Facteurs socio-économiques de la scolarisation des filles en Afrique centrale : Régression OLS / Clustering K-Means / ACP

1. Données et méthodologie

J'ai construit un panel avec des données UNESCO et Banque Mondiale pour 9 pays d'afrique centrale sur la période 2008-2024. Le plus difficile dans ce travail était la préparation des données. Les données Banque Mondiale sont en format large (une colonne par année). J'ai dû les transformer en format long pour pouvoir faire les fusions avec les données UNESCO. Il y avait aussi beaucoup de valeurs manquantes qu'il fallait gérer. Après nettoyage, j'avais 129 observations sur 9 pays.

Variable dépendante

- Taux de completion secondaire supérieur filles (UNESCO, indicateur CR.MOD.3.F) Variables explicatives (Banque Mondiale)
- Log PIB/habitant en parité de pouvoir d'achat (NY.GDP.PCAP.PP.CD)
- Pourcentage population rurale (SP.RUR.TOTL.ZS)
- Taux de fécondité adolescente naissances / 1000 femmes 15-19 ans (SP.ADO.TFRT)
- Dépenses publiques d'éducation en % du PIB (SE.XPD.TOTL.GD.ZS)
- Taux de croissance de la population (SP.POP.GROW)

Méthodes utilisées

- Régression OLS (statsmodels) : modèle linéaire pour estimer les effets de chaque variable
- K-Means k=3 (scikit-learn) : regroupement des pays selon leurs profils éducatifs et économiques
- ACP 2 composantes : réduction de dimension pour visualiser les clusters

Pays inclus dans le panel : Burundi, Cameroun, Centrafrique, Congo, Gabon, RD Congo, Rwanda, Sao Tomé, Tchad. 9 pays (la Guinée équatoriale a été exclue faute de données suffisantes sur les variables Banque Mondiale).

2. Résultats de la régression OLS

Le modèle explique 61,7 % de la variance du taux de complétion (R^2=0,617). La statistique F est significative (p=4.17e-24), donc le modèle dans son ensemble est utile pour expliquer les variations du taux d'achèvement des filles.

Variable	Coefficient	p-valeur	Interprétation
Log PIB/hab	+0.132	0.926	Non significatif
Pop. rurale (%)	-0.333	0.000	Barriere majeure
Fec ado.	-0.096	0.000	Grossesses précoces
Dép. éduc	+1.128	0.012	Investissement efficace

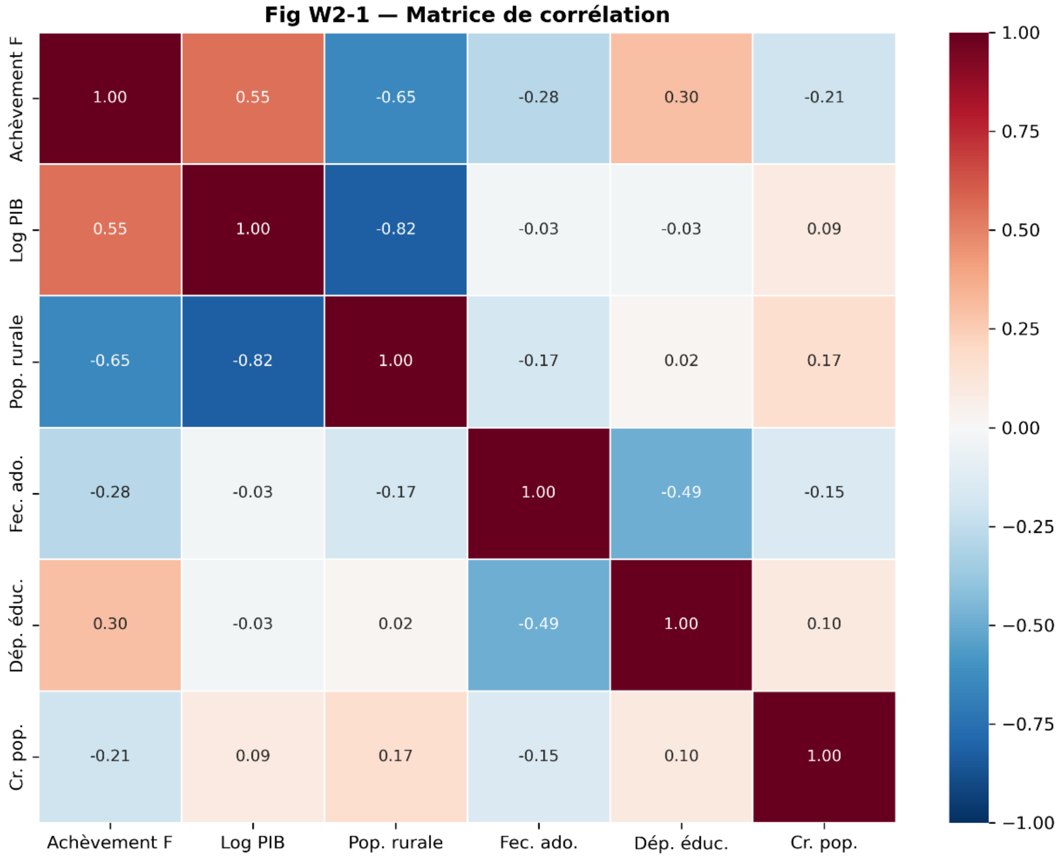
Cr pop.	-1.986	0.010	Pression sur l'offre

Le résultat qui m'a surprise : le PIB n'est pas significatif. (p = 0.926). Ce n'est pas la richesse brute qui détermine la scolarisation des filles, c'est plutôt la structure socio-démographique du pays : est-ce qu'il est rural ou urbain, est-ce qu'il y a beaucoup de grossesses précoces, est-ce qu'on investit vraiment dans l'éducation ? C'est-à-dire un pays relativement pauvre mais urbain, avec peu de grossesses adolescentes et des investissements en éducation, peut avoir de meilleurs résultats qu'un pays plus riche mais rural.

La corrélation entre Log PIB et population rurale est de -0,82. Ça veut dire que les deux variables capturent en partie la même chose. C'est pour ça que le PIB devient non significatif quand la population rurale est dans le modèle, c'est un phénomène de multicolinéarité.

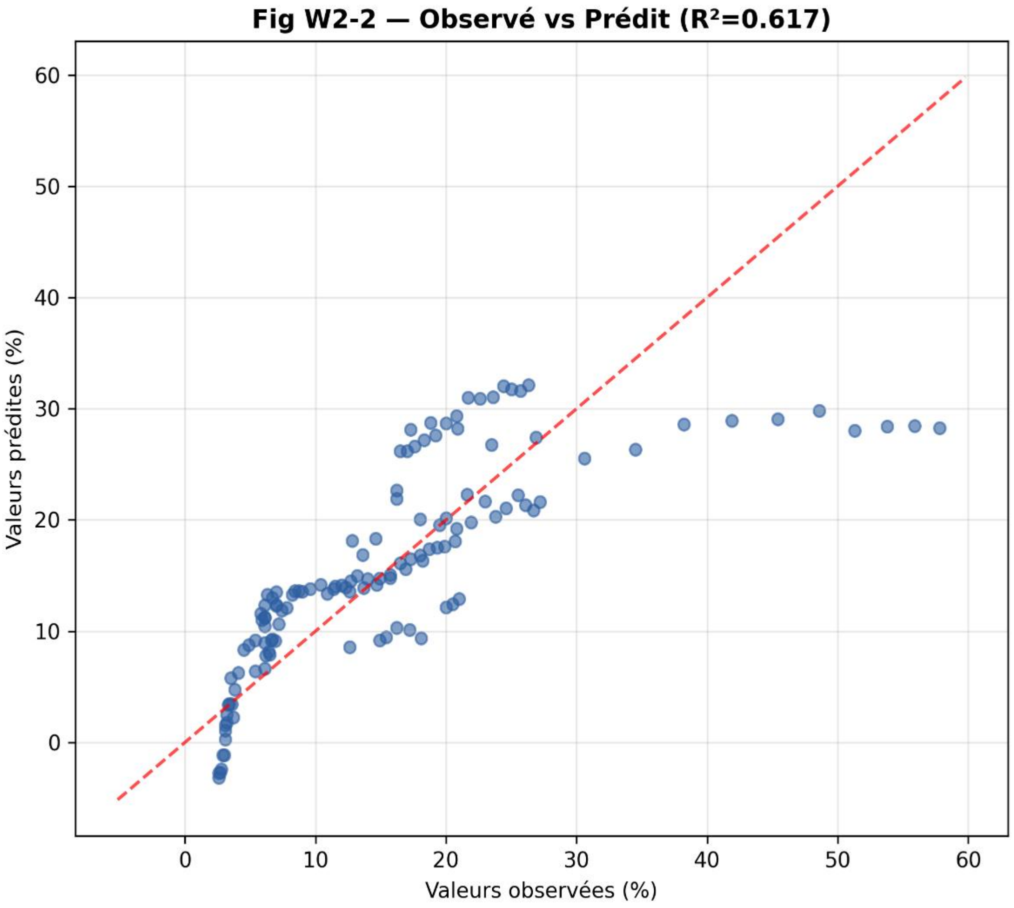
3. Figures et interprétations

Fig W2-1 — Matrice de corrélation



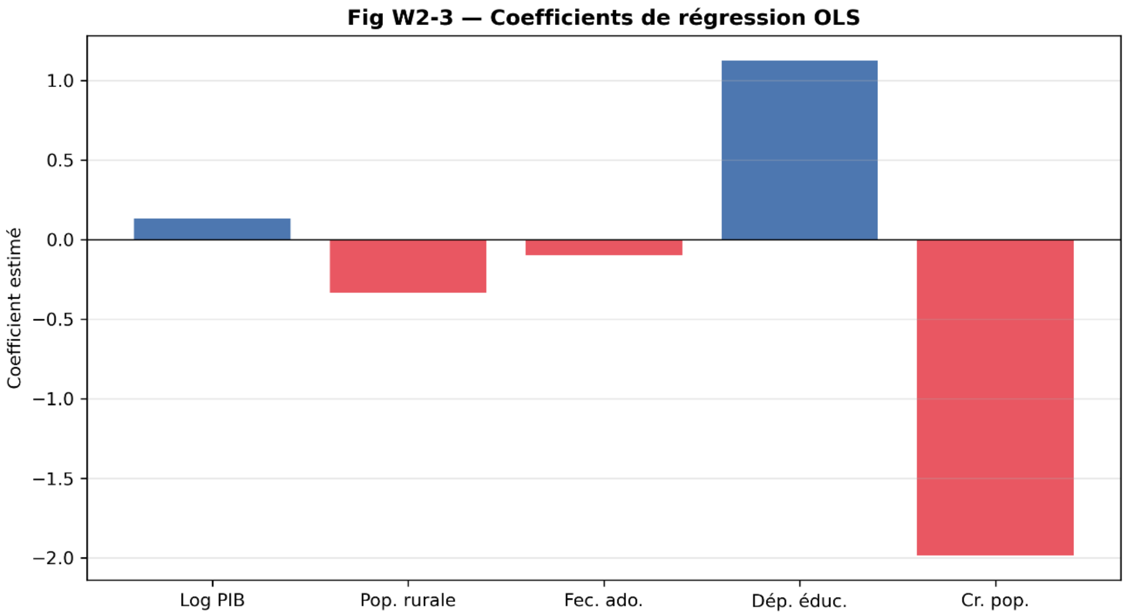
La corrélation la plus forte avec la completion est la population rurale (-0.65). Plus un pays est rural, moins les filles terminent le secondaire supérieur. La corrélation entre Log PIB et population rurale est -0.82, ce qui confirme le problème de multicolinéarité évoqué plus haut.

J'ai décidé de garder les deux variables dans le modèle pour montrer que le PIB devient non significatif quand on contrôle pour la ruralité. C'est une façon de décomposer l'effet brut du PIB en ses composantes réelles.



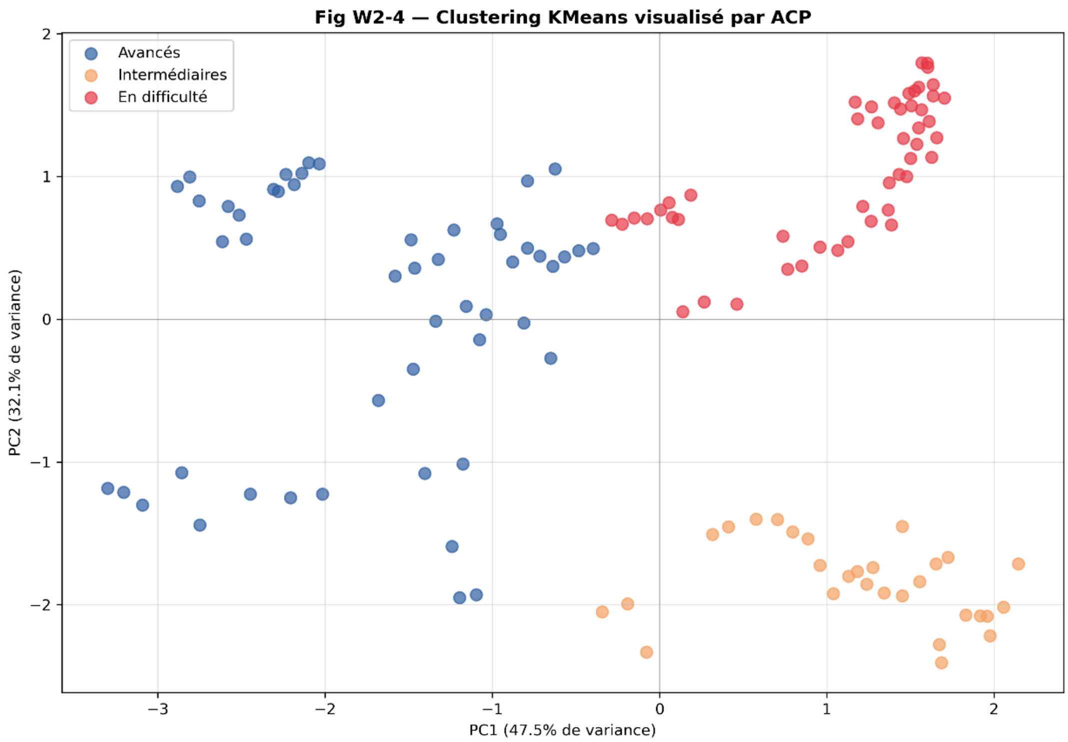
Les points proches de la ligne diagonale rouge indiquent une bonne prédiction. Pour les valeurs basses (0-15%), le modèle est assez précis. Mais pour les valeurs élevées (au-dessus de 30%), le modèle sous-estime systématiquement.

Ça veut dire qu'il y a des facteurs que je n'ai pas capturés , la qualité des enseignants, des politiques spécifiques à certains pays, la géographie, etc. Ces éléments pourraient expliquer pourquoi des pays comme Sao Tomé ou le Rwanda ont de meilleurs résultats que ce que le modèle prédit.



La croissance démographique a l'effet négatif le plus fort (-1.99). Quand la population augmente vite, l'offre scolaire ne suit pas. C'est un effet de capacité : trop d'enfants pour trop peu de salles de classe et d'enseignants.

Les dépenses d'éducation ont l'effet positif le plus fort (+1.13). Ce sont les deux leviers les plus importants selon ce modèle : réduire la croissance démographique et investir davantage dans l'éducation. Ces deux résultats sont cohérents avec la littérature économique sur le sujet. Le PIB (bleu, +0.13) est proche de zero et non significatif.



J'ai choisi k=3 parce que ça donnait des groupes interprétables. Les deux composantes principales expliquent 79.6% de la variance totale, ce qui est suffisant pour une visualisation fiable.

- Cluster bleu (Avancés) : Sao Tomé, Congo, Gabon, Cameroun, meilleurs résultats éducatifs, PIB plus élevé, population moins rurale
- Cluster orange (Intermédiaires) : Burundi, Rwanda , situation économique plus difficile mais dynamiques éducatives positives
- Cluster rouge (En difficulté) : Centrafrique, Tchad, RD Congo , fécondité adolescente élevée, fort taux de ruralité, achèvement filles très bas

Le score silhouette est de 0.459, ce qui indique une séparation modérée. C'est cohérent : ces pays ont des histoires et des géographies communes, donc il est normal qu'ils ne soient pas très bien séparés. Les frontières entre clusters sont floues, pas nettes.

4. Limites et ce que ce travail m'a appris

Limites du modèle OLS

- Causalité inverse possible : les pays qui scolarisent mieux peuvent aussi dépenser plus en éducation. Pour vraiment répondre à cette question il faudrait des variables instrumentales.
- Effets fixes pays non inclus : des caractéristiques propres à chaque pays (histoire, culture, géographie) ne sont pas capturées par le modèle. Un modèle à effets fixes permettrait de mieux contrôler ça.
- Multicolinéarité : la corrélation de -0.82 entre Log PIB et population rurale pose un problème d'interprétation des coefficients individuels.

Ce que j'ai appris

La différence entre corrélation et causalité n'est pas juste une formule qu'on apprend en cours. C'est une vraie contrainte qui change l'interprétation de chaque résultat. Et les données manquantes ne sont pas qu'un problème technique.

Sources de données :

- UNESCO Institute for Statistics
- Banque Mondiale — World Development Indicators

Outils

Python Jupyter · statsmodels, scikit-learn, seaborn