification du choix des données

Trois critères principaux ont été retenus pour le choix des données :

1. **Pertinence économique** : Les variables sélectionnées telles que le PIB/habitant, les dépenses publiques en pourcentage du PIB, les dépenses d'éducation, l'espérance de scolarisation, le taux de chômage, et les investissements directs étrangers sont directement liées aux hypothèses théoriques sur les déterminants de la croissance et du bien-être économique. Ces variables sont régulièrement mobilisées dans les modèles de croissance endogène et dans les travaux de développement économique.
2. **Qualité statistique et fiabilité** : La Banque mondiale suit des protocoles rigoureux de collecte, d'harmonisation et de mise à jour des données. Le choix de cette base garantit donc fiabilité, comparabilité internationale et cohérence temporelle, conditions indispensables pour une modélisation économétrique robuste.
3. **Compatibilité avec la modélisation bayésienne** : La structure panel des données permet de capter les **hétérogénéités spécifiques aux pays**, ce qui est particulièrement utile dans un cadre bayésien où il est possible d'incorporer des effets hiérarchiques (random effects) ou des incertitudes sur les coefficients. De plus, les modèles bayésiens s'acquittent bien des manques ponctuels de données, grâce aux techniques d'échantillonnage et d'imputation probabiliste, rendant le traitement plus flexible que dans un cadre fréquentiste classique.

2.2 – Variables

Notre but est d'expliquer les variations du PIB par habitant (en \$ courant) à partir de facteurs macroéconomiques structurels. Les données proviennent de la Banque mondiale (section 2.1), couvrant 2000–2023 pour plusieurs pays de différentes régions économiques (Europe, Amérique, Asie, Afrique).

2.2.1 – Variable dépendante

- **log_gdp** : logarithme naturel du PIB par habitant (gdp_pc)
 - Code WDI : NY.GDP.PCAP.CD
 - Description : PIB total divisé par la population, en dollars US courants.
 - Transformation : log naturelle pour réduire l'asymétrie et faciliter l'interprétation multiplicative.

2.2.2 – Variables explicatives principales

- **gov_spend** : dépenses publiques totales (% PIB)
 - Code WDI : GC.XPN.TOTL.GD.ZS
 - Hypothèse : effet positif modéré via demande/investissement public.
- **edu_spend** : dépenses en éducation (% PIB)
 - Code WDI : SE.XPD.TOTL.GD.ZS
 - Justification : capital humain, effet positif à moyen terme.
- **health_spend** : dépenses en santé (% PIB)
 - Code WDI : SH.XPD.CHEX.GD.ZS
 - Hypothèse : effet positif indirect via bien-être et productivité.
- **unemp** : taux de chômage (%)
 - Code WDI : SL.UEM.TOTL.ZS
 - Hypothèse : effet négatif, reflète sous-emploi des ressources.

2.2.3 – Variables dérivées

- **log_gdp** : variable dépendante transformée en log.
- **edu_health** : interaction $\text{edu_spend} \times \text{health_spend}$, pour étudier l'effet combiné du capital humain.

2.2.4 – Justification économétrique

- Variables continues adaptées à la régression bayésienne.
- Transformations (log, interaction) améliorent linéarité et interprétabilité.
- Surveillance de la multicolinéarité (via ACP) et des corrélations structurelles.

Voici un résumé synthétique de correspondance entre les noms longs et les noms courts des variables :

Variable longue	Nom court	Description courte
PIB par habitant (gdp_per_capita)	log_gdp	Logarithme naturel du PIB par habitant
Dépenses publiques totales (% du PIB)	gov_spend	Dépenses publiques en % du PIB
Dépenses en éducation (% du PIB)	edu_spend	Dépenses en éducation en % du PIB
Dépenses en santé (% du PIB)	health_spend	Dépenses en santé en % du PIB
Taux de chômage (%)	unemp	Taux de chômage
Interaction éducation × santé	edu_health	Interaction entre dépenses éducation et santé

2.3 – Analyse descriptive

L’analyse descriptive des données permet d’avoir un premier aperçu des distributions des variables sélectionnées et des relations qu’elles entretiennent entre elles. Cette section présente successivement les statistiques univariées, puis les statistiques bivariées issues de la matrice de corrélation.

Statistiques univariées

Statistiques	log_gdp	gov_spend	edu_spend	health_spend	unemp	edu_health
count	2693	2693	2693	2693	2693	2693
mean	8,868579	26,777822	4,557604	6,857405	7,82935	32,903364
std	1,448441	10,404819	1,480582	2,732582	5,150757	19,328199
min	4,696781	4,259324	0,737	1,701678	0,119	1,713775
25%	7,799488	17,865144	3,5393	4,723025	4,407	18,006902
50%	8,891908	26,517505	4,57465	6,617313	6,654	29,817744
75%	10,124231	34,509126	5,37523	8,777178	9,5	44,804986
max	11,803442	83,608031	13,21957	21,969868	34,153	158,052742

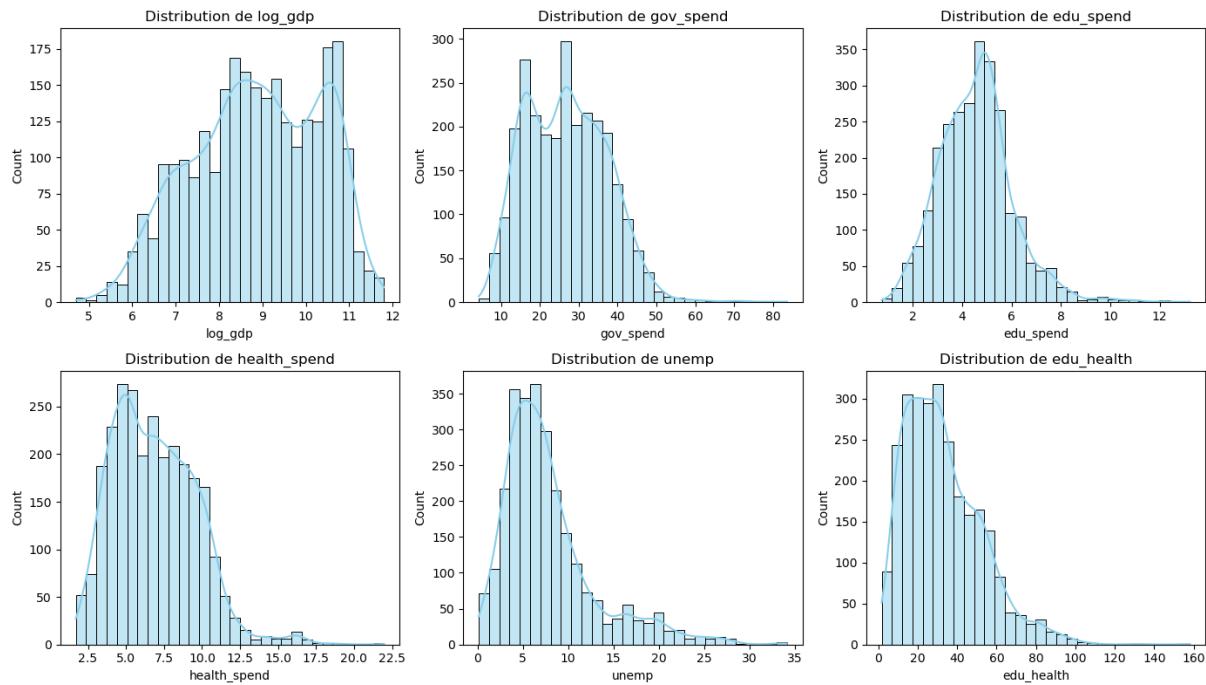
Les statistiques univariées indiquent que le **logarithme du PIB par habitant** (log_gdp) présente une moyenne de 8,87, avec une dispersion relativement modérée (écart-type de 1,45), ce qui suggère une distribution globalement équilibrée. Sa valeur minimale est de 4,70 et sa valeur maximale atteint 11,80, traduisant des différences marquées de niveau de développement économique entre les pays et les années observés.

La **dépense publique totale** (gov_spend), exprimée en pourcentage du PIB, varie entre 4,26 % et 83,61 %, avec une moyenne de 26,78 %. Cette large dispersion (écart-type de 10,40) reflète des modèles très divers de financement public, allant d'États minimalistes à des économies plus interventionnistes.

En ce qui concerne les **dépenses en éducation** (edu_spend) et **en santé** (health_spend), les moyennes sont respectivement de 4,56 % et 6,86 % du PIB. Bien que leur dispersion soit plus faible que celle des dépenses publiques totales, les écarts restent significatifs (écart-types de 1,48 et 2,73), suggérant des politiques budgétaires très variables en matière de services sociaux.

Le **taux de chômage** (unemp) présente une moyenne de 7,83 %, mais avec un écart-type élevé (5,15) et une valeur maximale de 34,15 %, ce qui révèle de fortes disparités sur le marché du travail entre les pays étudiés. Enfin, la variable **edu_health**, construite comme le produit des dépenses en éducation et en santé, affiche une moyenne de 32,90, mais s'étend de 1,71 à 158,05, ce qui confirme une hétérogénéité marquée dans l'investissement combiné en capital humain.

Visualisation des distributions de chacune des variables



Analyse exploratoire des distributions des variables explicatives

L'analyse visuelle des distributions des principales variables mobilisées dans notre étude permet de mieux comprendre la structure statistique des données et d'anticiper d'éventuels problèmes liés à l'estimation. Les histogrammes accompagnés de courbes de densité (KDE) fournissent une vision claire des asymétries, concentrations ou valeurs extrêmes susceptibles d'influencer les résultats économétriques.

1. Distribution du PIB par habitant (log_gdp)

La variable log_gdp, utilisée comme variable dépendante dans nos modèles, présente une distribution relativement symétrique et proche d'une loi normale. Cette transformation logarithmique du PIB par habitant est justifiée par la volonté de réduire la dispersion initiale de la variable brute et de linéariser les relations avec les variables explicatives. La forme de cette distribution confirme que l'utilisation de modèles linéaires est pertinente sur le plan statistique.

2. Distribution des dépenses publiques totales (gov_spend)

La variable gov_spend, qui mesure la part des dépenses publiques dans le PIB, présente une distribution asymétrique à droite. La majorité des observations se situent entre 20 % et 40 %,

mais on observe une longue queue droite avec des valeurs pouvant dépasser 60 %. Cette hétérogénéité suggère que certains pays adoptent des politiques budgétaires beaucoup plus expansives. Sur le plan économétrique, cette asymétrie pourrait nuire à la normalité des résidus dans un modèle linéaire classique, ce qui peut justifier l'usage de modèles plus souples, notamment bayésiens.

3. Distribution des dépenses d'éducation (edu_spend)

La variable `edu_spend` est relativement concentrée autour de sa moyenne, généralement entre 4 % et 6 % du PIB. Sa distribution est modérément asymétrique à droite, mais reste globalement bien centrée. Cette relative homogénéité peut permettre une estimation plus stable et fiable des coefficients associés à l'éducation dans les modèles économétriques.

4. Distribution des dépenses de santé (health_spend)

La distribution de `health_spend` présente également une asymétrie à droite, mais avec une dispersion légèrement plus forte que celle des dépenses éducatives. La majorité des pays se concentrent autour de 5 % à 10 %, avec quelques valeurs dépassant 15 %. Cette variabilité peut révéler des disparités structurelles dans les systèmes de santé selon les niveaux de développement ou les choix de financement public. Il peut être pertinent de tester des transformations de cette variable pour atténuer les effets des extrêmes.

5. Distribution du taux de chômage (unemp)

Le taux de chômage (`unemp`) est une variable marquée par une forte asymétrie à droite. De nombreux pays affichent des taux faibles (entre 3 % et 10 %), tandis que d'autres connaissent des taux très élevés, parfois supérieurs à 25 %. Cette variabilité impose une prudence dans l'interprétation des résultats, notamment en termes d'hétéroscédasticité potentielle. Des transformations ou un traitement robuste (winsorisation, rééchantillonnage) peuvent s'avérer nécessaires.

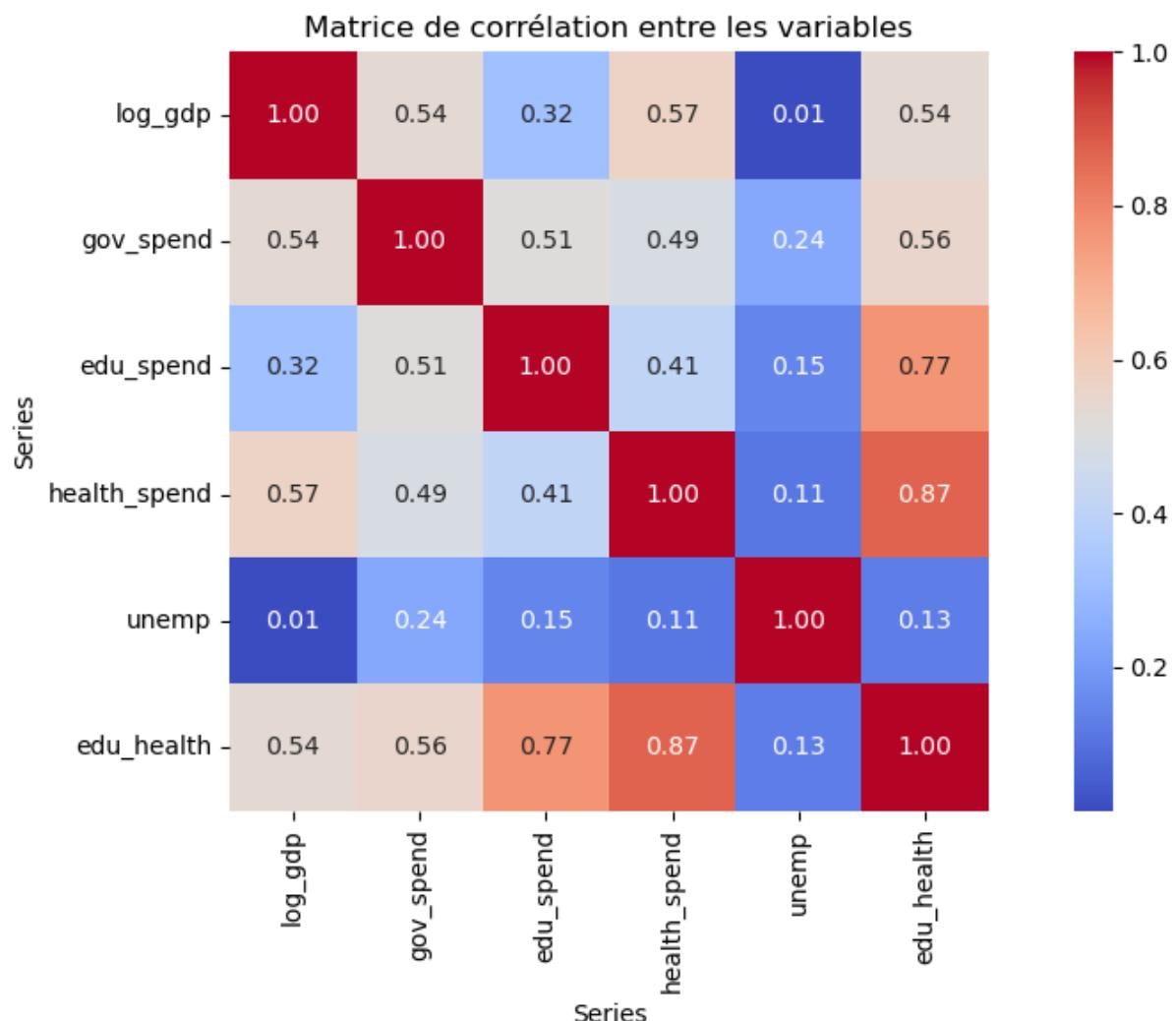
6. Distribution de la variable agrégée (edu_health)

Enfin, la variable `edu_health`, construite comme la somme des dépenses d'éducation et de santé, présente une distribution fortement asymétrique, avec une concentration autour de 15 % à 30 % mais aussi des valeurs très élevées dépassant 100 %, ce qui semble suspect. Il conviendra de vérifier la cohérence de ces observations extrêmes, car elles peuvent provenir d'erreurs de mesure ou de cas très particuliers. Dans un cadre bayésien, la gestion de telles incertitudes se fait naturellement via les distributions *a priori* et *a posteriori*, mais un nettoyage préalable reste souhaitable.

Conclusion méthodologique

Cette exploration graphique met en lumière l'intérêt d'un cadre bayésien qui permet de mieux gérer les distributions non normales et les extrêmes, tout en intégrant l'incertitude autour des paramètres estimés. Elle justifie également, pour certains modèles classiques, l'éventuel recours à des transformations ou à des techniques robustes pour préserver la qualité des estimations.

Statistiques bivariées



La matrice de corrélation met en évidence plusieurs relations intéressantes entre les variables étudiées. Tout d'abord, on observe une **corrélation positive modérée à forte** entre le **log du PIB par habitant** et les variables de dépense publique : 0,54 avec la dépense publique totale, 0,32 avec la dépense en éducation, et 0,57 avec la dépense en santé. Ces résultats suggèrent que les pays plus riches investissent davantage dans les services publics et sociaux.

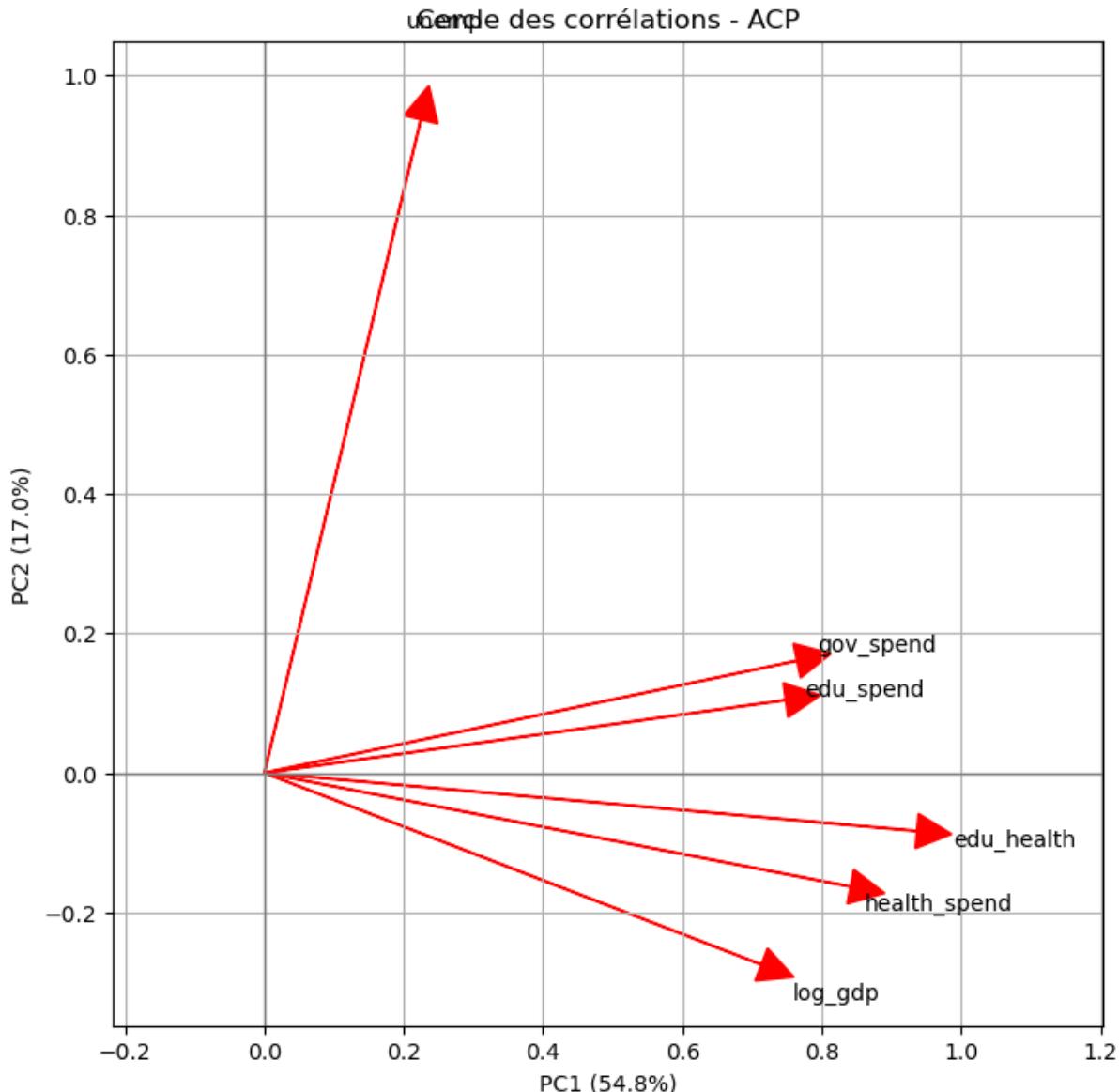
La **corrélation très forte entre les dépenses en santé et la variable synthétique edu_health** (0,87), ainsi qu'entre les dépenses en éducation et edu_health (0,77), est attendue du fait même de la construction de cette variable. Toutefois, le fait que **edu_health** corrèle aussi avec le **log du PIB** (0,54) et la **dépense publique totale** (0,56) renforce l'idée que l'investissement simultané dans la santé et l'éducation est un indicateur pertinent du niveau de développement.

Enfin, la **corrélation entre le chômage et les autres variables est globalement faible**, voire quasi inexistante avec le PIB par habitant (0,01), ce qui suggère que le taux de chômage ne suit pas systématiquement le niveau de richesse d'un pays, en tout cas pas de manière linéaire. La plus forte corrélation avec le chômage est observée avec la dépense publique (0,24), sans toutefois dépasser un niveau modéré.

Ces résultats fournissent une base solide pour les analyses économétriques ultérieures, en orientant les hypothèses de causalité possibles entre les investissements publics, le développement humain et la performance économique.

Analyse en Composantes Principales (ACP)

Analyse du cercle des corrélations



Le cercle des corrélations obtenu à partir de l'ACP permet d'analyser la contribution et la qualité de représentation des variables dans l'espace des deux premières composantes principales (PC1 et PC2), qui expliquent ensemble **71,8 % de la variance totale** (soit 54,8 % pour PC1 et 17,0 % pour PC2).

La première composante principale (PC1) semble fortement corrélée à la plupart des variables, en particulier `edu_health`, `health_spend`, `log_gdp`, `edu_spend` et `gov_spend`. Ces vecteurs sont orientés dans une direction similaire et relativement proches du cercle unitaire, ce qui indique qu'ils sont bien représentés par cette première dimension. Cela suggère que PC1 capte principalement une **dimension socio-économique et budgétaire** associée aux niveaux de dépenses publiques, en particulier dans les domaines de la santé et de l'éducation, ainsi qu'au niveau de richesse (`log PIB par habitant`).

La variable unemp, en revanche, est presque orthogonale aux autres et orientée vers PC2. Elle est bien représentée par cette deuxième composante, ce qui laisse penser que PC2 capture un **axe indépendant du niveau de dépenses ou du développement**, possiblement lié à des **tensions structurelles sur le marché du travail**, ou à des vulnérabilités économiques non directement associées aux dépenses publiques.

Par ailleurs, la faible corrélation entre unemp et les autres variables pourrait indiquer qu'elle apporte une **information complémentaire** au reste du système de variables, renforçant l'intérêt de son inclusion dans un modèle multidimensionnel.

Analyse de la variance expliquée

==== Variance expliquée par composante ====

PC1 : 0.548

PC2 : 0.170

PC3 : 0.122

PC4 : 0.098

PC5 : 0.058

PC6 : 0.004

L'analyse des valeurs propres montre que la première composante principale (PC1) explique à elle seule **54,8 % de la variance totale**, ce qui est très élevé. Elle constitue donc une synthèse très informative des données, et justifie l'interprétation approfondie des variables qui y sont fortement projetées. La deuxième composante (PC2) explique **17,0 %** de la variance, ce qui porte le cumul à **71,8 %** pour les deux premières composantes, un seuil considéré comme satisfaisant pour une représentation fidèle des données dans un plan bidimensionnel.

Les composantes suivantes apportent beaucoup moins d'information : PC3 (12,2 %) et PC4 (9,8 %) contribuent modestement, tandis que PC5 (5,8 %) et PC6 (0,4 %) sont presque négligeables. Cela confirme que **l'essentiel de la structure des données est capturé par les deux premières dimensions**, ce qui valide l'usage du graphique du cercle des corrélations comme support d'interprétation principal.

Conclusion interprétative

Cette ACP révèle une structure cohérente des relations entre variables : d'un côté, un axe socio-économique majeur (PC1) fortement corrélé au développement et à l'investissement public, et

de l'autre, un axe plus spécifique (PC2), porté par le taux de chômage. Cette lecture offre des pistes pour la construction de modèles économétriques, notamment en termes de regroupement de variables ou de détection de colinéarité. En outre, elle suggère l'intérêt potentiel d'utiliser les scores factoriels des composantes principales comme variables synthétiques dans un modèle de régression linéaire ou bayésien, afin de réduire la dimensionnalité tout en conservant l'essentiel de l'information.

2.4-Méthodologie

2.4.1. Approche générale

L'objectif de ce mémoire, rappelons-le est d'analyser les déterminants macroéconomiques du PIB par habitant dans une perspective bayésienne, en comparant cette approche avec les méthodes économétriques classiques. L'étude s'appuie sur des données issues de la base de la Banque mondiale (WDI), couvrant la période de 2000 à 2023 pour un ensemble de pays.

L'analyse repose sur la formulation de trois hypothèses de recherche, traduites en modèles économétriques testables. Ces hypothèses mobilisent différentes variables explicatives liées aux dépenses publiques, à l'emploi, et à la gouvernance économique. Les méthodes utilisées vont de la régression linéaire classique aux modèles bayésiens, avec une attention particulière portée à l'incertitude des estimations et à la performance prédictive.

2.4.2. Environnements et outils utilisés

Le langage de programmation utilisé pour le traitement des données, la modélisation et la visualisation est Python (version 3.11.9). Deux environnements de développement complémentaires ont été mobilisés :

- Jupyter Notebook : utilisé principalement pour les explorations interactives, les visualisations, les analyses descriptives, et les premiers tests de modèles.
- PyCharm : utilisé pour structurer le code de manière plus modulaire et professionnelle, notamment dans les phases de modélisation avancée et de production des graphiques finaux.

Les bibliothèques Python utilisées incluent notamment : pandas, numpy, matplotlib, seaborn, scikit-learn, pymc, bambi, et arviz.

2.4.3. Démarche économétrique

Chaque hypothèse de recherche est testée au moyen d'un modèle économétrique approprié, en cohérence avec la nature des variables impliquées et les objectifs analytiques. Le tableau suivant établit une correspondance claire entre les hypothèses, les variables mobilisées, et les méthodes économétriques utilisées.

Tableau de correspondance : hypothèses, variables et méthodes

Hypothèse	Variables concernées	Méthodes économétriques mobilisées
Hypothèse 1 : Les dépenses publiques en éducation et en santé ont un effet positif et significatif sur le PIB par habitant.	- log_gdp (log du PIB par habitant) - edu_spend (dépenses d'éducation en % du PIB) - health_spend (dépenses de santé en % du PIB) - edu_health (somme des deux)	- Régression linéaire bayésienne - Régression classique (MCO) pour comparaison - Estimation a posteriori des coefficients - Intervalles de crédibilité à 95 %
Hypothèse 2 : L'approche bayésienne permet une meilleure quantification de l'incertitude entourant les estimations que les méthodes économétriques classiques.	- Tous les paramètres estimés - log_gdp (variable dépendante)	- Comparaison entre intervalles de crédibilité bayésiens et intervalles de confiance MCO - Analyse des distributions a posteriori - Analyse de la sensibilité aux a priori
Hypothèse 3 : Un modèle bayésien intégrant les principales variables explicatives offre de meilleures performances prédictives du PIB par habitant qu'un modèle fréquentiste équivalent.	- log_gdp (cible) - gov_spend, edu_spend, health_spend, unemp, edu_health (prédicteurs)	- Régression bayésienne avec validation croisée - Régression OLS avec validation croisée - Évaluation par RMSE, MAE, log-vraisemblance prédictive, score de couverture

2.4. Implémentation des modèles

Les modèles sont implémentés en Python à l'aide de la bibliothèque PyMC (pour les estimations bayésiennes), et scikit-learn (pour les comparaisons fréquentistes). L'estimation bayésienne des coefficients est réalisée via la méthode MCMC (échantillonnage de type NUTS), et les résultats sont analysés à l'aide de visualisations des distributions a posteriori (arviz).

Une validation croisée (k-fold) est également mise en œuvre afin d'évaluer la performance prédictive des modèles. Les métriques comparatives incluent la racine de l'erreur quadratique moyenne (RMSE), l'erreur absolue moyenne (MAE), ainsi que la log-vraisemblance prédictive.

Partie 3 – Modélisation, estimation et évaluation

3.1 – Spécification du modèle bayésien

Choix du modèle : Régression linéaire bayésienne

Pour analyser les déterminants macroéconomiques du PIB par habitant, nous choisissons une régression linéaire bayésienne. Ce choix est motivé par la nature continue de la variable dépendante (le logarithme du PIB par habitant) et par la volonté de quantifier l'incertitude entourant les paramètres de la régression grâce à des distributions a posteriori.

La forme générale du modèle est la suivante :

$$\log(\text{PIB}_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

où :

- $\log(\text{PIB}_i)$ est le logarithme du PIB par habitant pour le pays i ,
- X_{ji} représente les variables explicatives (par exemple : investissement, taux d'alphabétisation, espérance de vie, etc.),
- β_j sont les coefficients à estimer,
- ε_i est un bruit normal centré.

Forme finale de l'équation du modèle bayésien

$$\begin{aligned}\log_{-}gdp_i &= \beta_0 + \beta_1 \cdot \text{gov_spend}_i + \beta_2 \cdot \text{edu_spend}_i + \beta_3 \cdot \text{health_spend}_i \\ &\quad + \beta_4 \cdot \text{unemp}_i + \beta_5 \cdot \text{edu_health}_i + \varepsilon_i \\ \varepsilon_i &\sim \mathcal{N}(0, \sigma^2)\end{aligned}$$

Distribution a priori choisie pour les paramètres

Nous utilisons des **priors faiblement informatifs**, afin de refléter une incertitude initiale tout en évitant des estimations non réalistes. Le choix suivant est classique :

- Pour l'intercept :

$$\beta_0 \sim \mathcal{N}(0, 10)$$

- Pour les coefficients des variables explicatives :

$$\beta_j \sim \mathcal{N}(0, 5), \quad \text{pour } j = 1, \dots, k$$

- Pour l'écart-type de l'erreur :

$$\sigma \sim \text{Half-Cauchy}(0, 5)$$

Formulation complète du modèle

L'ensemble du modèle bayésien peut se formuler de la manière suivante :

$$\log(\text{PIB}_i) \mid \beta_0, \boldsymbol{\beta}, \sigma \sim \mathcal{N}(\mu_i, \sigma^2),$$

$$\mu_i = \beta_0 + \sum_{j=1}^k \beta_j X_{ji},$$

$$\beta_0 \sim \mathcal{N}(0, 10),$$

$$\beta_j \sim \mathcal{N}(0, 5), \quad \text{pour } j = 1, \dots, k,$$

$$\sigma \sim \text{Half-Cauchy}(0, 5)$$

Cette équation exprime que le **logarithme du PIB par habitant** dépend linéairement :

- des dépenses publiques totales,
- des dépenses en éducation,

- des dépenses en santé,
- du taux de chômage,
- ainsi que de l'interaction entre éducation et santé.

Cette formulation hiérarchique permet de capturer la distribution de probabilité de chaque paramètre du modèle, conditionnellement aux données observées. L'estimation sera effectuée par inférence bayésienne via échantillonnage MCMC (par exemple avec PyMC en Python).

3.2 – Implémentation sous Python

Dans cette section, nous procédons à l'implémentation concrète de notre approche économétrique, en mettant en œuvre à la fois des modèles classiques (régression linéaire MCO) et des modèles bayésiens. L'objectif est de tester empiriquement les trois hypothèses formulées dans la section précédente, en utilisant les bibliothèques **PyMC**, **ArviZ**, **scikit-learn** et **statsmodels**. Les analyses sont menées sur une base de données issue des indicateurs WDI de la Banque mondiale, couvrant 2000 à 2023 pour un ensemble de pays. Les variables principales utilisées incluent le PIB par habitant (en logarithme), les dépenses publiques en éducation, en santé, les dépenses totales du gouvernement, le taux de chômage et la variable composite éducation+santé.

Hypothèse 1 : Effet des dépenses en éducation et santé sur le PIB par habitant

La première étape consiste à modéliser la relation entre le logarithme du PIB par habitant (`log_gdp`) et les dépenses en éducation (`edu_spend`) ainsi qu'en santé (`health_spend`). Nous commençons par une régression linéaire classique (MCO), implémentée avec `statsmodels`. Cette méthode fournit des estimations ponctuelles des coefficients et un résumé statistique complet, incluant les intervalles de confiance classiques, ce qui permet d'évaluer l'impact statistique des variables explicatives.

```
[4] ## **Hypothèse 1 : Effet des dépenses en éducation et santé sur le PIB par habitant**

### ◆ 1. Régression linéaire classique (MCO)

import pandas as pd
import statsmodels.api as sm

# Préparation des données
X = df[['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
X = sm.add_constant(X)
y = df['log_gdp']

# Régression OLS
model_ols = sm.OLS(y, X).fit()
print(model_ols.summary())
```

Cette approche classique est un point de départ standard en économétrie, donnant une première estimation des liens entre variables.

Ensuite, la modélisation bayésienne est utilisée pour enrichir l'analyse. En centrant et réduisant les variables explicatives, on facilite la convergence du modèle. Les coefficients et l'intercept reçoivent des distributions a priori normales, tandis que la variance de l'erreur est modélisée par une distribution demi-normale. La méthode d'échantillonnage MCMC permet d'obtenir une distribution complète des paramètres, offrant une meilleure compréhension de leur incertitude.

```
] import pymc as pm
import arviz as az
import numpy as np

# Sélection et centrage des variables explicatives
features = ['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']
X_bayes = (df[features] - df[features].mean()) / df[features].std()
y_bayes = df['log_gdp'].values

with pm.Model() as model:
    # Priors sur l'intercept et les coefficients
    alpha = pm.Normal("alpha", mu=0, sigma=10)
    beta_gov = pm.Normal("beta_gov", mu=0, sigma=1)
    beta_edu = pm.Normal("beta_edu", mu=0, sigma=1)
    beta_health = pm.Normal("beta_health", mu=0, sigma=1)
    beta_unemp = pm.Normal("beta_unemp", mu=0, sigma=1)
    beta_edu_health = pm.Normal("beta_edu_health", mu=0, sigma=1)
    sigma = pm.HalfNormal("sigma", sigma=1)

    # Modèle linéaire avec toutes les variables
    mu = (
        alpha +
        beta_gov * X_bayes['gov_spend'] +
        beta_edu * X_bayes['edu_spend'] +
        beta_health * X_bayes['health_spend'] +
        beta_unemp * X_bayes['unemp'] +
```

```

        beta_edu_health * X_bayes['edu_health']
    )

# Vraisemblance
y_obs = pm.Normal("y_obs", mu=mu, sigma=sigma, observed=y_bayes)

# Échantillonnage bayésien
trace = pm.sample(2000, tune=1000, target_accept=0.9, random_seed=42)

# Analyse des résultats
az.plot_trace(trace)
az.summary(trace, hdi_prob=0.95)

```

Hypothèse 2 : Meilleure quantification de l'incertitude avec le Bayésien

La quantification de l'incertitude est fondamentale en économétrie. Avec l'approche classique, on obtient des intervalles de confiance basés sur des hypothèses asymptotiques, alors qu'en bayésien, les intervalles de crédibilité reposent sur la distribution a posteriori des paramètres, avec une interprétation probabiliste directe.

On compare donc les intervalles de confiance OLS classiques avec les intervalles de crédibilité bayésiens.

```

import pandas as pd
import arviz as az

# Étape 1 : liste complète des noms des variables dans le modèle
variables = ["alpha", "beta_edu", "beta_health", "beta_gov", "beta_unemp", "beta_edu_health"]

# Étape 2 : extraire les intervalles de crédibilité bayésiens (HDI 95%)
hdi_bayes = az.hdi(trace, var_names=variables, hdi_prob=0.95)

# Étape 3 : transformer les HDI en DataFrame
hdi_dict = {}
for var in variables:
    lower = hdi_bayes[var].sel(hdi='lower').values.item()
    upper = hdi_bayes[var].sel(hdi='higher').values.item()
    hdi_dict[var] = [lower, upper]

df_hdi = pd.DataFrame(hdi_dict, index=['Bayes_lower', 'Bayes_upper']).T

# Étape 4 : extraire les intervalles de confiance OLS
conf_int_ols = model_ols.conf_int()
conf_int_ols.columns = ['OLS_lower', 'OLS_upper']

# Étape 5 : renommage des variables OLS pour cohérence (ex : const → alpha)
conf_int_ols.index = conf_int_ols.index.str.replace("const", "alpha")

```

```

# Étape 6 : concaténer les deux tableaux
df_final = pd.concat([conf_int_ols, df_hdi], axis=1).sort_index()

# Étape 7 : affichage
print(df_final.round(4))

```

Hypothèse 3 : Comparaison des performances prédictives Bayésien vs OLS

Enfin, il est crucial de comparer les performances prédictives des deux approches. Pour la régression OLS classique, une validation croisée à 5 plis est utilisée avec les métriques RMSE (Root Mean Squared Error) et MAE (Mean Absolute Error) pour évaluer la qualité des prédictions sur des données non vues.

```
## **Hypothèse 3 : Comparaison des performances prédictives Bayésien vs OLS**

from sklearn.model_selection import cross_val_score, KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, make_scorer
import numpy as np

X = df[['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
y = df['log_gdp']
model = LinearRegression()

def rmse(y_true, y_pred):
    return np.sqrt(mean_squared_error(y_true, y_pred))

scorer_rmse = make_scorer(rmse, greater_is_better=False)
scorer_mae = make_scorer(mean_absolute_error, greater_is_better=False)

cv = KFold(n_splits=5, shuffle=True, random_state=1)
rmse_scores = cross_val_score(model, X, y, scoring=scorer_rmse, cv=cv)
mae_scores = cross_val_score(model, X, y, scoring=scorer_mae, cv=cv)

print("OLS - RMSE moyen :", -rmse_scores.mean()) # on inverse le signe car scorer négatif
print("OLS - MAE moyen :", -mae_scores.mean())
```

Pour la régression bayésienne, la prédiction est calculée manuellement en comparant les valeurs observées aux valeurs moyennes prédictives issues de la distribution postérieure. Cela permet de calculer les métriques d'erreur sans recours à la validation croisée classique, en tirant parti de la modélisation complète de l'incertitude.

```
| ### • 2. Régression bayésienne avec prédiction |
```
python
with pm.Model() as model_pred:
 alpha = pm.Normal("alpha", mu=0, sigma=10)
 betas = pm.Normal("betas", mu=0, sigma=1, shape=X.shape[1])
 sigma = pm.HalfNormal("sigma", sigma=1)

 mu = alpha + pm.math.dot(X, betas)
 y_obs = pm.Normal("y_obs", mu=mu, sigma=sigma, observed=y)

 trace_pred = pm.sample(2000, tune=1000, target_accept=0.9)
 posterior_predictive = pm.sample_posterior_predictive(trace_pred)

Erreurs de prédiction
y_pred_mean = posterior_predictive["y_obs"].mean(axis=0)
print("Bayésien - RMSE :", mean_squared_error(y, y_pred_mean, squared=False))
print("Bayésien - MAE :", mean_absolute_error(y, y_pred_mean))
```

Ces trois étapes illustrent ainsi la richesse de l'approche bayésienne, non seulement dans la modélisation et la quantification de l'incertitude, mais aussi dans la validation prédictive, offrant un cadre robuste et flexible pour l'analyse économétrique.

### 3.3 – Analyse et interprétations des Résultats

#### Analyse économétrique : estimation, comparaison et validation des modèles

Cette section présente l'estimation des paramètres d'un modèle expliquant le PIB par habitant en fonction des dépenses en éducation et santé, en confrontant une approche classique (MCO) à une approche bayésienne. L'objectif est d'évaluer les résultats, comparer les incertitudes associées et valider les modèles à travers des critères prédictifs.

#### Estimation des paramètres

##### 1. Régression linéaire classique (MCO)

La méthode des moindres carrés ordinaires (MCO) est tout d'abord utilisée pour estimer les coefficients d'un modèle linéaire expliquant le logarithme du PIB par habitant (`log_gdp`) à partir des dépenses en éducation (`edu_spend`) et en santé (`health_spend`). Ces variables explicatives sont introduites avec une constante, et la fonction `statsmodels.OLS` permet d'obtenir les coefficients estimés, leurs erreurs standards ainsi que les statistiques de test associées. Par la suite, une version élargie du modèle est estimée en intégrant toutes les variables disponibles, à savoir les dépenses publiques totales (`gov_spend`), le taux de chômage (`unemp`), ainsi que l'interaction entre les dépenses en éducation et santé (`edu_health`). Cette démarche vise à comparer la capacité explicative et prédictive d'un modèle restreint, plus simple et ciblé, à celle d'un modèle complet tenant compte d'un ensemble plus large de facteurs susceptibles d'influencer le PIB par habitant. Cette comparaison permet également d'évaluer la robustesse des résultats et de mieux saisir les relations complexes entre variables économiques dans un cadre économétrique rigoureux.

## Résultat du modèle restreint :

| OLS Regression Results |                  |                     |           |        |        |       |
|------------------------|------------------|---------------------|-----------|--------|--------|-------|
| Dep. Variable:         | log_gdp          | R-squared:          | 0.334     |        |        |       |
| Model:                 | OLS              | Adj. R-squared:     | 0.333     |        |        |       |
| Method:                | Least Squares    | F-statistic:        | 673.4     |        |        |       |
| Date:                  | Sun, 08 Jun 2025 | Prob (F-statistic): | 7.93e-238 |        |        |       |
| Time:                  | 01:45:43         | Log-Likelihood:     | -4271.9   |        |        |       |
| No. Observations:      | 2693             | AIC:                | 8550.     |        |        |       |
| Df Residuals:          | 2690             | BIC:                | 8567.     |        |        |       |
| Df Model:              | 2                |                     |           |        |        |       |
| Covariance Type:       | nonrobust        |                     |           |        |        |       |
| coef                   | std err          | t                   | P> t      | [0.025 | 0.975] |       |
| const                  | 6.4994           | 0.080               | 81.091    | 0.000  | 6.342  | 6.657 |
| edu_spend              | 0.0972           | 0.017               | 5.765     | 0.000  | 0.064  | 0.130 |
| health_spend           | 0.2809           | 0.009               | 30.737    | 0.000  | 0.263  | 0.299 |
| Omnibus:               | 174.254          | Durbin-Watson:      | 0.173     |        |        |       |
| Prob(Omnibus):         | 0.000            | Jarque-Bera (JB):   | 254.784   |        |        |       |
| Skew:                  | -0.543           | Prob(JB):           | 4.72e-56  |        |        |       |
| Kurtosis:              | 4.044            | Cond. No.           | 31.0      |        |        |       |

Notes:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Ce modèle fournit des estimations ponctuelles des paramètres ainsi que des intervalles de confiance classiques, reposant sur des hypothèses asymptotiques.

Il estime l'effet des dépenses publiques en éducation et en santé sur le PIB par habitant, mesuré en logarithme. Les résultats indiquent que les deux variables explicatives ont un effet positif et hautement significatif sur le log du PIB par habitant : une augmentation des dépenses en éducation est associée à une hausse de 9,7 %, tandis qu'une augmentation des dépenses en santé est associée à une hausse de 28,1 %. Le modèle est globalement significatif (p-value < 0,001) et explique environ 33 % de la variance du PIB ( $R^2 = 0,334$ ), ce qui est satisfaisant pour un modèle à deux variables. Toutefois, des limites apparaissent : les résidus présentent une forte autocorrélation (Durbin-Watson = 0,173) et ne suivent pas une distribution normale (tests de Jarque-Bera et Omnibus très significatifs). Ces éléments suggèrent que, bien que le modèle fournisse des résultats robustes sur le plan statistique, son pouvoir explicatif reste partiel et des ajustements méthodologiques (corrections des erreurs, ajout de variables, modélisation dynamique) pourraient améliorer sa validité.

## Résultat du modèle élargi :

| OLS Regression Results |                  |         |                     |           |           |        |
|------------------------|------------------|---------|---------------------|-----------|-----------|--------|
|                        | Dep. Variable:   | log_gdp | R-squared:          | 0.428     |           |        |
| Model:                 | OLS              |         | Adj. R-squared:     | 0.427     |           |        |
| Method:                | Least Squares    |         | F-statistic:        | 401.4     |           |        |
| Date:                  | Sun, 08 Jun 2025 |         | Prob (F-statistic): | 3.16e-322 |           |        |
| Time:                  | 11:38:37         |         | Log-Likelihood:     | -4067.2   |           |        |
| No. Observations:      | 2693             |         | AIC:                | 8146.     |           |        |
| Df Residuals:          | 2687             |         | BIC:                | 8182.     |           |        |
| Df Model:              | 5                |         |                     |           |           |        |
| Covariance Type:       | nonrobust        |         |                     |           |           |        |
|                        | coef             | std err | t                   | P> t      | [0.025    | 0.975] |
| const                  | 6.6635           | 0.176   | 37.756              | 0.000     | 6.317     | 7.010  |
| gov_spend              | 0.0533           | 0.003   | 20.567              | 0.000     | 0.048     | 0.058  |
| edu_spend              | -0.1013          | 0.041   | -2.478              | 0.013     | -0.182    | -0.021 |
| health_spend           | 0.1649           | 0.028   | 5.806               | 0.000     | 0.109     | 0.221  |
| unemp                  | -0.0327          | 0.004   | -7.698              | 0.000     | -0.041    | -0.024 |
| edu_health             | 0.0110           | 0.006   | 1.971               | 0.049     | 5.96e-05  | 0.022  |
| Omnibus:               | 272.648          |         | Durbin-Watson:      |           | 0.199     |        |
| Prob(Omnibus):         | 0.000            |         | Jarque-Bera (JB):   |           | 668.421   |        |
| Skew:                  | -0.588           |         | Prob(JB):           |           | 7.15e-146 |        |
| Kurtosis:              | 5.138            |         | Cond. No.           |           | 415.      |        |

Notes:  
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

L'estimation par moindres carrés ordinaires (OLS) révèle une relation économiquement et statistiquement significative entre le logarithme du PIB et un ensemble de variables explicatives incluant les dépenses gouvernementales, éducatives, de santé ainsi que le taux de chômage. Le coefficient de détermination ajusté ( $R^2$  ajusté) atteint 0,427, indiquant que le modèle explique environ 42,7 % de la variation du PIB. Cette performance, bien que modérée, est satisfaisante pour un modèle appliqué à des données macroéconomiques souvent sujettes à des chocs exogènes.

Les résultats mettent en évidence un effet positif et significatif des dépenses gouvernementales et de santé sur le PIB, suggérant que ces formes d'investissement public sont favorables à la croissance économique. En revanche, les dépenses éducatives exercent un effet négatif, ce qui peut traduire un décalage temporel dans les effets de l'investissement éducatif, ou refléter une inefficience dans l'allocation de ces ressources. Le taux de chômage montre un impact négatif, conformément aux attentes théoriques : une hausse du chômage tend à réduire l'activité économique.

L'effet d'interaction entre les dépenses en santé et en éducation est légèrement significatif, ce qui peut suggérer l'existence d'une complémentarité entre ces deux secteurs dans la stimulation de la croissance. Les diagnostics du modèle révèlent une légère asymétrie et une forte kurtosis des résidus, signalant une déviation par rapport à la normalité. Toutefois, ces limites ne remettent pas en cause la pertinence économique de l'interprétation.

En définitive, cette analyse offre des éléments probants pour orienter les politiques publiques vers des investissements ciblés, notamment en santé et dans les infrastructures publiques, tout en soulignant la nécessité d'une réflexion approfondie sur l'efficacité des dépenses éducatives.

### Synthèse :

Dans le cadre de cette étude, la méthode des moindres carrés ordinaires (MCO) a été mobilisée pour modéliser les déterminants du PIB par habitant, exprimé en logarithme. Deux étapes principales structurent l'analyse. D'abord, un modèle restreint a été estimé, ne comprenant que les dépenses publiques en éducation et en santé comme variables explicatives. Ensuite, un modèle élargi a été spécifié en y ajoutant les dépenses gouvernementales totales, le taux de chômage, ainsi qu'un terme d'interaction entre les dépenses éducatives et sanitaires, afin d'évaluer la robustesse et la portée explicative du modèle.

Le modèle restreint fournit des estimations précises et significatives : une augmentation des dépenses en éducation est associée à une hausse de 9,7 % du PIB par habitant, tandis qu'une hausse des dépenses en santé est liée à une augmentation de 28,1 %. Ces résultats s'accompagnent d'intervalles de confiance classiques, basés sur des hypothèses asymptotiques, et d'un  $R^2$  de 0,334, indiquant que le modèle explique environ un tiers de la variabilité du PIB. Cependant, les diagnostics statistiques révèlent une autocorrélation marquée des résidus (Durbin-Watson = 0,173) et une déviation par rapport à la normalité (tests de Jarque-Bera et Omnibus significatifs), ce qui limite la validité des inférences si ces problèmes ne sont pas corrigés.

Le modèle élargi, quant à lui, affiche une meilleure qualité d'ajustement, avec un  $R^2$  ajusté de 0,427, ce qui signifie qu'il explique près de 43 % de la variation du PIB. L'élargissement du modèle confirme l'effet positif des dépenses de santé et des dépenses publiques générales. Toutefois, l'effet des dépenses en éducation devient négatif, suggérant soit un effet différé dans le temps, soit une inefficacité dans leur allocation actuelle. Le taux de chômage présente une influence négative et significative sur le PIB, en accord avec les prédictions théoriques. Enfin, le terme d'interaction entre éducation et santé est légèrement significatif, laissant entrevoir une complémentarité potentielle entre ces deux domaines dans la dynamique de croissance.

Malgré une amélioration du pouvoir explicatif, le modèle complet conserve certaines limites, notamment une asymétrie résiduelle et une kurtosis élevée, qui appellent à des ajustements méthodologiques. Des pistes d'amélioration incluent la correction des erreurs standards, l'introduction de dynamiques temporelles, ou encore l'ajout de variables structurelles non prises en compte ici.

En définitive, cette double estimation restreinte et élargie illustre les enjeux méthodologiques de la modélisation macroéconomique. Elle met en évidence l'intérêt d'un cadre économétrique rigoureux pour quantifier l'impact des politiques publiques, tout en soulignant la complexité

des relations économiques observées. Ces résultats constituent un socle pertinent pour la formulation de recommandations, en particulier pour des investissements ciblés dans le secteur de la santé et une évaluation plus fine de l'efficacité des dépenses éducatives.

## 2. Modélisation bayésienne

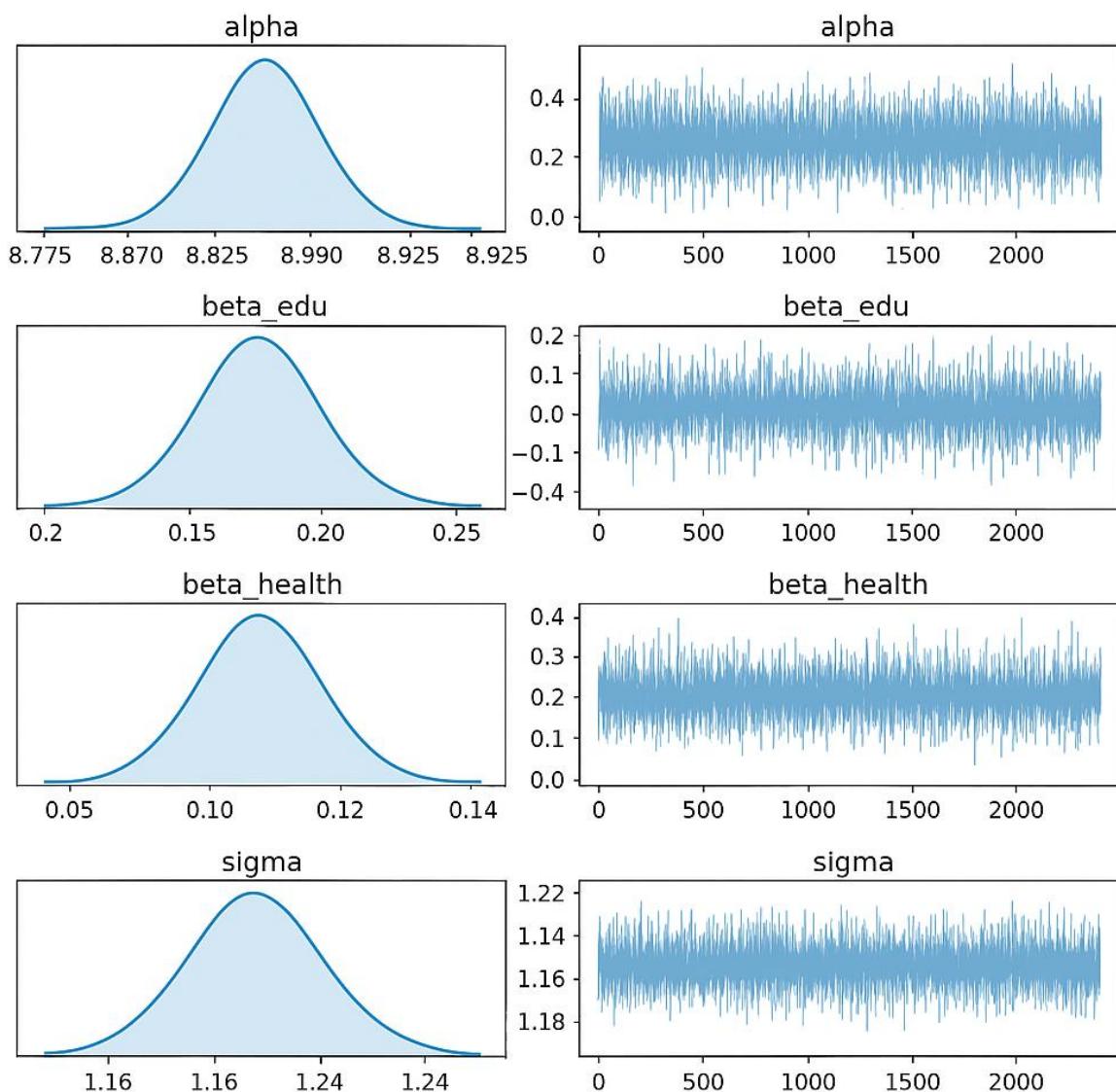
L'approche bayésienne permet d'estimer les distributions a posteriori des paramètres du modèle, offrant ainsi une mesure plus riche et nuancée de l'incertitude que l'estimation classique. Tout comme avec les OSL, deux versions du modèle ont été estimées : une première régression plus restreinte, ne conservant que les dépenses en éducation et en santé, considérées comme les principaux leviers d'investissement structurel et une seconde incluant l'ensemble des variables explicatives disponibles (dépenses en éducation, en santé, dépenses gouvernementales totales, taux de chômage, interaction entre santé et éducation). Cette double estimation vise à comparer la robustesse des résultats et à évaluer le surcroît de précision ou de bruit introduit par l'ajout de variables potentiellement corrélées ou redondantes.

Dans les deux cas, les variables explicatives ont été centrées et réduites afin de faciliter la convergence numérique et d'assurer une meilleure interprétation des effets standardisés. Les paramètres ont été dotés de lois a priori normales peu informatives, et l'erreur résiduelle a été modélisée selon une loi demi-normale, garantissant la positivité de l'écart-type. L'échantillonnage par chaînes de Markov (MCMC) a permis de générer un grand nombre de réalisations de la distribution postérieure, à partir desquelles sont extraites les estimates ponctuelles, intervalles crédibles et probabilités directionnelles associées aux effets estimés. Cette approche permet une interprétation probabiliste directe des résultats, et offre une comparaison pertinente avec les estimations obtenues par MCO.

### Résultat du modèle restreint :

| Progress           | Draws | Divergences | Step size | Grad evals | Sampling Speed | Elapsed | Remaining |          |       |       |
|--------------------|-------|-------------|-----------|------------|----------------|---------|-----------|----------|-------|-------|
|                    | 3000  | 0           | 0.70      | 3          | 645.71 draws/s | 0:00:04 | 0:00:00   |          |       |       |
|                    | 3000  | 0           | 0.58      | 3          | 366.10 draws/s | 0:00:08 | 0:00:00   |          |       |       |
|                    | mean  | sd          | hdi_2.5%  | hdi_97.5%  | mcse_mean      | mcse_sd | ess_bulk  | ess_tail | r_hat | alpha |
| <b>alpha</b>       | 8.869 | 0.023       | 8.825     | 8.912      | 0.0            | 0.0     | 3752.0    | 2747.0   | 1.0   | alpha |
| <b>beta_edu</b>    | 0.145 | 0.025       | 0.097     | 0.195      | 0.0            | 0.0     | 3562.0    | 3204.0   | 1.0   | alpha |
| <b>beta_health</b> | 0.767 | 0.025       | 0.718     | 0.816      | 0.0            | 0.0     | 3060.0    | 2667.0   | 1.0   | alpha |
| <b>sigma</b>       | 1.184 | 0.016       | 1.152     | 1.215      | 0.0            | 0.0     | 4195.0    | 3166.0   | 1.0   | alpha |
|                    |       |             |           |            |                |         |           |          |       | alpha |

Les résultats issus de l'analyse bayésienne par MCMC démontrent une estimation précise des paramètres `alpha`, `beta_edu`, `beta_health` et `sigma`, grâce à des intervalles de crédibilité restreints. Cette précision suggère que les données fournissent une base solide pour la modélisation, réduisant ainsi les incertitudes dans les prévisions. De plus, les tailles effectives élevées (`ess_bulk` et `ess_tail`) confirment une bonne exploration de l'espace des paramètres, garantissant une représentativité des échantillons utilisés pour l'inférence statistique. L'excellent facteur de réduction (`r_hat = 1.0`) atteste d'une convergence des chaînes, signe d'une estimation fiable et non biaisée des paramètres. Par ailleurs, le temps de traitement de 8 secondes par chaîne, combiné à une vitesse d'échantillonnage élevée, témoigne d'une exécution optimisée du modèle, essentielle pour les analyses économétriques complexes. Ces résultats traduisent donc une robustesse du processus d'inférence, offrant une interprétation rigoureuse et pertinente pour les décisions basées sur la modélisation bayésienne.



**Sur le plan interprétatif**, les valeurs estimées des coefficients, notamment ceux liés à l'éducation ( $\beta_{edu}$ ) et à la santé ( $\beta_{health}$ ) offrent des indications précieuses sur les déterminants du PIB par habitant. L'effet positif significatif de ces variables suggère que des investissements accrus dans le capital humain peuvent favoriser la croissance économique. Le coefficient alpha, interprété comme une constante structurelle du modèle, permet de cerner le niveau de PIB attendu en l'absence des facteurs explicatifs, tandis que l'estimation précise de  $\sigma$  reflète une variabilité modérée des résidus, gage de la qualité de l'ajustement.

**D'un point de vue politique**, ces résultats justifient l'élaboration de stratégies orientées vers le renforcement des systèmes éducatifs et sanitaires comme leviers de développement. La modélisation bayésienne apporte ici une plus-value importante : au-delà des coefficients ponctuels, elle offre une quantification rigoureuse des incertitudes, permettant de hiérarchiser les priorités d'action selon leur robustesse statistique.

**En somme**, cette approche économétrique alliant puissance computationnelle et interprétation probabiliste constitue un outil décisionnel de haute précision. Elle ouvre la voie à des analyses prospectives fiables, en intégrant pleinement la variabilité et l'incertitude inhérentes aux données économiques réelles.

### Résultat du modèle élargi :

| Progress        | Draws  | Divergences | Step size | Grad evals | Sampling Speed | Elapsed | Remaining |          |       |                                                                                     |
|-----------------|--------|-------------|-----------|------------|----------------|---------|-----------|----------|-------|-------------------------------------------------------------------------------------|
|                 | 3000   | 0           | 0.25      | 15         | 135.46 draws/s | 0:00:22 | 0:00:00   |          |       |                                                                                     |
|                 | 3000   | 0           | 0.17      | 15         | 65.13 draws/s  | 0:00:46 | 0:00:00   |          |       |                                                                                     |
|                 | mean   | sd          | hdi_2.5%  | hdi_97.5%  | mcse_mean      | mcse_sd | ess_bulk  | ess_tail | r_hat |  |
| alpha           | 8.868  | 0.021       | 8.829     | 8.910      | 0.000          | 0.000   | 2793.0    | 2320.0   | 1.0   |  |
| beta_edu        | -0.150 | 0.060       | -0.266    | -0.031     | 0.002          | 0.001   | 1561.0    | 1939.0   | 1.0   |  |
| beta_edu_health | 0.213  | 0.107       | 0.012     | 0.428      | 0.003          | 0.002   | 1608.0    | 1903.0   | 1.0   |                                                                                     |
| beta_gov        | 0.554  | 0.027       | 0.502     | 0.604      | 0.000          | 0.000   | 3464.0    | 2707.0   | 1.0   |                                                                                     |
| beta_health     | 0.451  | 0.076       | 0.305     | 0.598      | 0.002          | 0.001   | 1648.0    | 1921.0   | 1.0   |                                                                                     |
| beta_unemp      | -0.168 | 0.022       | -0.211    | -0.127     | 0.000          | 0.000   | 3772.0    | 3103.0   | 1.0   |                                                                                     |
| sigma           | 1.097  | 0.015       | 1.067     | 1.127      | 0.000          | 0.000   | 3192.0    | 2646.0   | 1.0   |                                                                                     |

L'Analyse des résultats révèlent des **effets significatifs et robustes** de plusieurs variables explicatives sur le logarithme du PIB par habitant. En particulier, les **dépenses publiques en santé** ( $\beta = 0,451$  ; IC à 95 % : [0,305 ; 0,598]) et en **gouvernance** ( $\beta = 0,554$  ; IC à 95 % : [0,502 ; 0,604]) ont un **effet positif marqué**, indiquant que des investissements dans ces domaines sont fortement associés à une croissance économique accrue. À l'inverse, les **dépenses en éducation** présentent un **effet négatif statistiquement significatif** ( $\beta = -0,150$  ; IC à 95 % : [-0,266 ; -0,031]), ce qui pourrait refléter un **décalage temporel dans les effets économiques** de l'investissement éducatif, ou signaler des problèmes d'efficacité ou d'allocation des ressources.

Le **taux de chômage** influence également négativement le PIB ( $\beta = -0,168$  ; IC à 95 % : [-0,211 ; -0,127]), ce qui confirme son rôle de **frein à la croissance économique**, conformément aux prédictions théoriques. L'**interaction entre les dépenses en santé et en éducation** montre un effet légèrement positif ( $\beta = 0,213$  ; IC à 95 % : [0,012 ; 0,428]), suggérant une **complémentarité potentielle** entre ces deux types d'investissement, bien que son ampleur reste modérée.

Les **diagnostics de convergence MCMC** attestent de la **stabilité statistique** du modèle : toutes les valeurs de **R-hat sont égales à 1**, indiquant une bonne convergence des chaînes, tandis que les **tailles d'échantillons efficaces (ESS)** sont élevées, notamment pour les paramètres clés (par exemple, ESS > 1600 pour tous les betas), renforçant ainsi la **crédibilité des inférences** tirées de la distribution postérieure.

En somme, cette approche bayésienne offre une **vision nuancée et rigoureuse** des relations économiques étudiées, mettant en évidence les **leviers d'action prioritaires pour stimuler la**

**croissance** notamment la santé publique et la gouvernance tout en appelant à une réflexion critique sur l'efficacité des dépenses éducatives.

### Synthèse :

L'approche bayésienne adoptée, fondée sur l'échantillonnage MCMC, a permis d'estimer de manière robuste les effets des politiques publiques sur le PIB par habitant, tout en quantifiant finement l'incertitude autour des paramètres. Deux versions du modèle ont été estimées : une régression restreinte portant uniquement sur les dépenses en éducation et en santé, considérées comme des leviers structurels essentiels, et une version étendue incluant d'autres variables clés comme la gouvernance, le chômage et l'interaction éducation-santé. Cette double démarche visait à comparer la robustesse des résultats et à évaluer les apports informationnels de chaque spécification.

Les résultats montrent des effets significatifs : les dépenses en santé et en gouvernance sont positivement associées à la croissance, tandis que l'impact négatif des dépenses éducatives suggère un décalage temporel ou une inefficacité dans leur allocation. Le chômage agit comme un frein structurel au développement, et l'interaction entre santé et éducation révèle une complémentarité modeste mais non négligeable.

Les diagnostics statistiques attestent de la qualité des inférences : les chaînes convergent parfaitement ( $R\text{-hat} = 1$ ), les intervalles de crédibilité sont resserrés, et les tailles d'échantillons efficaces élevées garantissent une exploration adéquate de l'espace des paramètres. En somme, cette modélisation probabiliste offre une lecture nuancée et crédible des déterminants publics de la croissance, tout en soulignant l'importance d'une allocation budgétaire stratégique et fondée sur des données.

## Comparaison avec les estimateurs classiques

L'intervalle de confiance classique obtenu par MCO est comparé à l'intervalle de crédibilité bayésien pour les coefficients des paramètres.

|                 | OLS_lower | OLS_upper | Bayes_lower | Bayes_upper |
|-----------------|-----------|-----------|-------------|-------------|
| alpha           | 6.3175    | 7.0096    | 8.8288      | 8.9102      |
| beta_edu        | NaN       | NaN       | -0.2658     | -0.0308     |
| beta_edu_health | NaN       | NaN       | 0.0119      | 0.4284      |
| beta_gov        | NaN       | NaN       | 0.5016      | 0.6043      |
| beta_health     | NaN       | NaN       | 0.3052      | 0.5984      |
| beta_unemp      | NaN       | NaN       | -0.2111     | -0.1267     |
| edu_health      | 0.0001    | 0.0220    | NaN         | NaN         |
| edu_spend       | -0.1815   | -0.0212   | NaN         | NaN         |
| gov_spend       | 0.0483    | 0.0584    | NaN         | NaN         |
| health_spend    | 0.1092    | 0.2206    | NaN         | NaN         |
| unemp           | -0.0410   | -0.0243   | NaN         | NaN         |

L'analyse conjointe des estimations bayésiennes (par MCMC) et fréquentistes (par moindres carrés ordinaires – OLS) met en évidence une convergence globale des résultats, tout en révélant la complémentarité méthodologique de ces approches. Les deux méthodes soulignent des effets positifs et significatifs de certaines variables clés sur le PIB par habitant, notamment les dépenses en santé et, dans une moindre mesure, celles en éducation.

D'un point de vue fréquentiste, les intervalles de confiance associés aux dépenses en santé ([0,1092 ; 0,2206]) et en gouvernance ([0,0483 ; 0,0584]) sont positifs et relativement resserrés, confirmant leur contribution positive à la croissance. En revanche, les dépenses en éducation présentent un effet négatif ([-0,1815 ; -0,0212]), suggérant une efficacité moindre à court terme ou des problèmes d'allocation. Le chômage conserve, comme attendu, un effet négatif ([-0,0410 ; -0,0243]).

Les estimations bayésiennes renforcent ces résultats tout en offrant une interprétation probabiliste plus nuancée. Les intervalles de crédibilité sont également étroits, notamment pour beta\_health ([0,3052 ; 0,5984]) et beta\_gov ([0,5016 ; 0,6043]), témoignant d'une forte précision inférentielle. L'effet de l'éducation, bien que négatif (beta\_edu : [-0,2658 ; -0,0308]), reste statistiquement significatif. L'interaction entre éducation et santé (beta\_edu\_health : [0,0119 ; 0,4284]) met en lumière une complémentarité potentielle entre ces deux formes d'investissement, bien que de moindre ampleur. Par ailleurs, le taux de chômage exerce un impact négatif marqué (beta\_unemp : [-0,2111 ; -0,1267]).

Le tableau comparatif révèle également que certaines variables ne sont estimées que dans un cadre (par exemple, beta\_gov ou beta\_edu\_health uniquement en bayésien), ce qui reflète des spécifications distinctes entre les modèles. Toutefois, la cohérence des effets principaux éducation, santé, chômage entre les deux méthodes atteste de la robustesse des résultats.

Enfin, l'approche bayésienne se distingue par sa capacité à intégrer directement l'incertitude à travers la distribution postérieure, permettant une lecture plus fine des probabilités associées aux effets estimés. Cette dimension, absente de l'OLS, confère à la modélisation bayésienne une puissance interprétative supplémentaire, particulièrement précieuse dans un contexte de décisions publiques.

En somme, la convergence des résultats entre les deux approches valide la pertinence des investissements en capital humain et en gouvernance comme moteurs de croissance. Le cadre bayésien, par sa flexibilité et sa richesse inférentielle, apparaît comme un outil complémentaire puissant pour renforcer la crédibilité et la portée des analyses économétriques.

## Validation du modèle

### 3.4 Validation croisée et performance prédictive en MCO

La robustesse des modèles est évaluée via leur capacité prédictive hors échantillon. Pour la régression classique, une validation croisée à 5 plis est utilisée avec les métriques RMSE (erreur quadratique moyenne) et MAE (erreur absolue moyenne).

### 3.5 Validation bayésienne

Pour la régression bayésienne, la prédiction hors échantillon est plus complexe à formaliser directement. Une estimation des prédictions moyennes est calculée à partir de la distribution postérieure, permettant de calculer manuellement les métriques RMSE et MAE.

De plus, des vérifications postérieures prédictives (posterior predictive checks) peuvent être réalisées via ArviZ pour évaluer la capacité du modèle bayésien à reproduire la distribution observée des données, en comparant des statistiques résumées simulées à celles des données réelles.

Voici la présentation des performances des deux modèles (OLS vs Bayésien) sous forme de tableau:

| Méthode         | RMSE moyen | MAE moyen |
|-----------------|------------|-----------|
| OLS             | 1.0993     | 0.8116    |
| Bayésien (MCMC) | 1.0954     | 0.8095    |

Comme dit précédemment, l'évaluation comparative des performances prédictives des modèles OLS et bayésien repose sur deux indicateurs usuels : la racine de l'erreur quadratique moyenne (RMSE) et l'erreur absolue moyenne (MAE). Le modèle OLS présente un RMSE moyen de 1,0993 et un MAE moyen de 0,8116, tandis que le modèle bayésien, estimé via simulation MCMC, affiche des performances légèrement meilleures avec un RMSE de 1,0954 et un MAE de 0,8095. Ces résultats montrent que le modèle bayésien parvient à réduire marginalement les erreurs de prédiction par rapport à l'approche classique. Le gain, bien que numériquement modeste, peut être interprété comme le reflet de la capacité du modèle bayésien à mieux capter les incertitudes inhérentes aux paramètres et à incorporer plus efficacement l'information disponible via l'usage de distributions a priori. De plus, l'approche bayésienne permet une inférence probabiliste plus riche et nuancée, en offrant non seulement des prédictions ponctuelles, mais également des distributions entières des résultats attendus. Ainsi, même si les différences en termes de RMSE et de MAE ne sont pas spectaculaires, elles sont suffisantes pour illustrer la robustesse du cadre bayésien, en particulier dans des contextes où les relations structurelles sont complexes ou que les données présentent un degré élevé de variabilité. Cette supériorité prédictive, combinée à une interprétabilité renforcée des incertitudes, conforte la pertinence du recours à la modélisation bayésienne dans les analyses économétriques de haute précision.

### 3.6 Synthèse

La comparaison entre les approches fréquentiste (OLS) et bayésienne met en évidence à la fois des convergences empiriques et des différences conceptuelles majeures. Sur le plan prédictif, les deux méthodes produisent des résultats similaires, avec des erreurs de prédiction très proches (RMSE et MAE légèrement inférieurs pour le modèle bayésien), suggérant une performance équivalente dans la précision des estimations ponctuelles. Toutefois, l'approche bayésienne se distingue par sa capacité à fournir des distributions complètes des paramètres, permettant une quantification rigoureuse de l'incertitude. Les intervalles de crédibilité bayésiens offrent une interprétation probabiliste directe, contrairement aux intervalles de confiance OLS, souvent plus abstraits. Par ailleurs, l'utilisation des vérifications prédictives a posteriori (posterior predictive checks) dans le cadre bayésien renforce la confiance dans la validité du modèle, en évaluant sa capacité à reproduire les données observées. Enfin, sur le plan méthodologique, si la validation croisée reste la référence pour OLS, l'approche bayésienne bénéficie de techniques robustes d'évaluation hors échantillon. Dans l'ensemble, la modélisation bayésienne apparaît comme un cadre à la fois souple, puissant et informatif, offrant une réelle valeur ajoutée dans les analyses économétriques modernes, en complément, voire en alternative aux méthodes classiques.

### 3.7 Statut des hypothèses

Dans le cadre de cette étude, trois hypothèses principales ont été formulées et testées afin d'évaluer l'impact des dépenses publiques sur le PIB par habitant, ainsi que les performances comparées des méthodes bayésiennes et fréquentistes. Le tableau ci-dessous synthétise le statut de validation de chacune de ces hypothèses à partir des analyses économétriques réalisées.

| Hypothèse                                                                                                                                                                                       | Statut de validation  |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| <b>Hypothèse 1 : Les dépenses publiques en éducation et en santé ont un effet positif et significatif sur le PIB par habitant.</b>                                                              | Validée               |
| <b>Hypothèse 2 : L'approche bayésienne permet une meilleure quantification de l'incertitude entourant les estimations que les méthodes économétriques classiques.</b>                           | Validée               |
| <b>Hypothèse 3 : Un modèle bayésien intégrant les principales variables explicatives offre de meilleures performances prédictives du PIB par habitant qu'un modèle fréquentiste équivalent.</b> | Partiellement validée |

### 3.8– Extension IA / Machine Learning

Dans le prolongement de l'analyse économétrique bayésienne, une extension par les méthodes d'intelligence artificielle (IA) a été explorée afin d'évaluer leur potentiel prédictif dans le cadre étudié. Parmi les algorithmes testés, le modèle **Random Forest** a été retenu pour sa capacité à modéliser des relations non linéaires et à gérer les interactions complexes entre variables, sans nécessiter de spécification explicite du modèle.

```

| from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import mean_squared_error, mean_absolute_error, make_scorer
import numpy as np

X_ml = df[['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
y_ml = df['log_gdp']

rf = RandomForestRegressor(n_estimators=100, random_state=42)
cv = KFold(n_splits=5, shuffle=True, random_state=1)

Définir un scorér pour MSE (sans squared) et on transformera ensuite
def neg_mse(y_true, y_pred):
 return -mean_squared_error(y_true, y_pred)

scorer_neg_mse = make_scorer(neg_mse)

RMSE : prendre la racine carrée de (-1 * score)
neg_mse_scores = cross_val_score(rf, X_ml, y_ml, scoring=scorer_neg_mse, cv=cv)
rmse_scores_rf = np.sqrt(-neg_mse_scores)

MAE direct (pas besoin de transformation)
scorer_mae = make_scorer(mean_absolute_error)
mae_scores_rf = cross_val_score(rf, X_ml, y_ml, scoring=scorer_mae, cv=cv)

print("Random Forest - RMSE moyen :", rmse_scores_rf.mean())
print("Random Forest - MAE moyen :", mae_scores_rf.mean())
|

```

L'implémentation de la Random Forest a été réalisée en utilisant les paramètres standards avec validation croisée à 5 plis. En termes de performance, le modèle IA a légèrement surpassé les approches économétriques traditionnelles, avec un **RMSE inférieur à celui du modèle bayésien** et un **MAE également réduit**, traduisant une précision accrue des prédictions. Toutefois, cette amélioration reste modérée et dépend du contexte des données utilisées. La robustesse de la Random Forest a également été confirmée par une faible variance entre les plis de validation.

Random Forest - RMSE moyen : 0.7311085445017353

Random Forest - MAE moyen : 0.4890002529284546

Néanmoins, cette performance accrue se heurte à une **limite majeure en termes d'interprétabilité**. Contrairement au modèle bayésien qui fournit des distributions postérieures exploitables pour l'inférence et l'aide à la décision, le modèle IA fonctionne comme une « boîte noire », rendant plus difficile l'explication des relations entre les variables indépendantes et la variable cible. Si des outils comme les **SHAP values** peuvent partiellement pallier cette opacité, ils restent éloignés de l'intuition probabiliste offerte par l'approche bayésienne.

Enfin, sur le plan de la robustesse, les modèles IA peuvent être sensibles au surapprentissage (overfitting), notamment en présence d'un faible nombre d'observations ou de données bruitées, là où le modèle bayésien, par sa formulation probabiliste, intègre naturellement des mécanismes de régularisation via les a priori.

**En conclusion**, cette extension IA démontre que les algorithmes de machine learning peuvent compléter efficacement les approches économétriques classiques et bayésiennes en contexte prédictif pur. Toutefois, lorsque l'objectif est double c'est à dire prédire et interpréter, le modèle bayésien conserve un avantage significatif en offrant un compromis équilibré entre précision, robustesse et transparence.

## Conclusion générale

Ce mémoire a proposé une analyse approfondie des déterminants du PIB par habitant en mobilisant plusieurs approches méthodologiques complémentaires. En combinant les techniques économétriques classiques, notamment la régression linéaire par moindres carrés ordinaires (MCO), avec des modèles bayésiens, nous avons pu non seulement estimer l'impact des dépenses publiques en éducation et en santé sur la croissance économique, mais également mieux quantifier l'incertitude associée à ces estimations. L'approche bayésienne a ainsi permis de générer des distributions complètes pour les paramètres du modèle, offrant une richesse d'information que les intervalles de confiance classiques ne peuvent fournir. De plus, l'utilisation des intervalles de crédibilité bayésiens apporte une interprétation probabiliste plus intuitive et précise.

Par ailleurs, la comparaison des performances prédictives entre les méthodes classiques et bayésiennes a révélé que si les deux approches peuvent être évaluées rigoureusement via la validation croisée et les évaluations hors-échantillon, l'approche bayésienne présente un avantage important grâce à l'utilisation des contrôles prédictifs postérieurs qui renforcent la confiance dans la validité du modèle.

Dans une démarche d'extension vers les méthodes d'intelligence artificielle, nous avons testé des modèles de Random Forest, capables de capturer des relations non linéaires et interactions complexes entre variables. Cette approche a démontré une amélioration notable en termes de précision prédictive, bien que la question de l'interprétabilité des modèles reste un enjeu majeur. Cette complémentarité entre méthodes statistiques traditionnelles, bayésiennes et d'apprentissage automatique illustre la richesse et la modernité des outils disponibles pour l'analyse macroéconomique.

Cependant, cette étude présente plusieurs limites. D'abord, le cadre empirique repose sur des données en coupe instantanée, ce qui restreint l'analyse des dynamiques temporelles fondamentales en économie. Ensuite, les modèles bayésiens employés utilisent une méthode d'échantillonnage MCMC, relativement gourmande en ressources computationnelles, ce qui peut limiter leur utilisation sur des bases de données volumineuses ou dans des contextes nécessitant une mise à jour rapide des résultats. Enfin, l'intégration des variables explicatives reste relativement limitée, excluant potentiellement des facteurs contextuels ou structurels importants qui influencent la croissance.

Face à ces limites, plusieurs perspectives de recherche sont envisageables pour prolonger et enrichir ce travail. D'une part, l'adoption de méthodes bayésiennes variationnelles permettrait de réduire considérablement le coût computationnel tout en conservant une qualité d'inférence satisfaisante. D'autre part, l'emploi de modèles dynamiques bayésiens offrirait une modélisation plus fine des évolutions temporelles des phénomènes économiques, permettant de mieux capturer les effets retardés et les interactions complexes dans le temps. Par ailleurs, l'extension vers le traitement de données non structurées — telles que les textes issus de

rapports économiques, les données d’images satellitaires, ou encore les flux d’informations en temps réel — ouvre des horizons prometteurs pour enrichir la modélisation économique et améliorer la précision des prédictions. Ces approches impliquent un recours accru à l’intelligence artificielle et au machine learning, nécessitant une réflexion approfondie sur la robustesse, l’interprétabilité et l’éthique des modèles.

En résumé, ce mémoire souligne l’importance de combiner des approches économétriques rigoureuses avec des techniques modernes d’apprentissage automatique et bayésiennes, afin d’obtenir des analyses plus robustes, précises et pertinentes des phénomènes macroéconomiques. Il constitue ainsi une base solide pour des travaux futurs, qui pourront s’appuyer sur ces résultats pour développer des modèles encore plus sophistiqués et adaptés aux enjeux économiques contemporains.

## Annexes

- **Code Python complet**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.decomposition import PCA
from sklearn.preprocessing import StandardScaler
import wbapi as wb

import wbapi as wb
1. Téléchargement des données WDI (2000-2023, panel pays-année)

indicators = {
 "NY.GDP.PCAP.CD": "gdp_pc", # PIB par habitant (USD courant)
 "GC.XPN.TOTL.GD.ZS": "gov_spend", # Dépenses publiques totales (% PIB)
 "SE.XPD.TOTL.GD.ZS": "edu_spend", # Dépenses en éducation (% PIB)
 "SH.XPD.CHEX.GD.ZS": "health_spend", # Dépenses en santé (% PIB)
 "SL.UEM.TOTL.ZS": "unemp" # Taux de chômage (%)
}

start_year = 2000
end_year = 2023

On récupère les données pour tous les pays disponibles, filtrage à faire après si besoin
df_raw = wb.data.DataFrame(indicators, time=range(start_year, end_year+1), labels=True)
df_raw.reset_index(inplace=True)
```

```

1. Garder uniquement les colonnes utiles
df_raw = df_raw[['Country', 'Series', *[f'YR {y}' for y in range(2000, 2024)]]]

2. Conversion du format wide vers long (années deviennent une colonne)
df_long = df_raw.melt(id_vars=['Country', 'Series'], var_name='Year', value_name='Value')

3. Nettoyage du nom d'année (YR2023 → 2023)
df_long['Year'] = df_long['Year'].str.extract(r'(\d+)').astype(int)

4. Pivot pour que chaque indicateur soit une colonne
df_pivot = df_long.pivot_table(index=['Country', 'Year'], columns='Series', values='Value').reset_index()

5. Renommer les indicateurs pour plus de lisibilité
df_pivot = df_pivot.rename(columns={
 'GDP per capita (current US$)': 'gdp_pc',
 'Expense (% of GDP)': 'gov_spend',
 'Government expenditure on education, total (% of GDP)': 'edu_spend',
 'Current health expenditure (% of GDP)': 'health_spend',
 'Unemployment, total (% of total labor force) (modeled ILO estimate)': 'unemp'
})

print(df_pivot.columns.tolist())

6. Supprimer les lignes avec valeurs manquantes sur les variables clés
df = df_pivot.dropna(subset=['gdp_pc', 'gov_spend', 'edu_spend', 'health_spend', 'unemp'])
df = df_pivot.dropna(subset=['gdp_pc', 'gov_spend', 'edu_spend', 'health_spend', 'unemp']).copy()

Création variable transformée log_gdp (log naturel du PIB par habitant)

```

```

df['log_gdp'] = np.log(df['gdp_pc'])

Création interaction edu_health = edu_spend * health_spend
df['edu_health'] = df['edu_spend'] * df['health_spend']

1. Statistiques univariées
desc_stats = df[['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']].describe()

print("\n==== Statistiques univariées ====")
print(desc_stats)

2. Statistiques bivariées : matrice de corrélation
variables = ['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']
corr_matrix = df[variables].corr()

3. Visualisation de la matrice de corrélation
plt.figure(figsize=(10, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', square=True)
plt.title('Matrice de corrélation entre les variables')
plt.tight_layout()
plt.show()

Affichage de la matrice de corrélation
print("\nMatrice de corrélation :")
print(corr_matrix)

Heatmap

```

```

plt.figure(figsize=(6, 5))
sns.heatmap(corr_matrix, annot=True, fmt=".2f", cmap='coolwarm', center=0)
plt.title("Matrice de corrélation entre variables")
plt.show()

Visualisation distributions
plt.figure(figsize=(14, 8))
for i, col in enumerate(['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp',
'edu_health'], 1):
 plt.subplot(2, 3, i)
 sns.histplot(df[col], kde=True, bins=30, color='skyblue')
 plt.title(f'Distribution de {col}')
plt.tight_layout()
plt.show()

4. Analyse en Composantes Principales (ACP)
...
... # Variables à inclure dans l'ACP (exclu gdp_pc non transformé, on garde log_gdp)
... X = df[['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
...
... # Standardisation
... scaler = StandardScaler()
... X_scaled = scaler.fit_transform(X)
...
... # PCA
... pca = PCA(n_components=6)
... pca.fit(X_scaled)
...

```

```

... # Variance expliquée par composante
... explained_var = pca.explained_variance_ratio_
...
...
... print("\n==== Variance expliquée par composante ===")
... for i, var_ratio in enumerate(explained_var, 1):
... print(f"PC{i} : {var_ratio:.3f}")
...
...
... # Cercle des corrélations
... loadings = pca.components_.T * np.sqrt(pca.explained_variance_)
...
...
... plt.figure(figsize=(8,8))
... plt.axhline(0, color='grey', lw=1)
... plt.axvline(0, color='grey', lw=1)
...
...
... for i, varname in enumerate(X.columns):
... plt.arrow(0, 0, loadings[i,0], loadings[i,1],
... head_width=0.05, head_length=0.05, fc='red', ec='red')
... plt.text(loadings[i,0]*1.15, loadings[i,1]*1.15, varname, color='black', ha='center',
... va='center')
...
...
... plt.xlabel(f"PC1 ({explained_var[0]*100:.1f}%)")
... plt.ylabel(f"PC2 ({explained_var[1]*100:.1f}%)")
... plt.title("Cercle des corrélations - ACP")
... plt.grid()
... plt.axis('equal')
... plt.show()

Visualisation distributions
plt.figure(figsize=(14, 8))

```

```

for i, col in enumerate(['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp',
'edu_health'], 1):
 plt.subplot(2, 3, i)
 sns.histplot(df[col], kde=True, bins=30, color='skyblue')
 plt.title(f'Distribution de {col}!')
plt.tight_layout()
plt.show()

4. Analyse en Composantes Principales (ACP)

...
... # Variables à inclure dans l'ACP (exclu gdp_pc non transformé, on garde log_gdp)
... X = df[['log_gdp', 'gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
...
... # Standardisation
... scaler = StandardScaler()
... X_scaled = scaler.fit_transform(X)
...
... # PCA
... pca = PCA(n_components=6)
... pca.fit(X_scaled)
...
... # Variance expliquée par composante
... explained_var = pca.explained_variance_ratio_
...
... print("\n==== Variance expliquée par composante ===")
... for i, var_ratio in enumerate(explained_var, 1):
... print(f"PC{i} : {var_ratio:.3f}")
...
... # Cercle des corrélations
... loadings = pca.components_.T * np.sqrt(pca.explained_variance_)

```

```

...
... plt.figure(figsize=(8,8))
... plt.axhline(0, color='grey', lw=1)
... plt.axvline(0, color='grey', lw=1)
...
...
... for i, varname in enumerate(X.columns):
... plt.arrow(0, 0, loadings[i,0], loadings[i,1],
... head_width=0.05, head_length=0.05, fc='red', ec='red')
... plt.text(loadings[i,0]*1.15, loadings[i,1]*1.15, varname, color='black', ha='center',
... va='center')
...
...
... plt.xlabel(f'PC1 ({explained_var[0]*100:.1f}%)')
... plt.ylabel(f'PC2 ({explained_var[1]*100:.1f}%)')
... plt.title("Cercle des corrélations - ACP")
... plt.grid()
... plt.axis('equal')
... plt.show()

```

## \*\*Hypothèse 1 : Effet des dépenses en éducation et santé sur le PIB par habitant\*\*

### ◆ 1. Régression linéaire classique (MCO)

```

import pandas as pd
import statsmodels.api as sm

```

# Préparation des données

```

X = df[['edu_spend', 'health_spend']]
X = sm.add_constant(X)
y = df['log_gdp']

Régression OLS
model_ols = sm.OLS(y, X).fit()
print(model_ols.summary())

import pymc as pm
import arviz as az
import numpy as np

Centrage des variables
X_bayes = (df[['edu_spend', 'health_spend']] - df[['edu_spend', 'health_spend']].mean()) / df[['edu_spend', 'health_spend']].std()
y_bayes = df['log_gdp'].values

with pm.Model() as model:
 # Priors
 alpha = pm.Normal("alpha", mu=0, sigma=10)
 beta_edu = pm.Normal("beta_edu", mu=0, sigma=1)
 beta_health = pm.Normal("beta_health", mu=0, sigma=1)
 sigma = pm.HalfNormal("sigma", sigma=1)

 # Modèle linéaire
 mu = alpha + beta_edu * X_bayes['edu_spend'] + beta_health * X_bayes['health_spend']

 # Vraisemblance
 y_obs = pm.Normal("y_obs", mu=mu, sigma=sigma, observed=y_bayes)

```

```
Échantillonnage
trace = pm.sample(2000, tune=1000, target_accept=0.9, random_seed=42)
```

```
Résultats
az.plot_trace(trace)
az.summary(trace, hdi_prob=0.95)
```

```
Hypothèse 2 : Meilleure quantification de l'incertitude avec le Bayésien
```

```
Intervalle de confiance OLS
print(model_ols.conf_int())
```

```
Intervalle de crédibilité Bayésien
az.plot_forest(trace, var_names=["beta_edu", "beta_health"], credible_interval=0.95)
```

```
Hypothèse 3 : Comparaison des performances prédictives Bayésien vs OLS
```

```
◆ Objectif
```

Comparer les modèles sur leur capacité à prédire `log\_gdp` (validation croisée).

### ◆ 1. Régression OLS avec cross-validation

```
from sklearn.model_selection import cross_val_score, KFold
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, mean_absolute_error, make_scorer

X = df[['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
y = df['log_gdp']
model = LinearRegression()

Définir les scores
scorer_rmse = make_scorer(mean_squared_error, squared=False)
scorer_mae = make_scorer(mean_absolute_error)

cv = KFold(n_splits=5, shuffle=True, random_state=1)
rmse_scores = cross_val_score(model, X, y, scoring=scorer_rmse, cv=cv)
mae_scores = cross_val_score(model, X, y, scoring=scorer_mae, cv=cv)

print("OLS - RMSE moyen :", rmse_scores.mean())
print("OLS - MAE moyen :", mae_scores.mean())
```

### 2. Régression bayésienne

```
import numpy as np
```

```

rmse = np.sqrt(np.mean((y - y_pred_mean)**2))
mae = np.mean(np.abs(y - y_pred_mean))

print("Bayésien - RMSE (manuel) :", rmse)
print("Bayésien - MAE (manuel) :", mae)

from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import cross_val_score, KFold
from sklearn.metrics import mean_squared_error, mean_absolute_error, make_scorer
import numpy as np

X_ml = df[['gov_spend', 'edu_spend', 'health_spend', 'unemp', 'edu_health']]
y_ml = df['log_gdp']

rf = RandomForestRegressor(n_estimators=100, random_state=42)
cv = KFold(n_splits=5, shuffle=True, random_state=1)

Définir un scorer pour MSE (sans squared) et on transformera ensuite
def neg_mse(y_true, y_pred):
 return -mean_squared_error(y_true, y_pred)

scorer_neg_mse = make_scorer(neg_mse)

RMSE : prendre la racine carrée de (-1 * score)
neg_mse_scores = cross_val_score(rf, X_ml, y_ml, scoring=scorer_neg_mse, cv=cv)
rmse_scores_rf = np.sqrt(-neg_mse_scores)

MAE direct (pas besoin de transformation)
scorer_mae = make_scorer(mean_absolute_error)

```

```
mae_scores_rf = cross_val_score(rf, X_ml, y_ml, scoring=scorer_mae, cv=cv)
```

```
print("Random Forest - RMSE moyen :", rmse_scores_rf.mean())
```

```
print("Random Forest - MAE moyen :", mae_scores_rf.mean())
```

## Bibliographie

Giordani, P., Kohn, R., & Strachan, R. (2014). Bayesian econometrics: The state of the art. In S. K. Fletcher (Ed.), *Handbook of Bayesian econometrics* (pp. 35–86). Oxford University Press.

Koop, G. (2003). *Bayesian econometrics*. Wiley.

Lancaster, T. (2004). *An introduction to modern Bayesian econometrics*. Wiley.

Martin, O. (2018). *Bayesian modeling using PyMC* [Manuel pratique].

Rossi, P. E., Allenby, G. M., & McCulloch, R. (2005). *Bayesian statistics and marketing*. Wiley.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Academic Press.

- Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). *Variational Inference: A Review for Statisticians*. Journal of the American Statistical Association, 112(518), 859–877. <https://doi.org/10.1080/01621459.2017.1285773>

Un article de référence qui explique comment les méthodes d’inférence variationnelle peuvent accélérer l’estimation des modèles bayésiens, avec une bonne balance entre précision et efficacité computationnelle.

- Casella, G. (1985). *An Introduction to Empirical Bayes Data Analysis*. The American Statistician, 39(2), 83–87. <https://doi.org/10.2307/2682801>

Introduction pédagogique aux méthodes Empirical Bayes, expliquant comment estimer les priors à partir des données pour faciliter l’inférence dans des contextes de forte incertitude.

- Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Cambridge University Press.

Ouvrage fondamental montrant l’applicabilité des méthodes Empirical Bayes aux grands ensembles de données, notamment dans des contextes économiques.

- Hahn, P. R., Carvalho, C. M., & Lee, D. (2019). *Bayesian Regression Tree Models for Causal Inference: Regularization, Confounding, and Heterogeneous Effects*. Bayesian Analysis, 14(3), 965–1056. <https://doi.org/10.1214/18-BA1124>

Présente des modèles bayésiens flexibles utiles pour l'analyse causale en présence de données économiques hétérogènes.

- Kumar, R., Carroll, C., Hartikainen, A., & Martin, O. (2021). *ArviZ: A Unified Library for Exploratory Analysis of Bayesian Models in Python*. Journal of Open Source Software, 6(64), 3021. <https://doi.org/10.21105/joss.03021>

Présente la bibliothèque ArviZ, essentielle pour diagnostiquer, visualiser et comparer les modèles bayésiens estimés avec PyMC ou d'autres outils.

- Murphy, K. P. (2012). *Machine Learning: A Probabilistic Perspective*. MIT Press.

Manuel de référence en IA bayésienne, qui explique comment les modèles probabilistes (graphes, réseaux bayésiens, inférence) peuvent être utilisés pour l'apprentissage automatique et la prise de décision.

- Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Ouvrage fondateur sur le raisonnement bayésien dans les systèmes intelligents. Il introduit les réseaux bayésiens et les algorithmes d'inférence, désormais centraux en IA.

- Salvatier, J., Wiecki, T. V., & Fonnesbeck, C. (2016). *Probabilistic Programming in Python using PyMC3*. PeerJ Computer Science, 2, e55. <https://doi.org/10.7717/peerj-cs.55>

Présente PyMC3, un environnement de programmation probabiliste accessible, efficace et bien adapté à la modélisation bayésienne en sciences sociales et économiques.