Économétrie

LEMKHAYER Imane - YILMAZ Ilayda - MOUSALLI Hafsa

18 January 2025

1 Présentation de la question de recherche

Dans ce projet, nous cherchons à répondre à la question de recherche suivante : **Quels sont les déterminants des ventes de cigares et comment les caractéristiques économiques, démographiques et les variations temporelles influencent-elles ces ventes ?**

L'objectif principal est de comprendre les principaux facteurs influençant les ventes de cigares et d'évaluer l'importance relative de chacun. Cette analyse permettra de mieux appréhender les comportements de consommation, les impacts des politiques tarifaires, et les dynamiques économiques et démographiques à travers les régions et les années.

Pour répondre à cette question, nous utilisons la base de données *cigar* et appliquons plusieurs modèles économétriques, notamment le modèle poolé, le modèle à effets fixes et le modèle à effets aléatoires. Cette approche méthodologique nous permettra de mettre en évidence les contributions respectives des facteurs économiques (prix, taxes, revenus), des caractéristiques démographiques (population) et des tendances temporelles sur les ventes de cigares.

1.1 Hypothèses de recherhe

- **H1 :** Une augmentation du prix moyen des cigares (*price*) réduit les ventes totales (*sales*) en raison de la sensibilité au prix des consommateurs. En effet, plus le prix est élevé, moins de consommateurs seront incités à consommer.
- **H2 :** Une population totale plus importante (*pop*) entraîne une augmentation des ventes totales, reflétant une demande agrégée plus élevée.
- **H3 :** La proportion de la population âgée de 16 ans et plus (*pop16*) a un effet positif sur les ventes, cette tranche d'âge étant plus susceptible de consommer des cigares.
- **H4 :** Une hausse du revenu national disponible (*ndi*) accroît les ventes.
- **H5 :** Une augmentation de l'indice des prix à la consommation (*cpi*) peut réduire les ventes, reflétant une diminution du pouvoir d'achat des consommateurs.
- **H6 :** La proportion de la population consommant des cigares (*pimin*) influence positivement les ventes totales.
- **H7** : Les caractéristiques géographiques influencent les ventes de cigares, par exemple, des ventes plus élevées dans les grandes villes.

2 Présentation des données

Notre base contient 1380 observations et 9 variables.

- state : Région ou localisation géographique
- year : Année de l'observation. Cette variable nous permet de capturer les variations temporelles (1963-1992).
- price : Prix moyen des cigares, en dollars américains par paquet
- pop : Population totale, en millions d'habitants.
- pop16 : Proportion de la population dans la tranche d'âge 16 ans et plus
- cpi : Indice des prix à la consommation
- ndi : Revenu national disponible en milliards de dollars américains
- sales : la valeur totale des ventes de cigares en millions de paquets.
- pimin : proportion de la population qui consomme des cigares ou produits similaires exprimée en pourcentage.

3 Statistiques descriptives

Dans un premier temps, nous avons fait un code qui calcule les statistiques descriptives (minimum, moyenne, maximum, médiane, écart-type) pour toutes les variables quantitatives, à l'exception des colonnes 'state' et 'year'. Puis les résultats sont affichés sous forme de tableau.

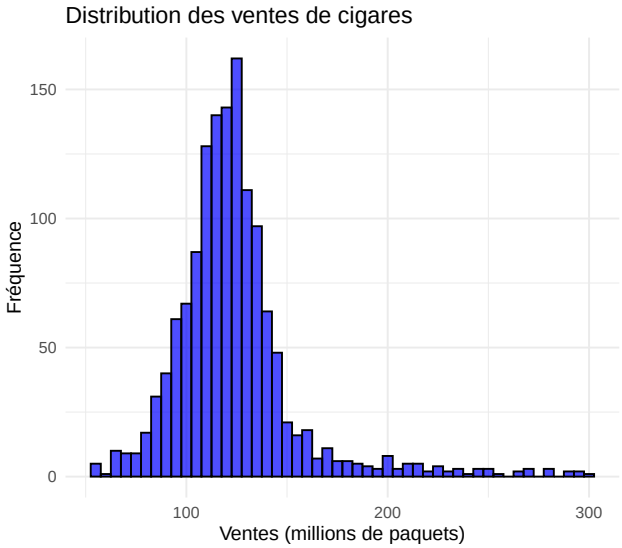
3.1 Fonction

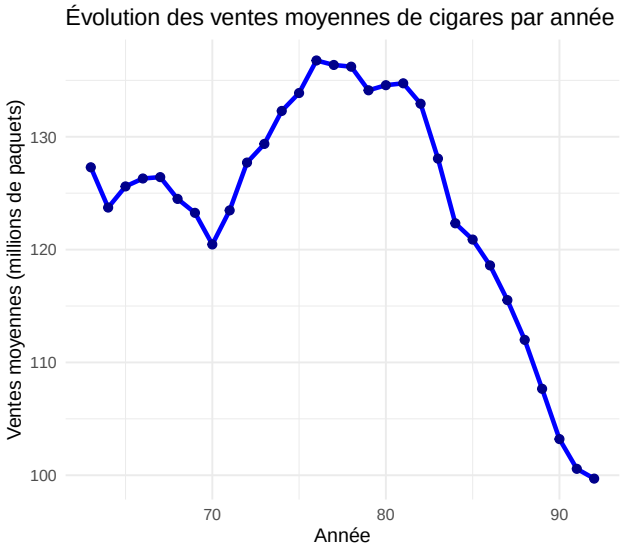
```
##           Min    Mean     Max Median      Sd
## price   23.40   68.70   201.9   52.3   41.99
## pop    319.00 4537.11 30703.3 3174.0 4828.84
## pop16  215.20 3366.62 22920.0 2315.3 3641.85
## cpi     30.60   73.60   140.3   62.9   36.53
## ndi   1322.57 7525.02 23074.0 6281.2 4747.86
## sales   53.40  123.95   297.9  121.2   30.99
## pimin   23.40   62.90   178.5   46.4   38.32
```

Interpretation :

- price : le prix varie de 23,4 à 201,9, ce qui pourrait influencer la consommation de manière significative.
- cpi : on a une moyenne de 73,6, un minimum de 30,6 et un maximum de 140,3. on peut voir qu'il y a une augmentation qui reflète l'inflation au cours des années.
- ndi : on a une moyenne de 7 525, avec un maximum de 23 074. Cette variable est importante pour analyser la demande, car elle représente le pouvoir d'achat.
- sales : Varie de 53,4 à 297,9 qui pourrait refléter des différences significatives de consommation entre les États ou les années.

3.2 Visualisation





Distribution des ventes de cigares :

- On observe une distribution asymétrique à droite, ce qui signifie qu'une minorité des observations représente des ventes très élevées.
- Cela peut refléter des différences entre les États en termes de consommation.

Évolution des ventes moyennes de cigares par année :

- Le graphique montre une tendance générale : les ventes augmentent dans les années 70, atteignent un pic, puis diminuent fortement dans les années 80 et 90.
- Cette diminution peut être due à des politiques publiques (taxes) ou à un changement de comportement des consommateurs.

4 Présentation des différentes Modèles

4.1 Modèle poolé

— Le modèle Pooled OLS considère toutes les observations comme une seule entité sans distinction entre les unités temporelles et géographiques. Il ne prend pas en compte les effets fixes ou aléatoires.

```
## Pooling Model
##
## Call:
## plm(formula = sales ~ price + pop + pop16 + cpi + ndi + pimin,
##     data = Cigar, effect = "individual", model = "pooling", index = c("state",
##         "year"))
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Residuals:
##     Min.  1st Qu.   Median  3rd Qu.     Max.
## -63.1547 -15.3291  -2.3784   9.2279 159.2068
##
## Coefficients:
##                Estimate  Std. Error  t-value  Pr(>|t|)
## (Intercept)  1.3862e+02  2.2470e+00  61.6920 < 2.2e-16 ***
## price       -1.5276e+00  1.1532e-01 -13.2466 < 2.2e-16 ***
## pop         -2.5018e-03  2.7692e-03  -0.9034   0.36645
## pop16        2.4162e-03  3.7002e-03   0.6530   0.51388
## cpi          1.4430e-01  8.1220e-02   1.7767   0.07584 .
## ndi          6.0417e-03  5.9718e-04  10.1171 < 2.2e-16 ***
## pimin        5.9477e-01  1.2745e-01   4.6667  3.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    1324500
## Residual Sum of Squares: 998450
## R-Squared:      0.24614
## Adj. R-Squared: 0.24285
## F-statistic: 74.7167 on 6 and 1373 DF, p-value: < 2.22e-16
```

Dans un premier temps, nous allons interpréter les coefficients des variables pour comprendre comment elles influencent les ventes de cigares.

— **Price** : Lorsque le prix moyen des cigares augmente de 1 dollar par paquet, en moyenne les ventes diminuent de 1.5276 millions de paquets.
— **NDI** : Lorsque le revenu national disponible augmente de 1 milliard de dollars, en moyenne les ventes augmentent de 0.00604 millions de paquets.
— **Pimin** : Lorsque la proportion minimale de la population consommatrice de cigares augmente de 1%, en moyenne les ventes augmentent de 0.59 millions de paquets.

Ensuite, nous examinons la p-value pour déterminer si la variable est significative, ce qui permet d'évaluer son impact réel dans le modèle.

— Dans notre modèle, **price**, **ndi** et **pimin** sont significative à 1%, ce qui confirme qu'elles ont un effet important sur les ventes.
— En revanche, **pop, pop16 et cpi** ne sont pas significatives, ce qui suggère qu'elles n'affectent pas directement les ventes dans ce modèle

Autres observations :

— **R2=0.24614**, il s'agit du ratio de la variance expliquée sur la variance totale. Plus il est proche de 1, plus il indique un bon ajustement

Dans notre cas, le modèle explique environ 24.6% de la variance des ventes de cigares. Ce score est relativement faible, indiquant que d'autres facteurs non inclus dans le modèle pourraient influencer les ventes.

— **R2ajusté= 0.2485**
— **F-statistic : 74.7167** : Le modèle dans son ensemble est statistiquement significatif car p=2.22e−16 est extrêment faible . La statistique de Fisher mesure la validité globale du modèle, c'est-à-dire si les variables indépendantes expliquent bien la variance de la variable dépendante par rapport à l'erreur aléatoire.

Conclusion :

— L'analyse des ventes de cigares en utilisant le modèle poolé nous a permis de mettre en avant les variables importantes et leur impact sur la demande de produits du tabac.

Le prix des cigares a un effet négatif et significatif sur les ventes ce qui montre l'importance du prix dans le processus d'achat. En effet, plus le prix du paquet de cigares augmente, moins les consommateurs en achèteront.

La proportion de la population âgée de 16 ans et plus ainsi que **le revenu national disponible** ont un impact positif sur les ventes de cigares. En effet, un revenu national disponible plus élevé augmente le pouvoir d'achat des consommateurs et leur permet l'achat des cigares. Une proportion plus importante de la population âgée de 16 ans et plus augmente le nombre potentiel d'acheteurs, car c'est généralement à partir de cet âge que l'on commence à fumer.

Toutefois, on a pu remarquer que la taille de la population totale (pop) n'était pas statistiquement significative. Cela peut surprendre car on pourrait penser qu'une population plus grande favorise les ventes de cigares. De plus, le fait que la population de 16 ans et plus soit significative montre que ce groupe a plus d'impact, tandis que les autres tranches d'âge semblent moins pertinentes dans ce modèle.

— Le faible R2 de 0,24614 s'explique par le fait que le modèle poolé ne capture pas toutes les variations spécifiques aux régions ou aux années, et par l'absence de variables potentiellement influentes comme le taux de taxation sur les cigares. Une taxation plus élevée augmente le prix des cigares, ce qui peut réduire leur consommation et donc les ventes. À l'inverse, une taxation faible pourrait encourager l'achat. Le niveau d'éducation est une variable omise, un niveau d'éducation plus élevé est souvent associé à une meilleure sensibilisation aux risques du tabac, ce qui pourrait réduire la consommation et donc les ventes de cigares.

4.2 Modèle à effets fixes (Within)

— Les différences non observées entre les individus (hétérogénéité individuelle) sont corrélées avec les variables explicatives.

```
## Oneway (individual) effect Within Model
##
## Call:
## plm(formula = sales ~ price + pop + pop16 + cpi + ndi + pimin -
##     1, data = Cigar, model = "within", index = c("state", "year"))
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Residuals:
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -57.81439  -6.16269  -0.43749   6.23805 113.91621
##
## Coefficients:
##          Estimate  Std. Error  t-value  Pr(>|t|)
## price -0.64685045  0.07529199  -8.5912 < 2.2e-16 ***
## pop   -0.00642573  0.00194877  -3.2973 0.0010019 **
## pop16  0.00862358  0.00228870   3.7679 0.0001718 ***
## cpi    0.93954489  0.05146855  18.2547 < 2.2e-16 ***
## ndi   -0.00532909  0.00047157 -11.3007 < 2.2e-16 ***
## pimin  0.21836004  0.08612167   2.5355 0.0113431 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    412530
## Residual Sum of Squares: 229940
## R-Squared:      0.4426
## Adj. R-Squared: 0.42119
## F-statistic: 175.748 on 6 and 1328 DF, p-value: < 2.22e-16
```

Dans ce modèle, la constante a été enlevé.

— **Price** : Lorsque le prix moyen des cigares augmente de 1 dollar par paquet, en moyenne les ventes diminuent de 0.647 millions de paquets.
— **Pop** : Lorsque la population totale augmente de 1 million d'habitants, en moyenne les ventes diminuent de 0.00643 millions de paquets.
— **Pop16** : Lorsque la proportion de la population âgée de 16 ans et plus augmente de 1%, en moyenne les ventes augmentent de 0.00862 millions de paquets.
— **CPI** : Lorsque l'indice des prix à la consommation augmente de 1 unité, en moyenne les ventes augmentent de 0.9395 millions de paquets.
— **NDI** : Lorsque le revenu national disponible augmente de 1 milliard de dollars, en moyenne les ventes diminuent de 0.00533 millions de paquets.
— **Pimin** : Lorsque la proportion minimale de la population consommatrice de cigares augmente de 1%, en moyenne les ventes augmentent de 0.21836 millions de paquets.
— Dans notre modèle, **toutes les variables sont significatives**, ce qui confirme qu'elles ont un effet important sur les ventes.

Autres observations :

— **R2=0.4426**, il s'agit du ratio de la variance expliquee sur la variance totale. Plus il est proche de 1, plus il indique un bon ajustement

Dans notre cas, le modèle explique environ 44.26% de la variance des ventes de cigares. Ce score est relati-

vement faible, indiquant que d'autres facteurs non inclus dans le modèle pourraient influencer les ventes.

- **R2ajusté= 0.42119**
- **F-statistic : 175.748** : Le modèle dans son ensemble est statistiquement significatif car p=2.22e−16 est extrêment faible .

En conclusion :

- Le modèle à effets fixes (EF) nous a permis d'obtenir une meilleure explication des ventes de cigares par rapport au modèle poolé, avec un R2 de 44.26%, ce qui indique que le modèle capte une proportion plus élevée de la variance des ventes.
- A l'inverse du modèle poolé pour qui seulement les variables : price, ndi et pimin étaient signifivatives, dans le modèle à EF, toutes nos variables sont désormais signifivatives. En contrôlant les différences spécifiques entre les observations, par exemple, les caractéristiques uniques des régions, les relations entre ces variables et la variable dépendante sont devenues plus claires et plus fiables. Le modèle à effets fixes a permis d'identifier des relations qui étaient masquées dans le modèle poolé en raison de l'omission de ces différences spécifiques
- On observe que la statistique de Fisher est plus élevée dans le modèle à effets fixes par rapport au modèle poolé, ce qui indique que le modèle à effets fixes est mieux adapté pour expliquer les variations des ventes de cigares. En effet, en tenant compte des différences spécifiques entre les individus (ou les unités, comme les régions ou les années), le modèle à effets fixes permet de capturer ces variations individuelles, ce qui améliore l'ajustement global du modèle. En revanche, le modèle poolé suppose que toutes les observations partagent la même structure, sans distinguer les différences entre les unités temporelles et géographiques, ce qui peut masquer des facteurs importants. Ainsi, le modèle à effets fixes, en prenant en compte ces spécificités, offre une explication plus précise des données et se traduit par une statistique F plus élevée.
- Ce modèle laisse supposer que d'autres facteurs, comme les politiques fiscales, les habitudes de consommation ou les campagnes de prévention sur le tabac, pourraient enrichir l'analyse et améliorer l'ajustement du modèle.
- Dans l'ensemble, bien que le modèle à effets fixes offre une meilleure explication que le modèle poolé, il reste insuffisant pour capturer l'intégralité des facteurs influençant la demande de cigares.

4.3 Modèle à effet aléatoire

— Le modèle Random Effects suppose que les variations entre unités sont aléatoires et indépendantes des autres unités.

```
## Oneway (individual) effect Random Effect Model
##    (Swamy-Arora's transformation)
##
## Call:
## plm(formula = sales ~ price + pop + pop16 + cpi + ndi + pimin -
##     1, data = Cigar, model = "random", index = c("state", "year"))
##
## Balanced Panel: n = 46, T = 30, N = 1380
##
## Effects:
##                  var std.dev share
## idiosyncratic 173.15   13.16 0.223
## individual    603.86   24.57 0.777
## theta: 0.9027
##
## Residuals:
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -46.99    1.30    7.93    9.14   15.58  130.22
##
## Coefficients:
##          Estimate  Std. Error  z-value  Pr(>|z|)
## price -0.69604738  0.09650655  -7.2124 5.496e-13 ***
## pop    0.01154438  0.00224182   5.1496 2.611e-07 ***
## pop16 -0.00803060  0.00279296  -2.8753  0.004036 **
## cpi    1.23182999  0.06458501  19.0730 < 2.2e-16 ***
## ndi   -0.00623541  0.00059987 -10.3946 < 2.2e-16 ***
## pimin  0.13871907  0.11036544   1.2569  0.208787
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    421160
## Residual Sum of Squares: 394600
## R-Squared:      0.3519
## Adj. R-Squared: 0.34954
## Chisq: 791.434 on 6 DF, p-value: < 2.22e-16
```

— **Price** : Lorsque le prix moyen des cigares augmente de 1 dollar par paquet, en moyenne les ventes diminuent de 0.69 millions de paquets.
— **Pop** : Lorsque la population totale augmente de 1 million d'habitants, en moyenne les ventes augmentent de 0.0115 millions de paquets.
— **Pop16** : Lorsque la proportion de la population âgée de 16 ans et plus augmente de 1%, en moyenne les ventes diminuent de 0.008 millions de paquets.
— **CPI** : Lorsque l'indice des prix à la consommation augmente de 1 unité, en moyenne les ventes augmentent de 1.23 millions de paquets.
— **NDI** : Lorsque le revenu national disponible augmente de 1 milliard de dollars, en moyenne les ventes diminuent de 0.0062 millions de paquets.

Autres observations :

— **R2=0.3519**, il s'agit du ratio de la variance expliquee sur la variance totale. Plus il est proche de 1, plus il indique un bon ajustement

Dans notre cas, le modèle explique environ 35.19% de la variance des ventes de cigares. Ce score est relativement faible, indiquant que d'autres facteurs non inclus dans le modèle pourraient influencer les ventes.

— **R2ajusté= 0.34954**

En conclusion :

— Le modèle à effets aléatoires montre que toutes les variables, sauf pimin ont un effet significatif sur les ventes de cigares. Le R2 de 0.3519 montre que le modèle n'explique qu'environ 35.19% de la variance des ventes, ce qui indique que de nombreux facteurs influençant la demande ne sont pas inclus dans l'analyse.
— En somme, bien que ce modèle apporte des informations utiles, il reste insuffisant pour capturer toute la complexité de la demande de cigares et nécessite une élargissement des variables prises en compte pour une meilleure compréhension.
— On peut également noter que, dans le modèle EA, certaines variables présentent des coefficients plus faibles que dans le modèle à EF.

5 Choix du modèle préféré et sa justification

5.1 Test de Hausman

```
##
##  Hausman Test
##
## data:  sales ~ price + pop + pop16 + cpi + ndi + pimin - 1
## chisq = 1340.5, df = 6, p-value < 2.2e-16
## alternative hypothesis: one model is inconsistent
```

Ho = les effets individuels ci ne sont pas corrélés aux variables explicatives

H1 = les effets individuels ci sont corrélés aux variables explicatives

Si on rejette H0, alors il faut préférer des modèles Within (différences premieres ou effets fixes)

Si on ne rejette pas H0, alors il faut préférer des modèles à effets aléatoires (random- fixed effects).

5.2 Test de Fisher

```
##
##  F test for individual effects
##
## data:  sales ~ price + pop + pop16 + cpi + ndi + pimin - 1
## F = 98.631, df1 = 45, df2 = 1328, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

H0 = pas d'effets de panel, tous les ci sont = 0

H1 = effets de panels à prendre en compte, au moins un ci est différent de 0

Si on rejette H0, alors il faut préférer un modele à effets fixes Si on ne rejette pas H0, alors il faut préférer des modèles MCO poolés.

5.3 Test de Breusch-Pagan

```
##
##  Lagrange Multiplier Test - (Breusch-Pagan)
##
## data:  sales ~ price + pop + pop16 + cpi + ndi + pimin
## chisq = 8310.8, df = 1, p-value < 2.2e-16
```

```
## alternative hypothesis: significant effects
```

H0 = pas d'effets de panel, tous les individus ont le même ci

H1 = effets de panels à prendre en compte

Si on rejette H0, alors il faut preferer un modele a effets aleatoires. Si on ne rejette pas H0, alors il faut preferer des modeles MCO pooles.

6 Conclusion

Le projet avait pour objectif de déterminer le modèle économétrique le plus adapté pour analyser les ventes de cigares en fonction des variables explicatives. Trois modèles ont été comparés :

- le modèle poolé : ce modèle regroupe toutes les observations sans tenir compte des différences entre les unités (comme les régions ou les années). Il repose sur l'hypothèse que les relations entre les variables explicatives et la variable dépendante sont identiques pour tous les individus et toutes les périodes,
- le modèle à effets fixes (EF) :il considère les spécificités des unités comme des constantes pouvant être corrélées aux variables explicatives. Il élimine les biais potentiels en capturant les hétérogénéités propres à chaque unité ou période
- le modèle à effets aléatoires (EA) : il suppose que les différences spécifiques entre unités sont aléatoires et non corrélées aux variables explicatives, ce qui permet d'estimer des effets communs tout en conservant les spécificités des unités comme une composante aléatoire

Nous avons réalisé divers tests qui nous ont permis de guider ce choix. Premièrement, le test de Hausman pour déterminer si les effets individuels sont corrélés ou non corrélés ? Pour cela, nous supposons que H0 est "les effets individuels ci ne sont pas corrélés aux variables explicatives" et que H1 est "les effets individuels ci sont corrélés aux variables explicatives". Dans notre cas, l'hypothèse nulle a été rejetée, donc le modèle à effets fixes est préféré car les effets individuels sont corrélés aux variables explicatives.

Ensuite, le test de Fisher pour déterminer si le modèle à effets fixes est meilleur que le modèle MCO poolés, ou inversement.Pour cela, nous supposons que H0 est " tous les ci sont égaux à zéro, donc pas d'effets de panels" et que H1 est " au moins un ci est différent de zéro, donc effet de panel à prendre en compte". Dans notre cas, le test nous a inqué que le modèle à effets fixes est préféré car les effets individuels sont significatifs et corrélés aux variables explicatives, donc H0 a été rejetée.

Enfin, le test de Breusch-Pagan pour déterminer si le modèle à effets aléatoires est meilleur que le modèle MCO poolés, ou inversement. Pour cela, nous supposons que H0 est " variance de ci est nulle, tous les individus t le m^me ci pas d'effets de panels" et que H1 est "variance de ci différent de zéro, effets de panels à prendre en compte". Ce test a montré une préférence pour le modèle EA par rapport au modèle poolé.

En termes de performance, le modèle EF a également démontré une supériorité notable avec un R2 de 0,4426, surpassant celui des modèles poolé (24,6 %) et EA (35,19 %). Ce modèle se distingue par sa capacité à prendre en compte les hétérogénéités spécifiques aux régions et aux années, tout en éliminant les biais liés aux effets inobservés constants dans le temps.

Contrairement au modèle EA, il ne suppose pas d'indépendance entre les effets spécifiques et les variables explicatives, ce qui le rend plus robuste dans des contextes où des facteurs non observés influencent les résultats.

De plus, toutes les variables explicatives du modèle EF se sont révélées significatives, renforçant encore sa pertinence pour cette étude.

Ainsi, le modèle à effets fixes est le meilleur modèle dans notre étude