Séance 7
Régression et corrélation statistique de deux variables

Honoka OYAMADA

Introduction

La séance 7 introduit les statistiques bivariées : corrélation et régression. On étudie la relation entre consommation d'énergie (X) et PIB (Y), d'abord pour 2022, puis sur la période 1962-2022.

Partie 1

Exercice de code – Réponses attendues

1) Sélection des colonnes 2022

J'ai sélectionné avec Pandas les deux colonnes demandées dans pib-vs-energie.csv : PIB_2022 et Utilisation_d_energie_2022.

2) Gestion des données censurées (couples complets uniquement)

J'ai converti les deux colonnes en numérique (valeurs non convertibles -> NaN), puis j'ai supprimé toutes les lignes incomplètes avec dropna(subset=...). Après filtrage, il reste n = 142 couples complets (PIB, énergie).

3) Régression linéaire simple (PIB expliqué par l'énergie)

La variable explicative est la consommation d'énergie (X) et la variable à expliquer est le PIB (Y). J'ai estimé la régression linéaire simple avec scipy.stats.linregress. Remarque de méthode : linregress attend x = variable explicative et y = variable à expliquer ; j'ai donc passé x = énergie et y = PIB.

4) Corrélation simple

J'ai calculé la corrélation linéaire de Pearson entre PIB_2022 et Utilisation_d_energie_2022 : r = 0.886543. Dans une régression linéaire simple, ce r correspond au r-value retourné par linregress (ici r = 0.886543).

5) Graphique de synthèse (nuage + droite de régression)

J'ai produit un nuage de points (énergie en abscisse, PIB en ordonnée) et j'ai superposé la droite de régression estimée.

6) Commentaire (1–2 paragraphes)

Le lien PIB–énergie en 2022 est positif et fort ($r \approx 0.887$). Autrement dit, les territoires consommant davantage d'énergie tendent à présenter un PIB plus élevé. Le coefficient de

détermination $R^2 = 0.786$ indique qu'une part importante de la variance du PIB est associée à une relation linéaire avec la consommation d'énergie, mais qu'il reste aussi une part non expliquée (résidus), liée à d'autres facteurs (taille démographique, structure productive, prix et change, spécialisation, etc.).
Limites : corrélation n'implique pas causalité. Le PIB peut augmenter avec l'énergie parce que les économies produisent plus, mais l'énergie peut aussi augmenter parce que le PIB augmente, et surtout les deux peuvent être entraînés par des variables cachées (population, degré d'industrialisation). Une amélioration possible est de travailler en log-log ou par habitant, ou d'étendre l'analyse à plusieurs années pour tester la stabilité du lien.

1. Questions de cours

1. Quel est l'intérêt de passer des statistiques univariées aux statistiques bivariées ?

L'univarié décrit une variable seule (centre, dispersion, forme). La bivarie permet d'étudier une relation : comment deux variables varient ensemble, si l'une peut expliquer (au sens statistique) l'autre, et quelle est la force de cette liaison. En géographie, cela sert à tester des hypothèses et à analyser les écarts au modèle via les résidus.

2. Quelles différences entre la corrélation et la correspondance ? Qu'est-ce qu'un rapport de corrélation ?

La corrélation/régression traite surtout des variables quantitatives connues individu par individu (nuage de points). Les correspondances (chi2, tableaux de contingence, AFC) traitent des variables qualitatives connues par classes/comptes d'effectifs. Un rapport de corrélation est un indicateur normalisé d'intensité de liaison fondé sur la part de variance expliquée (utilisé notamment quand on relie une variable quantitative à une variable qualitative).

3. Quelles différences entre valeurs marginales et conditionnelles ? Pourquoi les distinguer ?

Les valeurs marginales décrivent X et Y séparément (distributions, moyennes, variances). Les valeurs conditionnelles décrivent une variable en fonction de l'autre (ex. $E[Y|X=x]$ ou moyenne de Y par classes de X). Les distinguer permet de séparer le contexte global (marges) de la relation, qui s'exprime justement par des comportements conditionnels.

4. Quelles différences entre variance et covariance ?

La variance mesure la dispersion d'une variable autour de sa moyenne : $Var(X)=E[(X-E[X])^2]$. La covariance mesure la co-dispersion : $Cov(X,Y)=E[(X-E[X])*(Y-E[Y])]$. Le signe indique si les écarts vont dans le même sens (positive) ou en sens contraire (négative). La corrélation r est une covariance normalisée par les écarts-types.

5. Pourquoi mesurer la corrélation ou l'indépendance ?
Pour quantifier la force d'un lien, comparer plusieurs couples de variables, choisir des variables pertinentes pour expliquer/prédire, détecter des redondances (colinéarité), et vérifier si l'intensité observée est compatible avec le hasard (tests de significativité).

6. Quel est le principe de la méthode des moindres carres ? À quoi sert-elle ?
Elle choisit les paramètres du modèle (ex. droite) qui minimisent la somme des carres des résidus sum_i (y_i - y_hat_i)^2. Elle sert à ajuster une tendance moyenne, produire des prédictions et quantifier l'erreur d'ajustement.

7. Expliquez ce qu'est la théorie de la corrélation (simple).
Elle résume l'intensité et le sens de la liaison entre deux variables par un nombre. Le coefficient de Pearson r (entre -1 et +1) est positif si le nuage suit une droite croissante, négatif si décroissante, et proche de 0 s'il n'y a pas de liaison linéaire. Il est sans unité car il est normalisé.

8. En quoi consiste le piège de l'autocorrélation ?
Quand les observations ne sont pas indépendantes (temps/espace), deux séries peuvent sembler très corrélées car elles partagent une tendance ou un effet de voisinage. On obtient alors des r artificiellement élevés et des tests biaisés. Il faut traiter la dépendance (de-trend, différenciation, modèles temporels/spatiaux).

9. Expliquez ce qu'est une régression linéaire.
Une régression linéaire ajuste une relation affine Y_hat = aX + b pour expliquer Y par X. La pente a traduit la variation moyenne de Y pour +1 unité de X ; l'intersection b positionne la droite. Les résidus e_i = y_i - y_hat_i servent à évaluer la qualité de l'ajustement et à détecter des observations atypiques.

10. Différence entre coefficient de corrélation et coefficient de détermination
Le coefficient r mesure la liaison linéaire et son signe. Le coefficient de détermination R^2 est r^2 en régression linéaire simple : il mesure la part de variance de Y expliquée par X (entre 0 et 1).

11. Pourquoi faut-il tester les deux droites de régression ?
La régression de Y sur X minimise les erreurs verticales (sur Y), alors que la régression de X sur Y minimise les erreurs horizontales (sur X) : elles diffèrent en général. Estimer les deux rappelle que la corrélation n'impose pas une causalité directionnelle et aide à juger la robustesse de l'interprétation si l'on inverse les rôles.

2. Mise en œuvre avec Python : PIB vs énergie

2.1 Nettoyage (2022) et gestion des données censurées
On sélectionne PIB_2022 et Utilisation_d_energie_2022, on convertit en numérique (valeurs non convertibles -> NaN), puis on supprime les couples incomplets avec dropna sur les deux colonnes. Après nettoyage, n = 142 observations restent.

2.2 Régression linéaire simple (2022)

Modèle : PIB = a * énergie + b (X = énergie, Y = PIB).

Indicator	Valeur (2022)
Pente a	6.33728
Intercept b	7.31031e+10
r (linregress)	0.886543
R2	0.785959
p-value (test a=0)	1.034e-48
Erreur standard (a)	0.279504
Correlation Pearson (Pandas)	0.886543

2.3 Graphique 2022

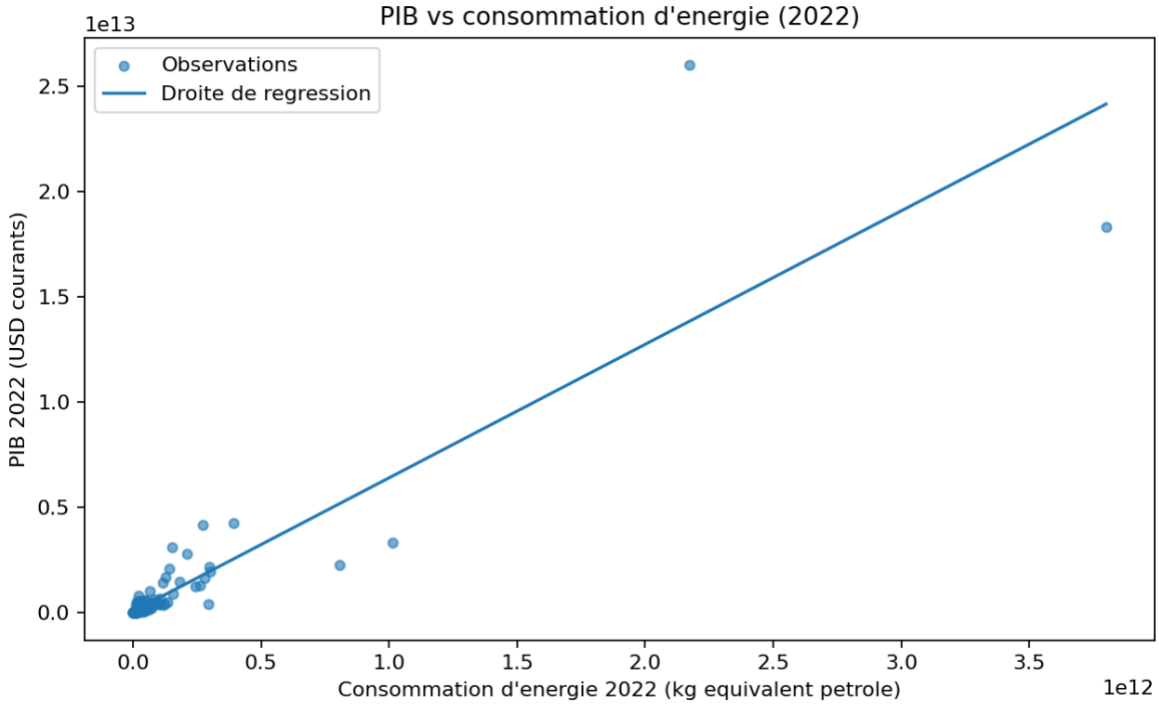

Figure 1 - Nuage de points et droite de régression (2022).

2.4 Commentaire (interprétation)

En 2022, la corrélation PIB-énergie est positive (r = 0.887) : les territoires qui consomment davantage d'énergie tendent à présenter un PIB plus élevé. La pente a indiqué la variation

moyenne du PIB associée à +1 kg d'équivalentes de pétrole consommés. Le coefficient de détermination $R^2 = 0,786$ représente la part de variance du PIB explicable par une relation linéaire avec l'énergie : une part importante reste résiduelle.

Limites : la corrélation ne signifie pas la causalité. La relation peut être influencée par des effets de taille (population), la structure productive, des valeurs extrêmes, ou des différences de mesure (dollars courants). Des modèles log-log, des indicateurs par habitant, ou l'ajout de covaria blés peuvent affiner l'analyse.

2.5 Bonus : généralisation 1962-2022

On répète le calcul de régression/corrélation pour chaque année disponible (1962-2022) afin d'observer l'évolution de r et de R^2. Cela permet de tester la stabilité du lien dans le temps, en gardant a l'esprit le risque d'autocorrélation temporelle.

Tableau 2 - Extrait (10 dernières années) des régressions annuelles.

Annee	n	Slope	r	R2	pvalue
2013	147	4.94	0.892	0.796	6.44e-52
2014	147	5.09	0.897	0.804	3.46e-53
2015	145	5.23	0.901	0.812	1.06e-53
2016	144	5.37	0.896	0.803	5.71e-52
2017	144	5.56	0.900	0.810	4.75e-53
2018	144	5.76	0.906	0.821	7.05e-55
2019	143	5.76	0.899	0.808	2.21e-52
2020	143	5.7	0.885	0.782	1.49e-48
2021	142	6.2	0.900	0.810	2.73e-52
2022	142	6.34	0.887	0.786	1.03e-48

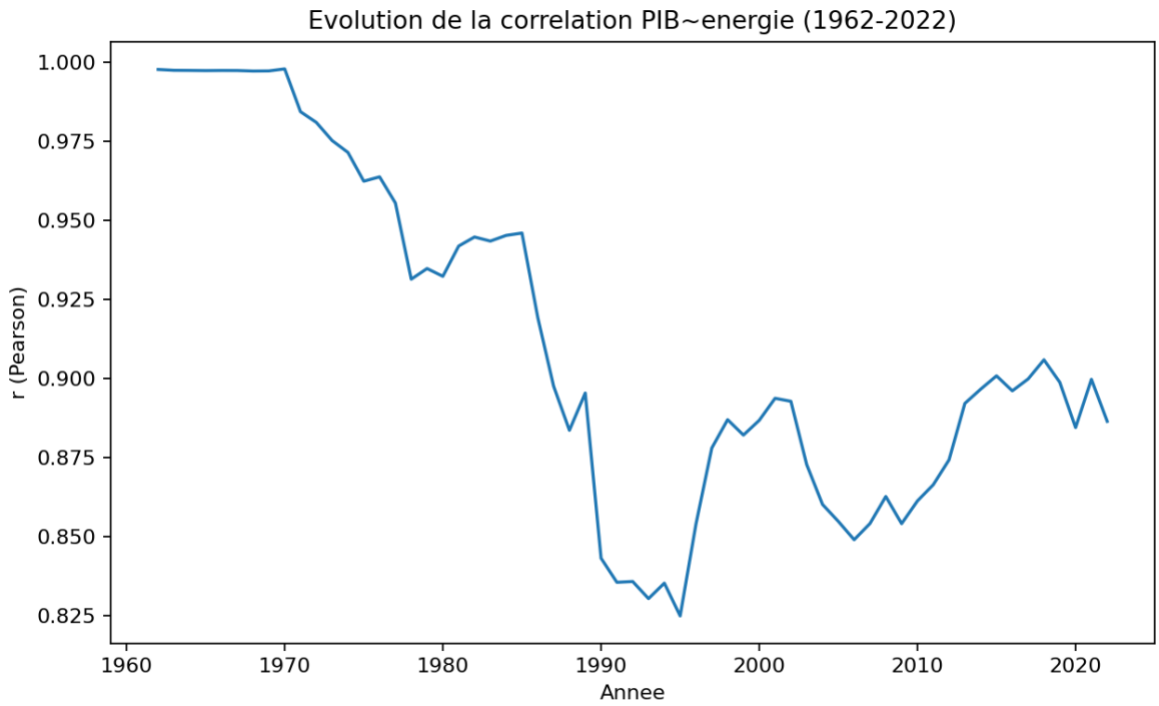

Figure 2 - Evolution de r (1962-2022).

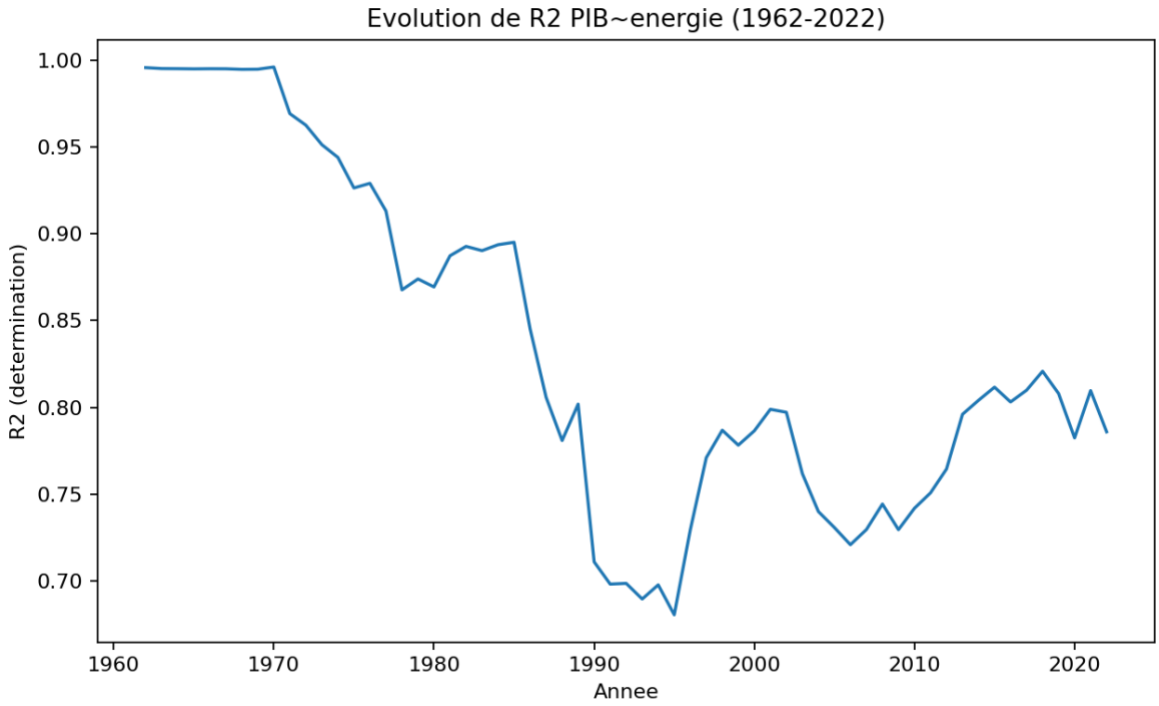

Figure 3 - Évolution de R2 (1962-2022).

Annexe mathématique (bonus)

Moindres carres (droite) : minimiser J(a,b)=sum_i (y_i-(a x_i+b))^2.
Estimations : a=cov(x,y)/var(x) ; b=y_bar-a x_bar.
Pearson : r=cov(x,y)/(s_x s_y).
Determination: R2=r^2 (simple regression).

Graphique de synthèse – Droite de régression (PIB expliqué par l'énergie, 2022)

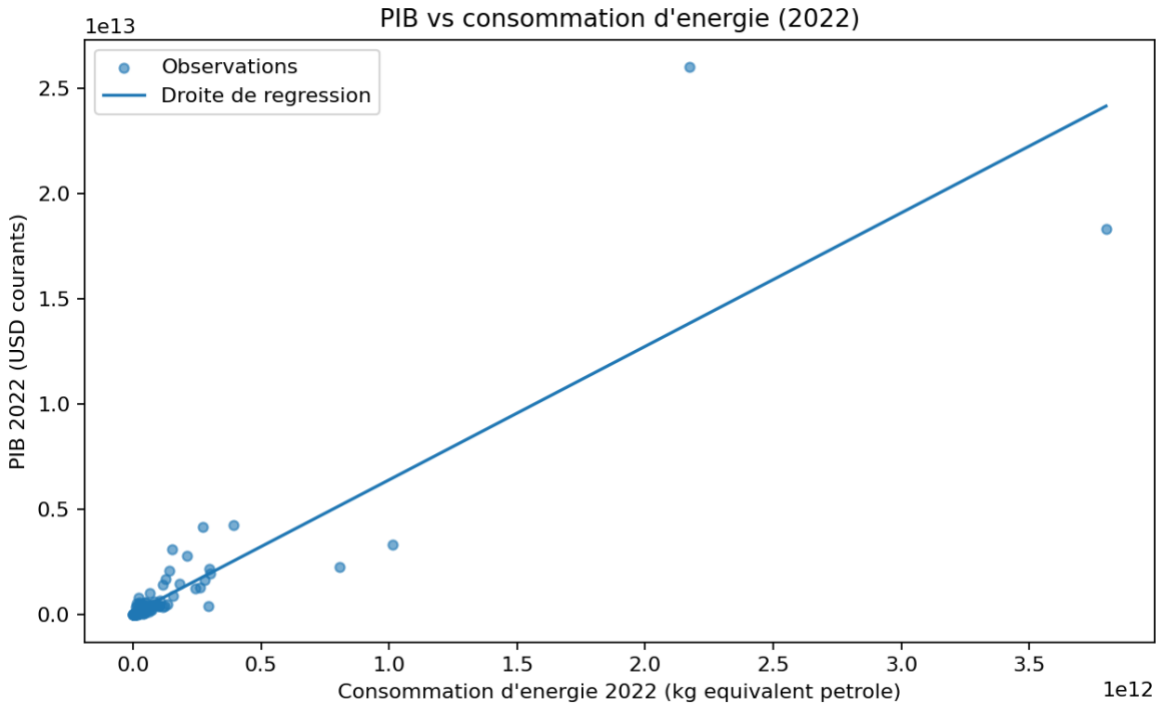

Figure – Nuage de points PIB (ordonnée) vs consommation d'énergie (abscisse) en 2022, avec droite de régression linéaire estimée.

<u>Partie 2</u>

Corrélation et régression linéaire : interprétation théorique à partir des figures

2. Analyse théorique de la corrélation et de la régression (à partir des figures)

Cette partie s'appuie sur l'ensemble des figures fournies afin d'expliquer visuellement et conceptuellement les notions de corrélation, de régression linéaire, de variance expliquée et de résidus. Les graphiques permettent de comprendre les fondements statistiques mobilisés dans l'analyse Python du lien entre PIB et consommation d'énergie.

2.1 Corrélation et forme du nuage de points

Les premières figures illustrent différents types de corrélation entre deux variables X et Y. En cas de corrélation nulle, le nuage de points est diffus et ne présente aucune orientation privilégiée : la connaissance de X n'apporte aucune information sur Y. Lorsque la corrélation est linéaire positive, les points s'alignent globalement selon une droite croissante : les valeurs élevées de X sont associées à des valeurs élevées de Y. À l'inverse, une corrélation linéaire négative se traduit par une droite décroissante : plus X augmente, plus Y diminue.

La figure du nuage bivarié avec ellipse de concentration ($\rho = 0{,}70$) montre que la corrélation se lit aussi par la forme et l'orientation de l'ellipse : plus l'ellipse est allongée et orientée diagonalement, plus la corrélation est forte. Lorsque l'ellipse est presque circulaire, la corrélation est faible ou nulle.

2.2 Centrage des données et lecture des quadrants

Le nuage centré autour des moyennes $\bar{X}$ et $\bar{Y}$ permet d'interpréter la corrélation en termes de signes des écarts à la moyenne. Les observations situées dans les quadrants I et III (écarts de même signe) contribuent positivement à la corrélation, tandis que celles situées dans les quadrants II et IV (écarts de signe opposé) contribuent négativement. La corrélation résulte donc d'un équilibre entre ces contributions.

2.3 Principe de la régression linéaire et méthode des moindres carrés

La régression linéaire vise à résumer la relation entre X et Y par une droite $\hat{Y} = aX + b$. La méthode des moindres carrés consiste à choisir les paramètres a et b qui minimisent la somme des carrés des écarts verticaux entre les valeurs observées Y et les valeurs prédites $\hat{Y}$. Chaque résidu correspond à un écart vertical entre un point observé et la droite de régression.

Les figures illustrent que la droite de régression ne passe pas nécessairement par tous les points, mais qu'elle représente une tendance moyenne conditionnelle de Y sachant X. Les résidus permettent d'évaluer la qualité de l'ajustement et de repérer des observations atypiques.

2.4 Variance totale, variance expliquée et variance résiduelle

Avant la régression, la variance totale de Y mesure la dispersion des observations autour de la moyenne $\bar{Y}$. La régression permet de décomposer cette variance en deux parties : la variance expliquée par le modèle (liée à la relation avec X) et la variance résiduelle (dispersion restante autour de la droite de régression).

Le coefficient de détermination R^2 correspond à la part de la variance totale de Y expliquée par la régression. Graphiquement, plus les points sont proches de la droite, plus la variance résiduelle est faible et plus R^2 est élevé. Ainsi, la régression permet de réduire l'incertitude sur Y en utilisant l'information contenue dans X.

2.5 Les deux droites de régression et l'axe principal

Les figures finales montrent qu'il existe en réalité deux droites de régression : la droite de Y en X, qui minimise les écarts verticaux, et la droite de X en Y, qui minimise les écarts horizontaux. Ces deux droites sont différentes sauf en cas de corrélation parfaite.

L’axe principal de l’ellipse de concentration correspond à la direction de dispersion maximale du nuage. Il est lié à l’analyse en composantes principales et représente une relation symétrique entre X et Y, contrairement aux droites de régression qui supposent une variable expliquée et une variable explicative. Cette distinction rappelle que la corrélation ne suffit pas à établir une causalité.

Dans le cadre de l’analyse PIB–énergie, la droite de régression de Y en X est utilisée pour prédire le PIB à partir de la consommation d’énergie, mais cette prédiction doit être interprétée comme une relation statistique moyenne et npon comme une relation causale directe.

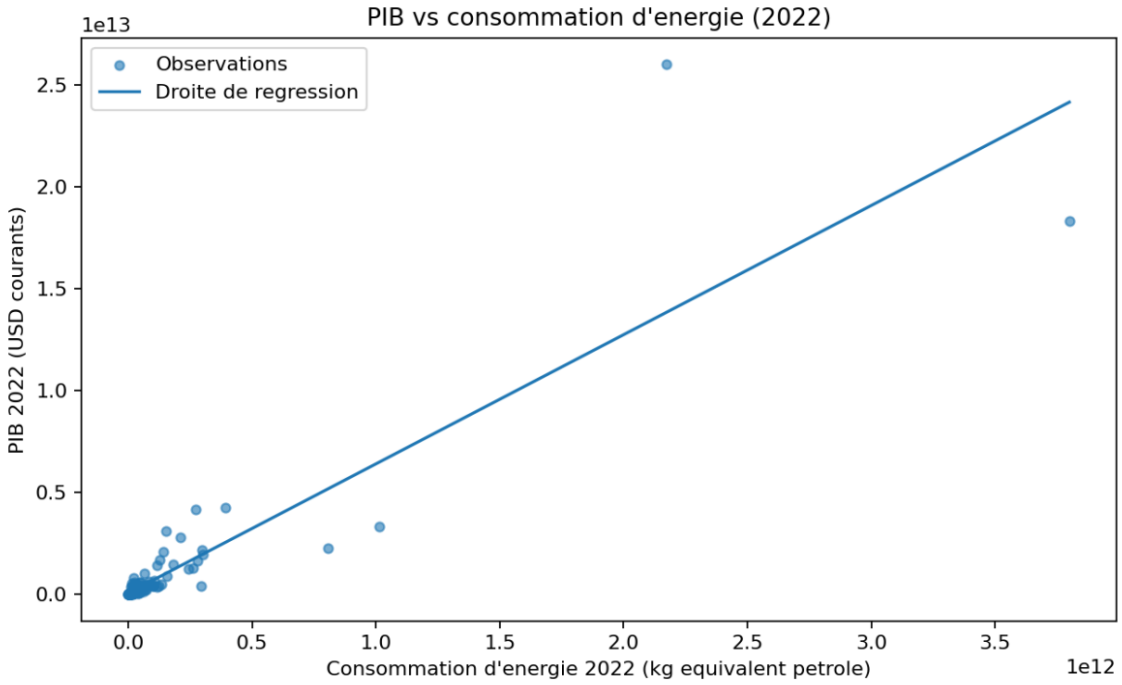

Figure – PIB vs consommation d'énergie (2022) avec droite de régression.