



ML SUR DONNEES DE PANEL

Prédiction des retards de vols aux USA

Master 2 : Ingénierie et évaluation économique | Université d'Angers | 2020-2021

**Projet final JEDHA**

**Réalisé par :** BARRY Mamadou Yaya

FOTSO Galus

SOMMAIRE

[Introduction 1](#_Toc61977914)

[I. Étude Économique 2](#_Toc61977915)

[II. Étude Économétrique 4](#_Toc61977916)

[II.1. Présentation des données 4](#_Toc61977917)

[II.2. Présentation des variables 5](#_Toc61977918)

[II.3. Spécification du modèle 6](#_Toc61977919)

[II.4. Statistiques descriptives 6](#_Toc61977920)

[II.5. Modèle à conserver (sous TSP) 8](#_Toc61977921)

[II.6. Estimation par la méthode *within* (sous STATA) 12](#_Toc61977922)

[II.6.1. Modèle à effet fixe individuel (One-way FE) 12](#_Toc61977923)

[II.6.2. Modèle within avec effet temporel 13](#_Toc61977924)

[II.6.3. Autocorrélation intra-individus (Test de Wald modifié) 14](#_Toc61977925)

[II.6.4. Test d’hétéroscédasticité inter-individus (Test de Wald modifié) 15](#_Toc61977926)

[II.7. Interprétations économiques du modèle final 17](#_Toc61977927)

[Conclusion 19](#_Toc61977928)

[Bibliographie 20](#_Toc61977929)

# Introduction

Le transport aérien est devenu l'un des moyens de transport les plus populaires au monde, offrant rapidité, commodité et confort pour les voyageurs. Cependant, malgré les progrès technologiques et l'amélioration des infrastructures aéroportuaires, les retards de vols demeurent un problème majeur pour l'industrie.

Selon les données de la Federal Aviation Administration (FAA), environ un quart des vols commerciaux aux États-Unis ont été retardés en 2019, avec un temps moyen de retard de 54 minutes. Ces retards peuvent avoir un impact important sur les voyageurs, qui peuvent manquer des correspondances, perdre des réservations d'hôtel ou subir des coûts supplémentaires.

De plus, les retards peuvent coûter très cher aux compagnies aériennes, avec des coûts opérationnels supplémentaires tels que la rémunération des membres d'équipage, le carburant, le temps d'occupation de l'aéronef et la perte de revenus due à la réduction de la capacité de vol. Selon une étude réalisée par le Bureau of Transportation Statistics (BTS), les retards de vols ont coûté environ 1,5 milliard de dollars aux compagnies aériennes américaines en 2019.

Historiquement, les retards de vols ont été un problème récurrent dans l'industrie du transport aérien. En 2007, les compagnies aériennes américaines ont connu l'une des pires années en termes de retards, avec un taux de ponctualité moyen de 73,4%. Les causes des retards sont diverses, allant de la congestion aérienne aux intempéries, en passant par les problèmes techniques de l'avion ou les retards dus à l'embarquement et au débarquement des passagers.

C'est pourquoi, dans le cadre de ce projet, nous allons nous concentrer sur la prédiction des retards de vols aux États-Unis. Nous allons développer un modèle de Machine Learning simple à déployer, une API et une application pour aider les compagnies aériennes à mieux anticiper les retards de vols et prendre les mesures nécessaires pour minimiser leur impact. Nous allons également explorer différentes techniques de nettoyage et de prétraitement des données, ainsi que différents modèles d'apprentissage automatique pour obtenir la meilleure précision possible.

En somme, ce projet ambitieux a pour objectif d'améliorer l'expérience de voyage des passagers, d'optimiser les opérations des compagnies aériennes et de faciliter la gestion des retards de vols pour l'ensemble de l'industrie du transport aérien.

Haut du formulaire

## I. Définition du cas d’usage

* L'objectif est de prédire les retards de vols pour aider les compagnies aériennes à mieux planifier les vols et à réduire les désagréments pour les passagers.
* Les données à disposition seront les données historiques de vols, les conditions météorologiques, les données de trafic aérien et les informations sur les avions.

### I.1. Concept fonctionnel

Le concept fonctionnel de ce projet consiste à développer un modèle de prédiction des retards de vols aux États-Unis en utilisant des techniques de Machine Learning. Ce modèle sera utilisé pour aider les compagnies aériennes, les voyageurs et les professionnels du transport aérien à mieux comprendre les causes des retards de vols et à prendre des mesures préventives pour minimiser les impacts négatifs sur leur activité.

### I.2. Définition des utilisateurs cibles

Les utilisateurs cibles de ce projet sont les compagnies aériennes, les voyageurs et les professionnels du transport aérien. Les compagnies aériennes peuvent utiliser le modèle pour planifier les horaires des vols, prévoir les retards et minimiser les perturbations pour les passagers. Les voyageurs peuvent utiliser l'application pour obtenir des informations en temps réel sur les retards de vols et planifier leurs itinéraires en conséquence. Les professionnels du transport aérien, tels que les régulateurs et les opérateurs d'aéroports, peuvent utiliser le modèle pour évaluer les performances des compagnies aériennes et les inciter à améliorer leurs services.

### I.3. Ecriture de 3 users stories

1. En tant que compagnie aérienne, Nous voulons être en mesure de planifier nos horaires de vol en fonction des prévisions de retard de vol, afin de minimiser l'impact sur mes activités.
2. En tant que voyageur : recevoir des notifications en temps réel sur les retards de vol afin de pouvoir planifier mon itinéraire en conséquence.
3. En tant que professionnel du transport aérien : avoir accès à des informations fiables sur les performances des compagnies aériennes afin de pouvoir réguler efficacement le marché.

### I.4. Définition des interfaces du produit (entrée/sortie)

Le produit sera constitué d'une API (Application Programming Interface) qui permettra aux utilisateurs d'interagir avec le modèle de prédiction de retard de vol. L'API aura une interface de sortie qui renverra les prévisions de retard de vol pour une compagnie aérienne donnée. L'interface d'entrée de l'API sera constituée des données de vol historiques et des données météorologiques actuelles. L'utilisateur devra fournir les données d'entrée via une requête HTTP à l'API.

### I.5. Définition du format des interfaces

Les données d'entrée seront fournies au format JSON, qui est un format de données couramment utilisé dans les applications Web. Les données de sortie seront également au format JSON, avec des champs spécifiques pour les prévisions de retard de vol. L'API sera accessible via des requêtes HTTP, ce qui permettra une intégration facile avec d'autres applications et systèmes.

## II. Étude Économétrique

Un modèle économique est une représentation simplifiée de la réalité économique ou d'une partie de celle-ci : par exemple la croissance, le commerce international, la monnaie, une entreprise ou un ménage.

Afin de mener à bien notre étude, nous pouvons établir notre modèle économique (qui nous permettra de décrire les liens divers entre les variables) comme suit :

Y = f (X1, X2, X3, …, Xn)

Notre variable d’intérêt (le taux de pauvreté) sera fonction de plusieurs variables explicatives issues d’une base de données établie à partir des données de la banque mondiale :

POV\_RATE = β1X1, i + β2X2, i + · · · + βP-1XP-1, i + βP + ui, i ∈ {1, 2, . . ., N}

Avec :

* POV\_RATE = Le taux de pauvreté (à 3,20$ PPP) qui est la variable expliquée.
* β1, β2, …, βP-1, βP les paramètres à estimer.
* X1, i ; X2, i ; … ; XP-1 = nombres de variables explicatives.

### II.1. Présentation des données

Les données utilisées dans cette étude proviennent du site de la Banque mondiale (World Bank Data[[1]](#footnote-1)). Données à partir desquelles, nous avons réalisé une base de données de quatre cent quatre-vingt-huit (480) observations après traitement des données brutes.

Elle est constituée de 23 variables toutes quantitatives (24 pays d’Amérique latine sur 20 années de 2000 à 2019) que nous avons réorganisées, regroupées puis transformées en fonction des pays afin d’avoir une base de données complète sur le taux de pauvreté.

Cependant, suite à la matrice V de cramer pour observer les relations de corrélations entre les variables, nous avons retenu onze (11) variables dont une expliquée et dix (10) autres qui sont les explicatives.

### II.2. Présentation des variables

Dans le cadre de notre étude, le choix des variables repose sur les différentes revues empiriques portant sur les déterminants de la pauvreté, surtout dans les pays d’Amériques latine, que nous avons pu consulter durant la réalisation de ce travail.

Ces déterminants sont divers et variés. Entre autres, nous pouvons citer les variables que nous avons retenues pour cette étude dans le tableau suivant :

|  |  |  |
| --- | --- | --- |
| Noms | Description | Sigles |
| Taux de pauvreté | La proportion des populations du seuil de 3,20$ par jour. | *Pov\_rate* |
| Taux d’emploi | Taux d’insertion dans l’activité économique. | *emp\_rate* |
| [Indice GINI](https://donnees.banquemondiale.org/indicateur/SI.POV.GINI) | Une mesure statistique permettant de mesurer le niveau d'inégalité de la répartition d'une variable dans la population. | *gini* |
| Dépenses courantes de santé | Niveau des dépenses de santé courantes exprimé en pourcentage du PIB. | *curr\_health\_exp* |
| Crédit intérieur fourni au secteur privé | Désigne les ressources financières fournies au secteur privé par les sociétés financières. | *crdpriv\_gdp* |
| Industrie, valeur ajoutée | La valeur ajoutée des activités industrielles. | *indusva\_gdp* |
| [Agriculture, valeur ajoutée](https://donnees.banquemondiale.org/indicateur/NV.AGR.TOTL.ZS) | La valeur ajoutée des activités agricoles. | *agrva\_gdp* |
| [Prévalence de la sous-alimentation](https://donnees.banquemondiale.org/indicateur/SN.ITK.DEFC.ZS) | Indique le pourcentage de la population dont l'apport alimentaire est insuffisant pour satisfaire les besoins en énergie alimentaire de façon continue. | *undernour\_prev* |
| [Services, valeur ajoutée](https://donnees.banquemondiale.org/indicateur/NV.SRV.TOTL.KN) | La valeur ajoutée de secteur de services | *servva\_gdp* |
| Chômage | La population active sans emploi | *unemprate\_tot* |
| [Taux de participation à la population active](https://donnees.banquemondiale.org/indicateur/SL.TLF.CACT.ZS) | Toutes les personnes qui fournissent du travail pour la production de biens et de services au cours d'une période donnée. | *Labf\_tot3* |

### II.3. Spécification du modèle

Après établissement du modèle économique, il conviendra de le transformer en modèle économétrique. Un modèle économétrique est un ensemble d’équations permettant de déterminer les valeurs prises par un groupe de variables (variables-résultats ou endogènes) en fonction des valeurs attribuées à un groupe de variables (variables-causes ou exogènes).

Cette forme nous permettra non seulement d’étudier les variables qui nous ont été imposées par la théorie pour cette étude, mais également d’observer les variables qui n’ont pas été prises en compte ; c’est-à-dire qu’il y a de nombreuses autres variables qui peuvent avoir un effet sur le taux de pauvreté et que l’on ne peut pas toutes les lister ni observer.

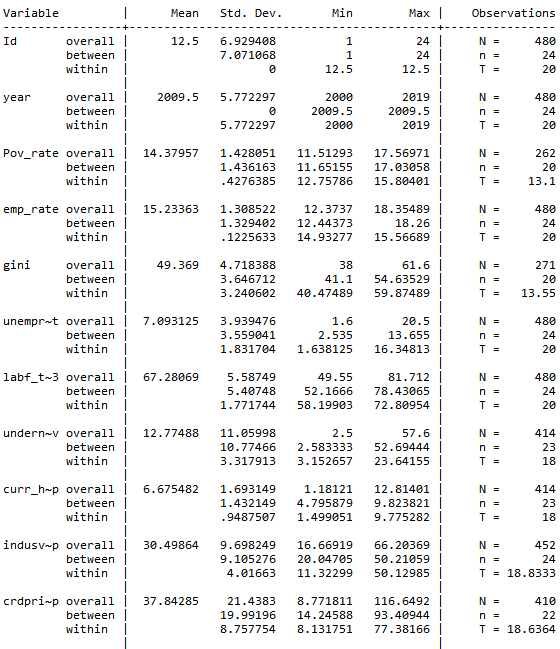
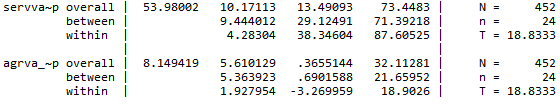
Notre modèle économétrique se présente comme suit :

POV\_RATE = Cste + β1EMP\_RATE + β2GINI + … + β10LABF\_TOT3 + Ɛt

Le but ici, est de faire l’analyse statistique tout en décrivant la structure des données, c'est-à-dire définir la composition des variables, la significativité des variables, les nuages de points et les corrélations. Dans un premier temps nous ferons une analyse des statistiques descriptives de l’ensemble des variables (moyenne, médiane, maximum, minimum, nombre d’observations), ensuite nous ferons les tests nécessaires pour notre étude afin de déterminer le modèle correspondant.

### II.4. Statistiques descriptives

Les statistiques descriptives des données obtenues avec les logiciels SAS, R et celles obtenues avec les logiciels TSP, STATA et Gretl sont identiques. Ce pendant avec STATA, on a plus de modèles par rapport aux autres.

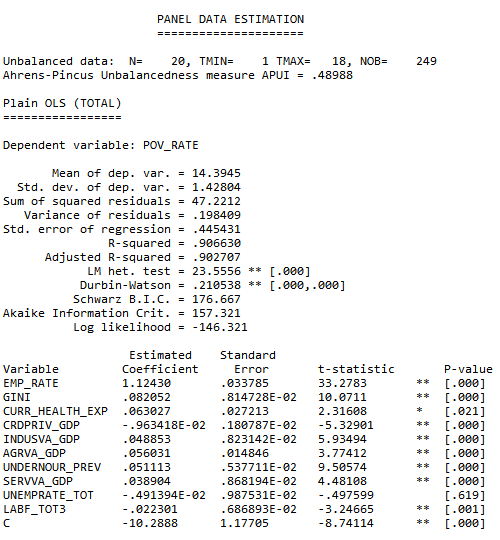


En réalisant une statistique descriptive sur les variables, nous pouvons observer que nous avons effectivement 480 observations avec 24 pays sur une période de 20 ans (de 2000 à 2019). Cependant, suite à la présence des données manquantes (d’autres variables sont mieux renseignées que d’autres) dans notre base, *le panel* est *dit non cylindré* (*unbalanced*) : donc présence de bruit blanc. Malgré cela, le nombre total de nos données reste supérieur à 3000 (480\*11 = 5280 -n\**NA* >3000).

### II.5. Modèle à conserver (sous TSP)

A l’issue de cette procédure (utilisant la commande *panel* de TSP), on dispose des réalisations des estimateurs *Pooled*, *Between*, des estimateurs du modèle à effets individuels fixes (*Within*), du modèle à effets individuels aléatoires (*Error Component Model*), des résultats de trois tests de Fischer, d’un estimateur de la variance des effets individuels, d’un estimateur de la variance totale, de l’estimateur d’un paramètre de pondération et de la statistique du test d’Hausman. Voilà ainsi résumés tous les éléments que nous nous proposons d’étudier tout au long de notre étude.

Méthode Pooled ou des Moindres Carrés Ordinaires (MCO)

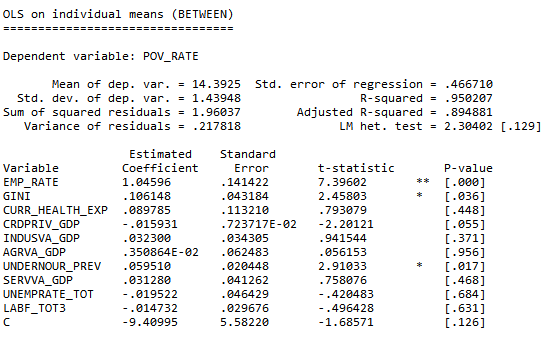


Le OLS nous indique que toutes les variables à l’exception de *unemprate\_tot* sont significatives à 5%.

Toutes les variables significatives, sauf ces deux (*labf\_tot3* et *crdpriv\_gdp*) présentent *ceteris paribus* (toutes choses étant égales par ailleurs) des effets marginaux positifs sur le taux de pauvreté.

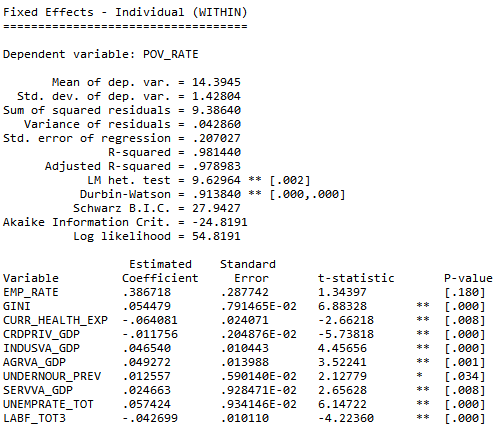
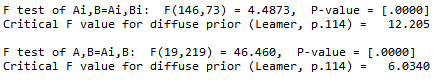
Le R² et le R²-ajusté sont à plus de 90% : ce qui signifie que plus de 90% de la distribution (variation du taux de pauvreté) est pris en charge par le modèle. Cependant le modèle rencontre un problème d’hétéroscédasticité des coefficients.

La *p-value* associée à ce test est très largement inférieure au seuil de 5%. Donc pour ce test, on rejette l’hypothèse nulle *Ho1* d’homogénéité globale : il convient alors de tester l’hypothèse *Ho2* des coefficients (associés aux variables explicatives) entre les pays.

**Estimation par le modèle Between

Le modèle *between*, malgré une bonne variabilité du modèle, reste un mauvais modèle car la plupart des variables ne sont pas représentatives. Il s’emble s’éloigné de l’objectif souligné par le OLS.

Estimation par le modèle à effets fixes - Within

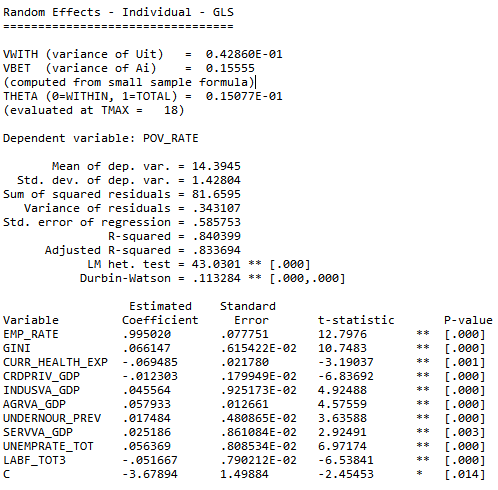
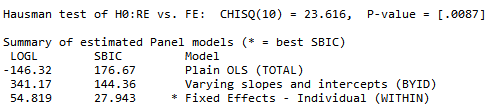


Les *p-value* du *Ho2* et *Ho3* (F test of *Ai, B =Ai, Bi* etF test of *A, B =Ai, B*) d’égalité des coefficients et de constantes individuelles sont très largement inférieurs à 5%. Les hypothèses sont rejetées. Il est donc nécessaire d’introduire des effets individuels.

Nos variables sont toutes significatives à 5%. Seule exception pour le taux d’emploi (*emp\_rate*).

NB : le modèle *within* ne comporte pas de constante (à supprimer sur les autres logiciels tels que STATA).

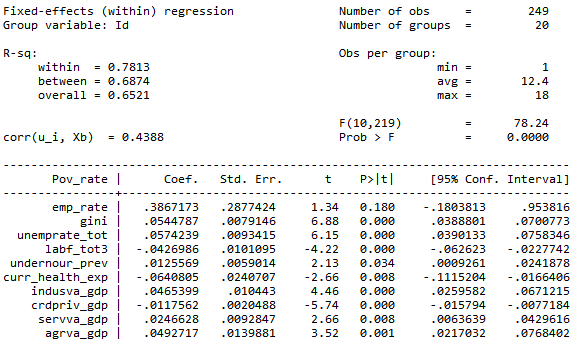
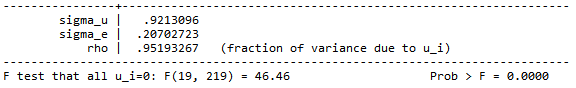
Estimation par le modèle à effets aléatoires – MCG



Ce modèle à effet aléatoire présente des variables toutes significatives à 5%, avec une très bonne variabilité expliquée du modèle. Cependant, un test de spécification d’Hausman permettant de déterminer si les coefficients des deux estimations (fixe et aléatoire) sont statistiquement différents, est nécessaire. L’idée de ce test est que, sous l’hypothèse nulle d’indépendance entre les erreurs et les variables explicatives, les deux estimateurs sont non biaisés, donc les coefficients estimés devraient peu différer. Il résulte donc de nos résultats TSP que le modèle à estimer sera un modèle à effets fixes individuels avec présence d’hétéroscédasticité.

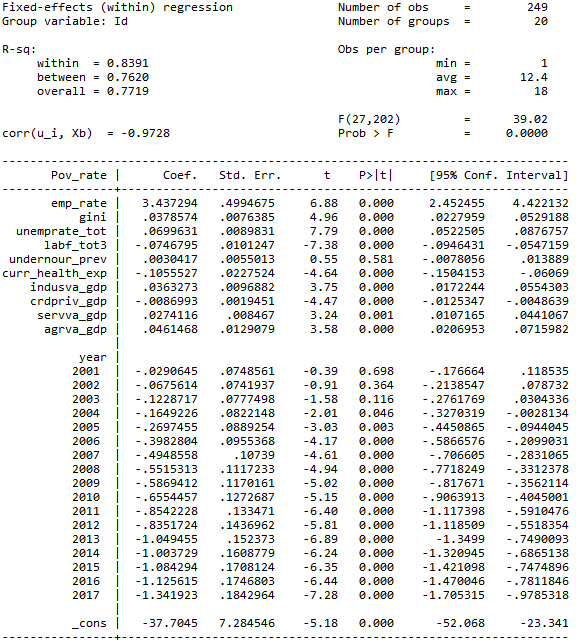
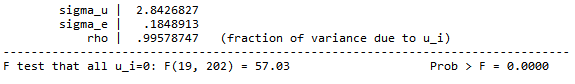
### II.6. Estimation par la méthode *within* (sous STATA)

#### II.6.1. Modèle à effet fixe individuel (One-way FE)



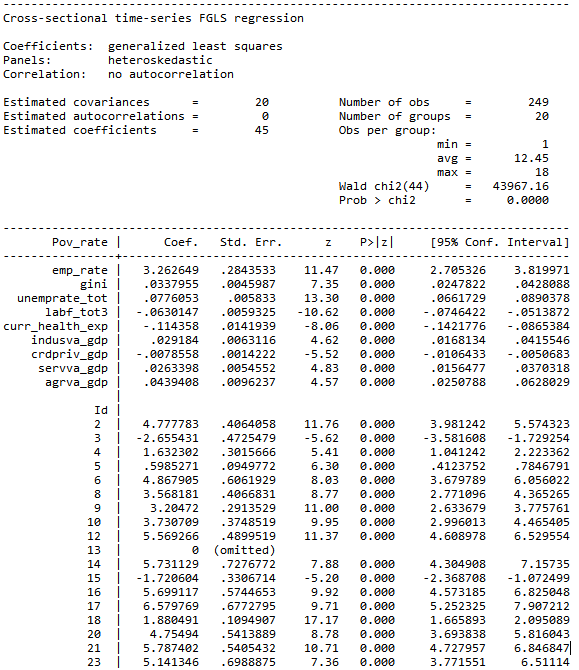
Le test reste le même que sur le logiciel TSP. Cependant, comme le soulignait le test d’Hausman, il y a bel et bien une forte corrélation positive entre les effets individuels et les variables explicatives de notre modèle.

#### II.6.2. Modèle within avec effet temporel

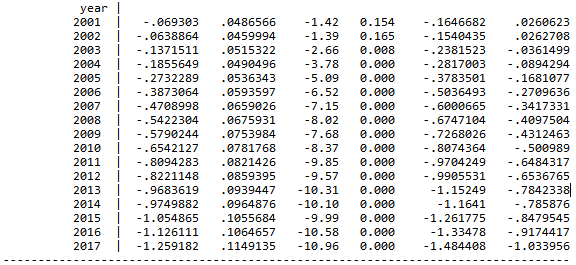


Notre test présente un effet temporel (la p-value de Fisher reste très largement inférieure à 5%. Donc le model est globalement significatif). Dans ce cas, il y a une variabilité dans le temps pour chaque individu (pays) et le problème de colinéarité reste persistant.

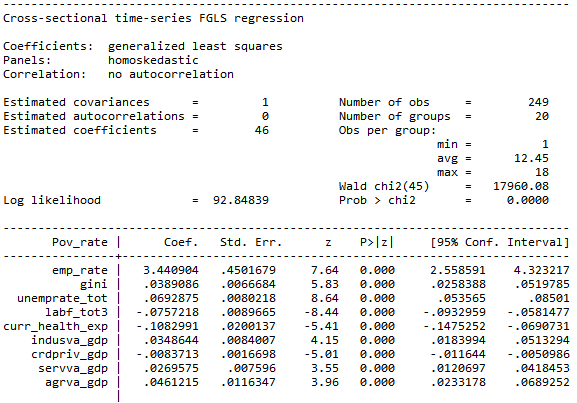
#### II.6.3. Autocorrélation intra-individus (Test de Wald modifié)



Dans ce test de Wald modifié (sous STATA), on ne peut rejeter l’hypothèse nulle d’absence d’autocorrélation des erreurs. Le modèle est globalement significatif, c’est donc un modèle à effets fixes (individuel et temporel) hétéroscédastique.

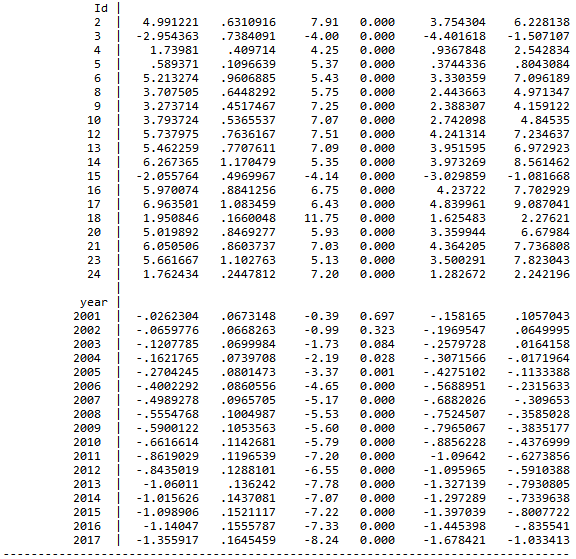


#### II.6.4. Test d’hétéroscédasticité inter-individus (Test de Wald modifié)



Après avoir corrigé le problème d’autocrrrélation, c’est autour de l’hétéroscédasticité.

Le modèle reste très largemment significatif au seuil de 5%. On ne peut rejeter l’hypothèse nulle qui stipule que la variance des erreurs est la même pour tous les individus. Le modèle demeure un *within* à effets fixes individuel et temporel.

z

Pour la composante temporelle : tous les individus sont significatifs à l’execption des années 2001-2003. Sans doute dû à des des périodes de crises.

### II.7. Interprétations économiques du modèle final

Dans cette partie, nous allons exposer ce qui ressort de notre étude (cf. II.6.4) et expliquer les différentes relations de causes à effets vis-à-vis des différentes littératures sur le sujet.

Les effets marginaux négatifs :

* « labf\_tot3 », la proportion de la population active ave une élasticité d’environ 0,08% ;
* « curr\_health\_exp », avec une élasticité d’environ 0,11% pour les dépenses de santé publique (en % PIB) ;
* « crdpriv\_gdp », et enfin une élasticité de 0,008% pour le crédit intérieur contracté par le secteur privé.

Une augmentation de 1% des ces différentes variables, toutes choses étant égales par ailleurs (ceteris paribus) entraine un effet marginal négatif sur le taux de pauvreté. C’est-à-dire que toutes les politiques publiques visant à améliorer le niveau de santé et de relancer l’activité économique, baissent le niveau de pauvreté des populations latino-américaines ayant un PPP (*Purchasing Power Pariy : Parité de Pouvoir d’Achat* ) de 3,20$ par jour.

Les variables telles que l’indice de gini et le taux de chômage, présentent ceteris paribus, des effets marginaux positifs sur le sur taux de pauvreté. Leur augmentation de 1%, entraine respectivement une accentuation des inégalités entre pauvres et riches. L’écart se creuse d’environ 0,039% entre les différentes classes sociales et de 0,069% entre individus de la population active.

Par contre, le reste des variables (le taux d’emploi et les différentes valeurs ajoutées dans le secteur primaire et tertiaire) présentent également des effets marginaux positifs. Cela ne semble-t-il pas contradictoire ? La logique nous aurrait laissé penser ainsi. Mais non, car ces élastiscités renferment une subtilité qui s’apparente à deux cas de phénomènes majeurs que rencontrent ces pays :

* Premièrement les pays d’Amérique latine tels que le Mexique, le Brésil et l’Argentine, ont libéralisé leurs économies : c’est une nouvelle Amérique latine qui émerge dans les années ‘‘2000’’.
* Deuxièmement, les économies latino-américaines sont beaucoup influencées par les différents courants politiques notamment avec l’avènement des régimes populistes d’un côté ; et de l’autre, une faible croissance de l’emploi et des sous-emplois malgré l’essor des secteurs miniers et agro-industriels qui favorisent la force de travail rurale.

C’est suite à cela, que malgré leur faible taux d’augmentation, le taux de pauvreté s’accentue. Il y a une mauvaisae redistribution des richesses. Seules les grandes firmes trans et multinationales en tirent profit. A cela, s’ajoute les problèmes de corruption et d’insécurité.

Quant aux composantes temporelles (période avant 2004) : la non significativité de ces 3 variables se traduit par le faite qu’en 2004 (notamment la période de 2004-2008 où les économies latino-américaines connaissent un essor) on sortait d’une période de récession mondiale.

En palliant, résultats économétriques et littératures économiques, nous avons réussi à s’assurer de la véracité du modèle : but ultime de l’économètre.

# Conclusion

En somme, nous retenons que l’aboutissement à un bon modèle se trouve au préalable dans la construction et la gestion des données. Cette étape nous permet d’avoir, notamment avec un nuage de point ou une matrice V de cramer, une idée sur le choix des variables visant à expliquer le phénomène d’étude.

En effet, la méthodologie de travail utilisée dans ce projet, visait au-delà d’une simple analyse économétrique. L’objectif était de réaliser des relations de causes à effets afin de confirmer ou infirmer certaines littératures qui le plus souvent des cas, sont basées sur des cas factuels et non sur la recherche de causalités via des méthodes qui pourraient exposer des phénomènes non apparents.

Les niveaux de pauvreté et de chômage galopants sont, de nos jours, des problèmes de sociétés majeurs devant impliquer toutes les différentes couches sociales.

Des politiques de relance économiques et budgétaires sont vivement recommandées, notamment en cette période de crise sanitaire afin d’éviter une implosion des classes sociales déficitaires qui se soldera par des troubles d’ordre sociaux où personne ne sera épargné.

# Bibliographie

BOSSIO ROTONDO, J. C. « Amérique latine: pauvreté, inégalités, chômage et orientations politiques. », *Vie économique* 1.3 (2010).

ÉPAULARD, Anne. « Croissance et réduction de la pauvreté dans les pays en développement et les pays en transition », *Reflets et perspectives de la vie économique*, vol. tome xlii, no. 2, 2003, pp. 9-20.

VALDES, A., and MISTIAEN J. A. « Pauvreté rurale en Amérique Latine: tendances récentes et nouveaux enjeux. », (2001).

Webographie

<https://www.banquemondiale.org/fr/region/lac/overview>, consulté le 12/01/2021

https://www.lepoint.fr/monde/l-extreme-pauvrete-touche-plus-10-de-la-population-en-amerique-latine-15-01-2019-2286104\_24.php, consulté le 12/01/2021

1. https://donnees.banquemondiale.org/ [↑](#footnote-ref-1)