



FACTEURS AFFECTANT LA DEMANDE DE L'ASSURANCE-VIE DANS LE MONDE

SEMESTRE 2 – DATA MINING 2

MASTER 1 – DATA SCIENCE

- ASHUZA CIRUMANGA Destin
- GOVENDASAMY Elvina
- GHERNAOUT Nadia

RESUME

Le marché de l'assurance-vie a connu une évolution fluctuante caractérisée par des hausses et des baisses de demandes de produits d'investissements durant les trois dernières décennies.

L'objectif de cette étude est d'analyser les effets des facteurs socio-économiques et démographiques sur la demande en assurance-vie dans le monde. Nous utilisons la densité de l'assurance-vie comme variable réponse pour mesurer le développement de ce type de marché dans divers pays.

Pour répondre à la problématique, nous commençons par faire une analyse économique qui sert à présenter les variables choisies dans l'étude. Ensuite une analyse exploratoire est réalisée à l'aide d'une analyse en composantes principales et d'une classification ascendante hiérarchique. Ces deux analyses nous permettent de mieux comprendre le regroupement des variables et des individus, dans notre cas, les pays. Tandis que l'analyse en composantes principales nous aide à voir clairement que les variables sont classées en termes de niveau de développement du pays ou de villes, la classification hiérarchique regroupe les pays par rapport au niveau du taux d'inflation, le développement des villes, le taux de chômage et finalement, le niveau de développement des pays.

À la suite de l'analyse exploratoire, nous faisons une analyse économétrique où, une régression multiple, une régression par composantes principales et finalement une méthode de PLS (Partial Least Squares) sont implémentées, testées puis comparées. Finalement, c'est ce dernier modèle que nous considérons le meilleur car il explique une plus grande variabilité de la variable réponse, la densité de l'assurance-vie.

Toutes ces analyses nous permettent d'avoir des informations pertinentes : savoir notamment que plus un pays est riche, plus la population est instruite et en bonne santé, plus la demande de produits de l'assurance-vie est importante. Le développement de ce marché sera ainsi plus propice dans les pays développés.

ABSTRACT

The evolution of life insurance market has been volatile, a fluctuation that has been characterized by ups and downs in demand of investment products during the last three decades.

The aim of this study is to analyze the effects of socio-economic et demographic factors impacting global demand in life insurance products. Life insurance density, used as the response variable, serves to properly measure the development of this type of market across the world.

To solve this issue, an economic analysis is first made to present the variables selected for the study. It is followed by an exploratory analysis made by Principle Component Analysis and Hierarchical Ascending Classification. These two methods allow us to properly regroup variables and observations (here, in the form of countries). While the Principal Component Analysis help is to classify variables in terms of the level of development of countries or by the level of development of cities, the Hierarchical Classification regroups countries in terms of inflation rate, development of cities and countries, as well as the unemployment rate.

Following the exploratory analysis, an econometric analysis is performed, where multiple regressions, Principal Composant Regression and finally Partial Least Squares method are implemented, tested and then compared. Finally, the last model is selected as being the best one since it explains the greatest variation of the response variable, life insurance density.

All these analyses allow us to obtain valuable information, namely, that the richer a country is, the more its inhabitants are educated and in good health, the greater the demand in life insurance products. Therefore, the development of life insurance market fairs better in developed countries.

INDEX

TABLES DE MATIÈRES

- I. INTRODUCTION
- II. REVUE BIBLIOGRAPHIQUE
- III. ANALYSE ÉCONOMIQUE
- IV. ANALYSE EXPLORATOIRE
- V. ANALYSE ÉCONOMÉTRIQUE
- VI. CONCLUSION
- VII. BIBLIOGRAPHY
- VIII. ANNEXE

TABLE DES MATIERES

I- INTRODUCTION	7
II- REVUE BIBLIOGRAPHIQUE	9
III- ANALYSE ÉCONOMIQUE	13
PRESENTATION DES DONNÉES	13
PRESENTATION DES VARIABLES	13
La variable à expliquer	14
Les variables explicatives	14
IV- ANALYSE EXPLORATOIRE	21
Analyse univariée	21
Analyse bivarié	22
Observations atypiques et abérantes	24
Réduction de dimension	25
HIERARCHICAL AGGLOMERATIVE CLUSTERING	27
V- ANALYSE ECONOMÉTRIQUE	30
Estimation du modèle	30
Régression linéaire – aic	30
Régression sur composantes principales	37
Méthode partial least squares (pls)	41
VI- CONCLUSION	49
<i>Bibliographie</i>	51
<i>ANNEXE</i>	52

I- INTRODUCTION

Une police d'assurance-vie est un contrat établi entre un individu et une compagnie d'assurance. Ce premier réalise des paiements sous forme de primes et en retour la compagnie verse un montant forfaitaire ou rentes à une date ultérieure. L'objectif de cette assurance est de couvrir les risques associés au décès ou la survie d'un individu et correspond donc à un produit d'épargne qui s'étend en général sur le moyen ou long terme.

La demande en assurance-vie a beaucoup évolué depuis son émergence. Il a fallu, toutefois, attendre de nouvelles régulations, la mutualisation des firmes, des crises financières et la deuxième guerre mondiale pour que celle-ci augmente. Nous notons également que les États-Unis et le Japon furent les premiers à avoir la plus grosse part du marché mondiale en 1986, suivis de quelques pays Européens, tels que la Grande-Bretagne, la France et l'Allemagne de l'Ouest.¹ D'ailleurs, à ce jour, l'appétence et la sensibilisation de la population Nord-Américaine dépassent encore celles de l'Europe continentale. Cette disparité observée entre les différents pays nous pousse à émettre notre hypothèse, soit que la demande en assurance-vie serait impactée par plusieurs facteurs socio-économiques.

Devant l'instabilité économique qu'a connue le monde durant cette dernière décennie, notamment avec la crise des *subprimes* et la crise de la dette souveraine des pays européens de nouveaux questionnements quant à l'impact des facteurs économiques sur la demande en assurance-vie se soulèvent. Alors que la demande en assurance-vie se voyait grandir, ces crises économiques ont eu un effet dévastateur pour beaucoup de firmes. Nous notons d'ailleurs que lors de la crise des *subprimes* en 2008-2009, certaines compagnies d'assurance-vie aux États-Unis, tel que AIG ont perdu presque des centaines de milliards de dollars en 2008.² Ces pertes initialement liées aux marchés financiers ont eu un impact majeur sur les demandeurs. L'assurance-vie, souvent connue comme étant un outil d'épargne longue vie ne l'était plus.

¹ Collett, Robert L., Abkemeier, Noel J., Bonach, Edwar J., and Papasavvas, Demos K. Evolution of Life Insurance Industry Throughout the World. 1990. *Record of Society of Actuaries*.

² Jérôme Bonnard. Les conséquences des crises financières de 2008/2009 et 2011/2012 sur l'assurance. 2012. <https://halshs.archives-ouvertes.fr/halshs-00655657/document>

De plus, les données de l'OCDE obtenues en 2004³ nous montrent l'urgence d'investir davantage sur les fonds de pension en France qui comptent 0.5% du PIB contrairement à 159% aux Pays-Bas ou encore 83% aux États-Unis.

Ainsi, avec les changements drastiques, à la hausse tant qu'à la baisse, observés durant ces trente dernières années, et avec le vieillissement de la population contribuant à l'épuisement des fonds de régimes, nous cherchons à analyser comment des facteurs socio-économiques auraient un impact sur la demande d'assurance vie.

Lors de notre étude nous nous intéresserons à mesurer les effets de ces facteurs sur la demande en assurance vie dans les différents pays du monde. Nous analyserons ainsi plusieurs variables qui nous semblent pertinentes, basées sur des articles ou recherches antérieures, afin de voir s'il existe une corrélation entre la demande en assurance vie et ces facteurs. Nous démontrerons également que certains auteurs d'articles émettent des hypothèses différentes, nous poussant ainsi à faire notre propre analyse en nous basant sur nos propres données. Lors de l'étude nous verrons si oui ou non les variables choisies ont vraiment un lien avec la demande en assurance-vie. Nous chercherons, par la suite, à effectuer différents tests afin de voir la validité, par la significativité de nos variables. Nous tenterons de chercher les sources d'erreurs et différences par rapport à nos attentes, et finalement de proposer, s'il y a lieu, le modèle optimum pouvant expliquer le mieux les tendances.

³ Secrétariat général du Conseil d'orientation des retraites. Le point sur les fonds de pensions – synthèse des travaux de L'OCDE. 2016. *Vieillissement, emploi et retraite : panorama international*. Téléchargement : <http://www.cor-retraites.fr/IMG/pdf/doc-3178.pdf>

II- REVUE BIBLIOGRAPHIQUE

Avant de débuter notre analyse nous voulons porter une attention particulière sur des recherches et articles faits ultérieurement sur ce sujet.

Nous classerons les articles en deux catégories en prenant en compte la période à laquelle ils ont été présentés.

Nous avons ainsi l'article le plus ancien, datant de 1990, où l'approche économétrique était différente. Aucune preuve ou test n'a été fait, les facteurs qui pouvaient influencer la demande dans le secteur de l'assurance vie étaient présentés d'une telle façon à ce que nous notions les différences entre la demande entre différents pays. Nous nous sommes basés sur cette revue pour avoir notre première idée des facteurs que nous voulions chercher.

Les revues plus récentes démontraient des similitudes en termes de facteurs étudiés. Ils utilisaient également des méthodes d'analyses économétriques plus adaptées à notre étude. Nous nous sommes basés sur ces articles pour faire nos hypothèses sur ces variables. Toutefois, nous noterons des différences entre les résultats observés entre différentes études, nous encourageant ainsi, à faire nos propres recherches en utilisant nos propres données.

Revue plus ancienne :

En 1990, un article intitulé : '*Evolution of Life Insurance Industry*'⁴ évoque l'évolution de l'assurance vie et les facteurs qui influençaient sa demande durant les années qui précédaient sa publication. Le but de l'article était de montrer les pays qui présentaient les meilleurs 'pronostics' par rapport au développement de ce secteur pour finalement voir si une globalisation de cette industrie était possible. Ainsi des comparaisons par rapport aux pays, et non entre les facteurs. Toutefois, les facteurs étaient intéressants à analyser : nous comptons

⁴ Collett, Robert L., Abkemeier, Noel J., Bonach, Edward J., and Papasavvas, Demos K. *Evolution of Life Insurance Industry Throughout the World*. 1990. *Record of Society of Actuaries* <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjC3ufEiqHeAhUBQBoKHZY-DDQQFjAAeqQIBBAC&url=https%3A%2F%2Fwww.soa.org%2Flibrary%2Fproceedings%2Frecord-of-the-society-of-actuaries%2F1990-99%2F1990%2Fjanuary%2FRSA90V16N4A18.PDF&usg=AOvVaw3OKktOsijOzIYRf4S3NZcs>

la réputation et crédibilité des compagnies d'assurance-vie, l'inflation, la progression de la profession des actuaires, la superstition, notamment dans certains pays d'Europe du Sud, les revenus des habitants, le type de ménage, la mentalité des gens quant à la notion d'économiser ou emprunter de l'argent, l'économie du pays, l'établissement des fonds de retraite, les législations et taxation dans les pays, comme une taxation sur la prime en France qui décourageait les individus à investir. Nous notons également des facteurs comme les différences de cultures entre différents pays qui impactaient l'aversion aux risques des individus. Bien que beaucoup de ces variables soient intéressantes, nous avons utilisé uniquement l'inflation et le revenu des habitants (sous la forme de PIB). Faute de données suffisantes nous n'avons malheureusement pas pu étudier certaines variables présentées dans cette revue comme les facteurs suivants: l'établissement des fonds de retraite, les législation et taxation. Nous n'avons également pas considéré les variables : mentalité quant à la notion d'argent, superstition, ou cultures car il s'agissait de variables qualitatives, difficiles à mesurer.

Articles plus récents :

Nous nous basons sur trois articles plus récents pour la recherche de nouvelles variables. Les articles sont '*The determinants of the demand of life insurance in an emerging economy – the case of China*'⁵, '*Les déterminants de la consommation d'assurance-vie: le cas de L'UEMOA*'⁶, et finalement '*Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries*'⁷.

⁵ Tienyu, Hwang. The determinants of the demand for life insurance in an emerging economy – the case of China. 2003. Managerial Finance. Vol 29. p82-96

⁶ Dieng, Momar S., et Fall, Mouhamadou. Les déterminants de la consommation d'assurance-vie : le cas de l'UEMOA. 2015. Revue d'Économie Théorique et appliquée. (juin). p 15-36

⁷ Thorsten, Beck and Webb, Ian. Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. 2002. World Bank and International Insurance Foundation.

Nous notons que les deux premières revues sont basées sur des marchés plus ‘micros’ et regardent de façon transversale pour l’un et temporelle pour l’autre comment certains facteurs affectent la demande à l’intérieur d’un pays ou zone géographique du monde en voie de développement. Le troisième article fait une étude temporelle sur plusieurs pays du monde, soit 68 pays sur une période de 1961 à 2000.

Nous verrons plus tard lors de la présentation de notre sujet que nous avons choisi d’évaluer la demande en assurance vie avec comme outil de mesure : la densité de l’assurance, soit le montant de prime en dollars par habitant. Nous garderons en tête que les études mentionnées ci-dessus utilisent souvent le taux de pénétration soit, le ratio de prime par rapport au PIB, ce qui implique que nous pourrons ne pas obtenir les mêmes résultats.

Nous avons notamment l’inflation, qui selon les 3 articles, aurait une corrélation négative avec la demande et la croissance de l’assurance vie. Le développement du secteur bancaire et financier, ainsi que le revenu par habitant auraient quant à eux chacun des liens positifs avec la demande en assurance vie.

Nous notons toutefois de grosses différences au niveau de la significativité entre certains articles. L’éducation, le ratio de jeunes dépendants, l’espérance de vie ne seraient pas significatifs tandis que l’inflation le serait, selon l’article de Thorsten, Beck et Ian, Webb (2003).⁸ Toutefois, celui de Dieng, Momar S., et Fall, Mouhamadou (2015)⁹ révèle un manque de données au niveau des années pour la variable éducation ainsi ne pouvant pas donner des résultats au niveau de sa significativité, et affirme que l’espérance de vie et le ratio de jeunes dépendants, tous les deux négativement corrélés à la croissance de l’assurance vie seraient des facteurs significatifs, tandis que l’inflation ne le serait pas, et cela même selon Tienya,

⁸ Thorsten, Beck and Webb, Ian. Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. 2002. World Bank and International Insurance Foundation. Téléchargement: <https://pdfs.semanticscholar.org/36f8/b2b87f98257f82964cc20b86443df2dfff20.pdf>

⁹ Dieng, Momar S., et Fall, Mouhamadou. Les déterminants de la consommation d’assurance-vie : le cas de l’UEMOA. 2015. Revue d’Économie Théorique et appliquée. (juin). p 15-36

Hwang (2002)¹⁰. Cette dernière sera d'ailleurs la seule auteure à trouver que la variable éducation soit significative par rapport à la demande en assurance vie en Chine.

Pourquoi de telles différences ?

Comme cela a été noté au début de cette section, certains articles basent leurs recherches sur des études temporelles, pour la plupart, tandis que d'autres non. Or même si les mêmes types d'études temporelles ont été réalisées, les années observées sont différentes, pouvant causer des différences entre les résultats. Une autre causalité serait que certaines recherches sont faites à l'intérieur d'un pays ou certaine zone géographique du monde, tandis que, Thorsten, Beck et Ian, Webb (2003)¹¹ basent leur étude sur plusieurs pays du monde. Sur cela s'ajoute le fait que Tienyu, Hwang (2002)¹² et Dieng, Momar S., et Fall, Mouhamadou (2015)¹³ font leurs études sur des pays en voie de développement tandis que Thorsten, Beck et Ian, Webb (2003)¹⁴ utilisent une population plus variée et large, possible source d'hétéroscédasticité. Cela pourrait d'ailleurs expliquer pourquoi Tienyu, Hwang (2002)¹⁵ et Dieng, Momar S., et Fall, Mouhamadou (2015)¹⁶ ont eu les mêmes résultats par rapport à la variable éducation c'est-à-dire, que l'impact de l'éducation pour les pays en voie de développements serait plus significatif sur la croissance de l'assurance vie.

¹⁰ Tienyu, Hwang. The determinants of the demand for life insurance in an emerging economy – the case of China. 2003. Managerial Finance. Vol 29. p82-96

¹¹ Thorsten, Beck and Webb, Ian. Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. 2002. World Bank and International Insurance Foundation. Téléchargement: <https://pdfs.semanticscholar.org/36f8/b2b87f98257f82964cc20b86443df2dfff20.pdf>

¹² Tienyu, Hwang. The determinants of the demand for life insurance in an emerging economy – the case of China. 2003. Managerial Finance. Vol 29. p82-96

¹³ Dieng, Momar S., et Fall, Mouhamadou. Les déterminants de la consommation d'assurance-vie : le cas de l'UEMOA. 2015. Revue d'Économie Théorique et appliquée. (juin). p 15-36

¹⁴ Thorsten, Beck and Webb, Ian. Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. 2002. World Bank and International Insurance Foundation. Téléchargement: <https://pdfs.semanticscholar.org/36f8/b2b87f98257f82964cc20b86443df2dfff20.pdf>

¹⁵ Tienyu, Hwang. The determinants of the demand for life insurance in an emerging economy – the case of China. 2003. Managerial Finance. Vol 29. p82-96

¹⁶ Dieng, Momar S., et Fall, Mouhamadou. Les déterminants de la consommation d'assurance-vie : le cas de l'UEMOA. 2015. Revue d'Économie Théorique et appliquée. (juin). p 15-36

III- ANALYSE ÉCONOMIQUE

PRESENTATION DES DONNÉES

Les données ont été extraites des bases de 2013 de l'OCDE¹⁷ et de WorldBank¹⁸, une année qui est considérée comme suffisamment récente et contenant moins de données manquantes. La base initiale contient 56 pays à travers le monde. Afin de pouvoir utiliser la méthode de moindres carrées, les observations ayant des valeurs manquantes ont été enlevées de la base.

PRESENTATION DES VARIABLES

La base contient 11 variables explicatives, que nous utilisons pour présenter le modèle initial :

$$Y = \alpha + \beta_1(LITERACY) + \beta_2(GOOD_HEALTH) + \beta_3(URBAN_POP) + \beta_4(LIFE_EXP) + \beta_5(UNEMP) + \beta_6(OLD_DEP) + \beta_7(INFLATION) + \beta_8(FINANC_DEV) + \beta_9(GDP) + \beta_{10}(YOUNG_DEP) + \beta_{11}(GNI) + \varepsilon, \text{ où :}$$

- **Y : INS_DEN** : La densité de l'assurance, en \$ US
- **LITERACY** : Taux d'alphabétisation (% de la population totale)
- **GOOD_HEALTH** : Pourcentage de la population en bonne santé
- **URBAN_POP** : Population urbaine (% de la population totale)
- **LIFE_EXP** : Espérance de vie
- **UNEMP** : Taux de chômage
- **OLD_DEP** : Ratio des inactifs par rapport aux actifs
- **INFLATION** : Taux d'inflation
- **FINANC_DEV** : Développement du secteur financier (ratio M2/PIB)
- **PIB/GDP** : PIB par habitant (en US\$). (OCDE Statistics, s.d.)
- **YOUNG_DEP** : Ratio des jeunes inactifs (<18) par rapport à la population active

¹⁷ OCDE Statistics. <https://stats.oecd.org> (2020/02/02)

¹⁸ WorldBank:Data Bank. <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS> (2020/02/02)

➤ **GNI** : Revenu national brut en \$ US

LA VARIABLE A EXPLIQUER

INS_DEN : DENSITE DE L'ASSURANCE-VIE (EN \$ US)

Afin de pouvoir mesurer la demande en assurance vie par pays il a fallu faire un choix afin de trouver l'outil de mesure le plus représentatif. Cette mesure représente notre variable à expliquer, Y, c'est-à-dire la variable pour laquelle nous souhaitons décrire les changements en fonction des variables explicatives.

Il existe deux mesures du développement du marché de l'assurance-vie¹⁹ :

- 1) Le taux de pénétration dans une économie,
- 2) La densité de l'assurance, en dollars américains

Les deux mesures sont proches. La première est équivalente au nombre total de primes d'assurance-vie en pourcentage du PIB (primes totales / PIB) tandis que la deuxième mesure la proportion de ces primes par habitant (primes totales / habitant) (Dieng & Fall, 2012). Plusieurs études, tel que détaillé dans la revue bibliographique, ont utilisées ces deux mesures afin de quantifier le développement, l'expansion du marché de l'assurance, qui reste un marché d'investissement. Afin d'éviter tout risque d'endogénéité avec la variable explicative PIB – l'endogénéité étant le risque que cette variable explicative influence également la variable à expliquer – le meilleur choix semble être donc la densité de l'assurance.

LES VARIABLES EXPLICATIVES

LITERACY

Le taux d'alphabétisation chez les adultes est le « pourcentage de la population âgée de 18 ans et plus qui peuvent, en comprenant, lire et écrire des phrases courtes et simples »

¹⁹ Mesurer le développement de l'assurance : au-delà du taux de pénétration de l'assurance. Compte rendu de la 21e Consultation téléphonique A2ii-AICA. 2017. URL:

https://a2ii.org/sites/default/files/reports/21_consultation_call_fr_web.pdf

(UNESCO, 2019). Depuis les années 90, le taux d'individus pouvant lire et écrire a augmenté considérablement parmi les pays développés (Roser & Ortiz-Ospina, 2013). Cette variable est ainsi considérée comme une bonne mesure de développement d'un pays, nous poussant à chercher s'il existe une éventuelle relation avec le développement du marché d'assurance-vie. Effectivement, des individus pouvant mieux lire et écrire seraient plus disposés à avoir des polices d'assurance vie, car non seulement seraient-ils capables de lire les contrats d'assurance mais ils auraient un meilleur niveau d'éducation et une plus grande sensibilisation à l'assurance vie et ses implications. De ce fait, nous supposons que ces deux variables sont positivement corrélées. De plus, plus un pays serait développé, c'est-à-dire, plus le PBI, GNI ou encore le pourcentage de la population urbaine grandirait, plus le taux d'alphabétisation devrait augmenter.

GOOD HEALTH

Nous distinguons deux types de contrats d'assurance vie, soit les contrats en cas de décès et ceux en cas de vie.

Regardons la police d'assurance vie en cas en de vie plus en détails. Ce type de contrat est équivalent à un placement ou une épargne car elle représente une garantie de versement de rentes ou montant forfaitaire lorsque l'assuré survit au-delà d'une certaine date. Nous aurons ainsi pensé que le niveau de santé d'un individu aurait un lien avec sa demande d'assurance vie. Afin d'établir le niveau de santé de la population d'un pays, nous avons choisi comme variable le pourcentage de la population en bonne santé.

La mesure de l'état de santé des individus reste toutefois sensible car les données restent limitées et peu accessibles au public.²⁰ La disparité entre l'obtention des données dans les pays en développement et les pays développés nous pousse à être prudent avec cette variable choisie. Il est également aussi important de noter que la notion de 'bonne' santé reste un point subjectif et relatif, impliquant qu'un individu dans un pays ne serait pas nécessairement

²⁰ https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthy_life_years_statistics/fr
(Healthy life years statistics, s.d.) 2020/02/10

en ‘bonne’ santé dans un autre. Cela dépend du médecin qui s’occupe du patient et même de la stabilité économique du pays, notamment un pays qui vit une période de guerre ou non.

Nous nous intéressons néanmoins à cette variable, malgré son manque de ‘fiabilité’ au niveau des données, car il serait intéressant de voir s’il existe un lien entre la santé d’un individu et son envie d’investir en supposant qu’il survivra suffisamment longtemps pour récolter les bénéfices de son placement.

LIFE_EXP

Selon l’INSEE, l’espérance de vie à la naissance est la durée de vie moyenne, l’âge moyen jusqu’au décès. Cette espérance est donc le nombre moyen d’années restantes à vivre pour une génération quelconque²¹.

Il est possible que le décès intervienne avant ou après l’âge attendu. Cependant, les progrès de la médecine depuis de nombreuses années vont permettre de guérir une maladie qui autrefois était très difficile à guérir. Ces derniers vont donc augmenter la durée de vie d’un individu, mais non la qualité de vie.²²

L’augmentation de l’espérance de vie a un impact direct sur les risques de longévité²³, qui, à leur tour, ont un impact direct sur la demande d’assurance-vie. Effectivement, nous supposons qu’il existe une relation positive entre ces deux variables car une plus grande anticipation de longévité entraînerait une plus grande consommation de produits d’épargne long-terme, comme l’assurance-vie.

(BROWN & GUY C.)

²¹ <https://www.insee.fr/fr/metadonnees/definition/c1374>. (INSEE, s.d.)2020/02/10.

²² Brown, Guy C. 2015. Living too Long. EMBO Reports. Vol 16. No 2. URL:
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4328740/pdf/embr0016-0137.pdf>

²³ National Association of Insurance Commissioners. Longevity Risk. URL:
https://www.naic.org/cipr_topics/topic_longevity_risk.htm (2020/02/10)

URBAN_POP

La 'population urbaine' est une variable démographique affichant le pourcentage de la population d'un pays vivant dans les zones urbaines.

L'accroissement de cette population affecte différemment les marchés émergents des pays Européens, où il existe un système d'économie de marché, et ceux d'Asie où il existe une économie planifiée²⁴. Ainsi, nous supposons que le développement des zones urbaines aurait un impact sur le développement du marché financier et ainsi sur le marché de l'assurance-vie.

De plus, nous chercherons, lors de cette étude, à établir un lien éventuel entre le niveau de santé global des individus dans un pays, ou encore le niveau d'alphabétisation et la taille de la population urbaine. Ainsi, cela nous permettrait de confirmer s'il existe un lien entre cette dernière et le développement du marché de l'assurance-vie, tel que spécifié par certains auteurs, notamment (Thorsten & Webb, 2002), qui concluaient trouver un lien significatif et négatif et (Dieng & Fall, 2012) au contraire avaient trouvé un lien non significatif mais positif.

UNEMP

Selon OCDE²⁵, le taux de chômage est la proportion d'individus sans emploi sur la population active, soit celle en 'droit' de travailler. Le chômage affecte tous les pays du monde, et le taux est indicateur du pouvoir d'achat d'une population. C'est ainsi que nous supposons qu'il existe une relation significative et négative entre cette variable et le PIB ou encore le Revenu National Brut (variable GNI). Les individus ayant moins de revenus ne peuvent pas se permettre d'épargner surtout à long-terme. Ainsi, nous supposons, dans cette étude, que les pays ayant un haut taux de chômage auraient un marché d'assurance-vie moins développé.

²⁴ Dragos Simona. 2014. Life and non-life insurance demand: The different effects of influence factors in emerging countries from Europe and Asia. *Economic Research* 27(1) (November) :169-180.

²⁵ OCDE Statistics. <https://stats.oecd.org> (2020/02/02)

INFLATION

Selon l'INSEE, on définit l'inflation comme une perte de pouvoir d'achat de la monnaie et qui va s'expliquer par une augmentation des prix qui sera générale et durable. Le taux d'inflation est ainsi un facteur du taux de croissance et il est évalué à partir de l'IPC (Indice des prix à la Consommation).²⁶ Le taux d'inflation est ainsi considéré comme une bonne mesure de la 'santé' d'une économie, et nous supposant qu'il est négativement corrélé au PIB ou le Revenu National brut dans d'un pays. De ce fait, plus il augmenterait, moins les habitants d'un pays aurait la capacité d'acheter des primes d'assurances-vie. Nous estimons, dans ce cas, un ralentissement ou une 'rétraction' dans le développement de ce marché. L'étude de la Société des Actuaires et de l'Institut Canadienne des Actuaires²⁷ renforce notre hypothèse, en expliquant qu'il existe une relation directe et même significative entre une autre forme d'assurance, soit celle de l'automobile et habitation et le taux d'inflation.

FINANC_DEV

Le développement du marché financier peut être mesuré par la proportion de l'offre de la monnaie (M2) et le PIB et son unité est le dollar américain. Le choix de cette variable est basé sur certaines recherches déjà effectuées, notamment celle de (Dieng & Fall, 2012) qui établit un lien positif, bien que non significatif, avec la demande en assurance-vie. La mesure du développement du marché financier est obtenue de WorldBank²⁸ et elle démontre le développement des institutions, instruments et marchés du secteur financier, ainsi expliquant pourquoi nous supposons que plus elle est élevée, plus le PIB, Revenu National brut seraient élevés, et le plus bas seraient le taux d'inflation ou le taux de chômage dans un pays.

²⁶ <https://www.insee.fr/fr/metadonnees/definition/c1473>. 2020/02/10.

²⁷ Ahlgrim, Kevin C., et D'Arcy, Stephen P. 2012. The effect of Deflation or High Inflation on the Insurance Industry. Casualty Actuarial Society, Canadian Institute of Actuaries, Society of Actuaries. (February). URL: <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjS05e-OAhVKUxoKHbb9BHsQFjAAegQIAEAC&url=https%3A%2F%2Fwww.soa.org%2FFiles%2FResearch%2FProjects%2Fresearch-2012-02-effect-deflation-report.pdf&usg=AOvVaw0GhuCooZumDbh00HLoXbb2>

²⁸ WorldBank:Data Bank. <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS> (2020/02/02) (Group, s.d.)

GDP / PIB

Le produit intérieur brut est un indicateur économique qui nous donne des informations précises sur la richesse produite au cours de l'année dans un pays quelconque. Plus précisément, cet indicateur va nous permettre de connaître la valeur ajoutée totale des biens et des services produits sur un territoire national. On l'utilise lorsque l'on veut connaître la croissance économique d'un pays. Le rapport PIB par habitant qui se calcul en faisant la somme des différentes valeurs ajoutées de tous les agents économiques, nous donne ainsi des informations sur le niveau de vie des habitants²⁹. Également utilisé comme variable explicative par (Dieng & Fall, 2012), ces derniers concluent, lors de leur analyse, qu'elle a une corrélation positive mais non significative avec le développement du marché financier. Ceci justifie le choix de cette variable dans notre étude.

YOUNG_DEP

Cette variable représente la proportion d'individus de moins de 18 ans sur la population active, soit celle âgée entre 18 et 64 ans. Le choix de cette variable est basé sur les études de (Thorsten & Webb, 2002) qui ont trouvé qu'elle avait un impact significatif sur la demande de l'assurance-vie. Effectivement, lorsque la proportion de 'jeunes' individus dans un pays est grande, il y a moins d'individus majeurs en mesure d'acheter des produits d'assurances. De ce fait, nous supposons qu'il existe une relation négative entre cette variable et notre variable à expliquer.

OLD_DEP

La proportion d'individus retraités, âgée de plus de 64 ans sur la population active (entre 18 et 64 ans) est une variable considérée comme ayant un impact significatif et positif sur la demande et ainsi que le développement du marché d'assurance-vie. Cette hypothèse est émise de par le fait que les pays ayant un plus gros pourcentage de personnes de plus de 64 ans seraient les pays où la population est en meilleure santé et a donc une meilleure espérance de vie. De plus, tel que démontré par (Dieng & Fall, 2012), un pourcentage plus

²⁹ <https://www.futura-sciences.com/planete/definitions/developpement-durable-pib-6295/>.2020/02/10.

élevé de ce groupe d'individus impliquerait une plus grande consommation de produit d'assurance-vie, notamment, dans le cas de décès. Ceci entraînerait une plus grande demande et une expansion de ce marché.

GNI

Le Revenu National Brut est la somme du PIB et des revenus étrangers nets. Contrairement au PIB, elle ne considère ainsi pas uniquement les revenus des individus et entreprises intérieurs mais également ceux gagnés à l'extérieur du pays en question. Ainsi, bien que nous estimions voir une très grande corrélation entre le PIB et notre variable GNI, nous utiliserons les deux mesures dans notre étude afin de mieux évaluer laquelle aurait le plus d'impact ou de significativité sur notre variable à expliquer.

IV- ANALYSE EXPLORATOIRE

ANALYSE UNIVARIEE

Toutes les variables utilisées dans cette étude sont quantitatives et continues.

En analysant les statistiques descriptives des données (Figure 1) nous remarquons que pour certaines variables, comme LIFE_EXP, GOOD_HEALTH, il y a peu de dispersion par rapport à la moyenne. En effet cette moyenne est très proche de la médiane. À l'inverse les variables affichant le plus de dispersion sont GDP et GNI, ce qui n'est pas très étonnant compte tenu des variations entre les niveaux de développement des pays que nous analysons dans cette étude. Ainsi, nous comparons le développement du marché de l'assurance-vie entre, la Norvège qui a le PIB le plus élevé (81 100 \$) et Porto Rico qui a le PIB le plus bas (2 939\$).

Figure 1 : Statistiques descriptives

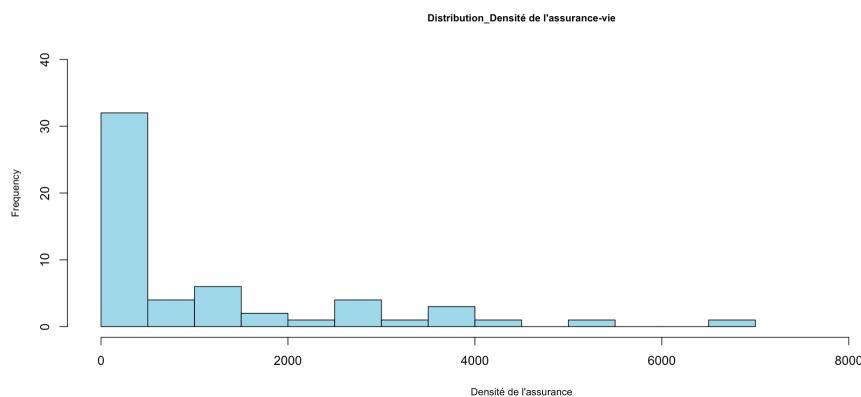
COUNTRY	INS_DEN	LITERACY	GOOD_HEALTH	URBAN_POP	LIFE_EXP
Length:56	Min. : 2.72	Min. : 63.00	Min. : 54.40	Min. : 48.30	Min. : 65.00
Class :character	1st Qu.: 61.27	1st Qu.: 90.09	1st Qu.: 66.42	1st Qu.: 63.63	1st Qu.: 74.79
Mode :character	Median : 304.24	Median : 93.73	Median : 69.60	Median : 75.98	Median : 78.84
	Mean : 1128.41	Mean : 92.41	Mean : 68.88	Mean : 74.47	Mean : 77.79
	3rd Qu.: 1596.18	3rd Qu.: 96.00	3rd Qu.: 71.92	3rd Qu.: 85.61	3rd Qu.: 81.33
	Max. : 6896.94	Max. : 100.00	Max. : 76.00	Max. : 100.00	Max. : 83.33
UNEMP	OLD_DEPENDENTS	INFLATION	FINANC_DEV	GNI	YOUNG_DEPENDENTS
Min. : 1.593	Min. : 7.052	Min. : -0.3242	Min. : 25.85	Min. : 1520	Min. : 19.95
1st Qu.: 3.529	1st Qu.: 11.636	1st Qu.: 1.2821	1st Qu.: 44.50	1st Qu.: 9960	1st Qu.: 22.60
Median : 5.571	Median : 21.919	Median : 2.0115	Median : 58.72	Median : 18725	Median : 28.19
Mean : 6.323	Mean : 20.489	Mean : 2.8070	Mean : 74.18	Mean : 28508	Mean : 31.60
3rd Qu.: 7.481	3rd Qu.: 28.030	3rd Qu.: 3.8260	3rd Qu.: 100.00	3rd Qu.: 47402	3rd Qu.: 39.97
Max. : 23.000	Max. : 42.653	Max. : 10.9076	Max. : 170.62	Max. : 104340	Max. : 61.14
GDP					
Min. : 2939					
1st Qu.: 16608					
Median : 35343					
Mean : 41874					
3rd Qu.: 70851					
Max. : 81100					

Source : Sortie R

Ces observations sont confirmées avec l'utilisation de boxplots, (Annexe 1), où nous voyons de plus que certaines variables comme le taux d'inflation (INFLATION) ou encore les ratios de jeunes (YOUNG_DEPENDENTS) et des personnes âgées (OLD_DEPENDENTS) ont des données moins centrées au niveau de la moyenne, mais plutôt réparties du quartile inférieur.

En regardant de plus près la variable à expliquer, INS_DEN (Figure 2), nous remarquons un écart relativement important entre la moyenne (1128 \$) et le médiane (304\$). En utilisant l'histogramme ci-dessous nous voyons qu'effectivement plus de la moitié des pays affiche une densité d'assurance-vie de moins de 500\$, suggérant que pour au moins la moitié des pays, les primes d'assurance-vie payées par les habitants sur toute l'année restent relativement bas. D'ailleurs, nous voyons que très peu de pays (à peu près 1% des pays) affichent une densité supérieure à 6000 \$ par habitant.

Figure 2



Source : Sortie R

ANALYSE BIVARIE

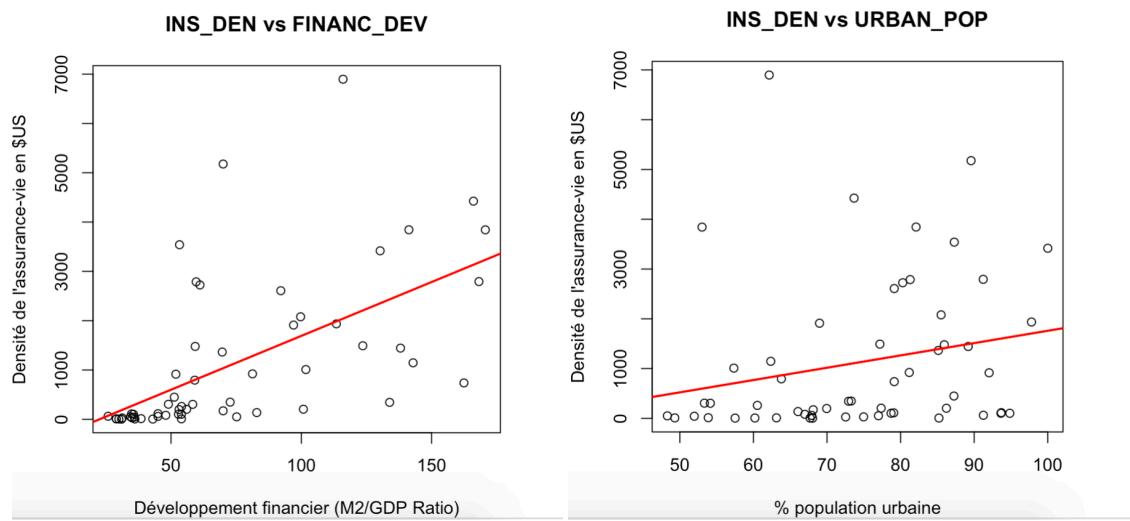
Lors de l'analyse économique certaines hypothèses ont été émises sur la corrélation entre les différentes variables utilisées, et cela, à partir de différentes études faites dans le passé. Dans cette section nous analyserons plus en détail les vraies relations entre ces variables.

En nous basant sur les nuages de points (Annexe 2), nous voyons que les variables FINANC_DEV, GOOD_HEALTH, LIFE_EXP et GNI sont toutes positivement corrélées entre elles. Ainsi, une augmentation dans le développement du secteur financier augmente le revenu des habitants du pays (Annexe 4), ainsi que le pourcentage de personnes en bonne santé et leur espérance de vie. À l'inverse, il n'est pas surprenant de voir que le revenu national brut soit

négativement corrélé au taux d'inflation, car plus ce dernier augmente plus le pouvoir d'achat des habitants diminue.

En regardant la relation entre les variables explicatives au développement du marché de l'assurance-vie, nous pouvons établir que plus un pays est développé et a ainsi un PIB plus élevé (Annexe 3), plus son secteur bancaire est développé ou encore plus il a un grand pourcentage de population urbaine (Figure 3), plus la demande des produits de l'assurance-vie est grande. Le pouvoir d'achat des habitants étant plus grand, ces derniers sont plus aptes à investir sur du long-terme.

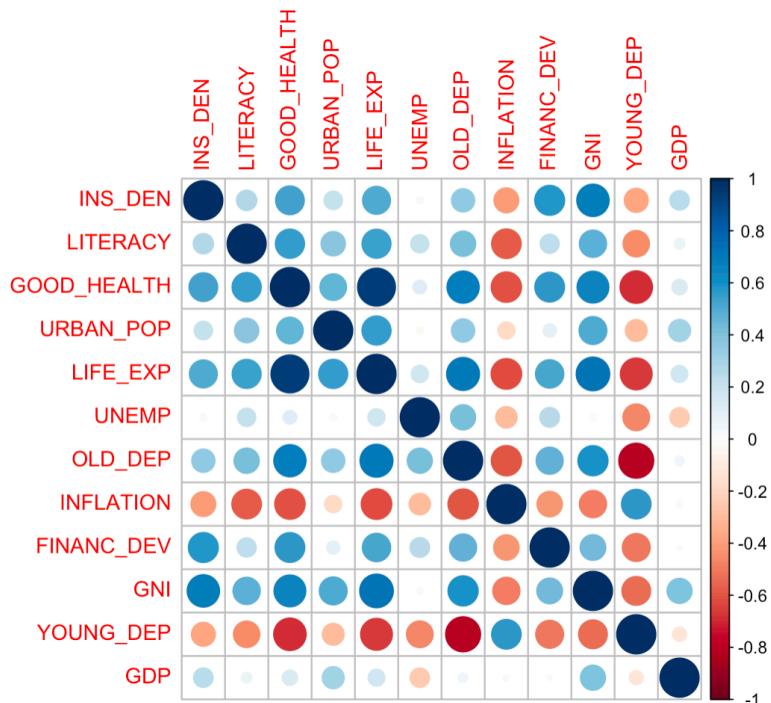
Figure 3: Nuage de points



Source : Sortie R

Le tableau de corrélation en faisant le test de Spearman (Annexe 2) nous confirme qu'il existe des corrélations significatives (au seuil à 5%) entre l'espérance de vie et certaines variables notamment, le pourcentage de personnes en bonne santé, ou encore la proportion de personnes âgées. D'ailleurs cette dernière est significativement négativement corrélée à la proportion de personnes moins de 18 ans, toujours au seuil de 5%. Le revenu national brut a également beaucoup de corrélations significatives au même seuil, avec les variables comme le PIB, le niveau de santé de la population et l'espérance de vie. Ce tableau nous indique que beaucoup de nos variables présentent un fort risque d'autocorrélation. Ceci devra être traité avec soin lors de la modélisation.

Figure 4 : Tableau de corrélation entre variables



Source : Sortie R

OBSERVATIONS ATYPIQUES ET ABERRANTES

Dans cette section, nous tentons de détecter la présence et d'éliminer, si possible, les valeurs aberrantes pouvant biaiser les résultats du modèle, et cela sans supprimer trop de données, crucial à la création d'un bon modèle.

Nous utilisons les boîtes à moustaches pour détecter les valeurs atypiques. L'annexe 1 nous indique que les variables LITERACY, GOOD_HEALTH, UNEMP, INFLATION et GNI ont des points atypiques. Afin de connaître plus précisément les observations aberrantes, nous utilisons deux tests : le test de Grubbs, qui est utilisé lorsque la boîte à moustache met un seul point en évidence (ce qui indique qu'il y a 1 point atypique), et le test de ESD (« Extrem Studentized Deviation ») qui est un test de déviation extrême de Student nous permettant de détecter simultanément plusieurs observations atypiques en même temps. Ce dernier est utilisé lorsque la boîte à moustache nous affiche plusieurs points ainsi indiquant un risque éventuel d'avoir plusieurs 'outliers' présents dans notre variable.

En faisant les tests nous trouvons qu'il y a 5 observations atypiques au total, que nous décidons d'enlever car elles représentent moins de 10% de la base initiale.

C'est à partir de cette nouvelle base finale, ayant maintenant 51 pays et toujours 12 variables (incluant la variable illustrative) que nous nous lançons dans l'analyse économétrique de l'étude.

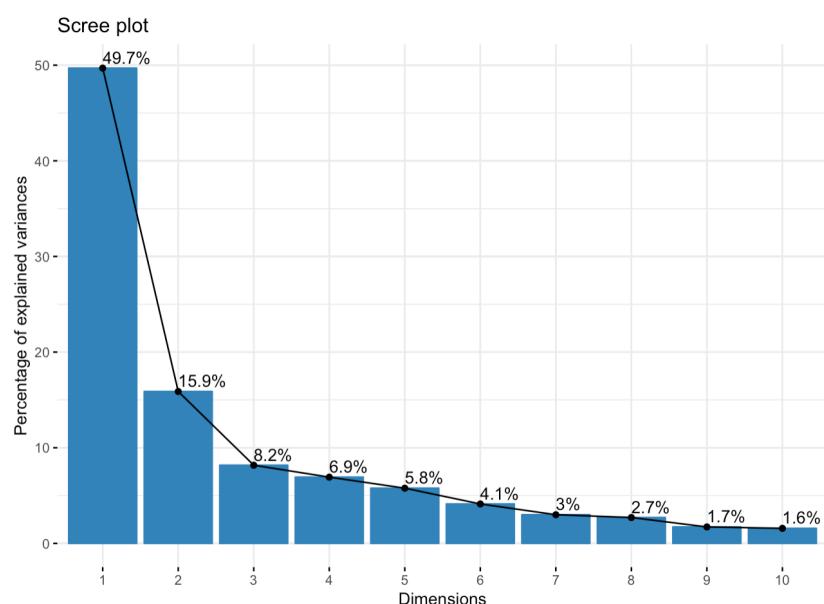
REDUCTION DE DIMENSION

Lors du test de corrélation de Spearman, plusieurs variables étaient significativement corrélées entre elles. Afin de diminuer les risques d'autocorrélation et de ne garder que les variables les plus significatives par rapport à la variable expliquée, nous cherchons, dans cette section, à réduire le nombre de dimensions en utilisant une analyse en composantes principales. L'objectif de cette analyse est de réduire le nombre de variables tout en déformant le moins possible l'information, l'inertie totale initiale.

Le tableau de valeurs propres (Annexe 5) nous affiche que seulement 69.06% de l'inertie totale est expliquée par les deux premières composantes principales, et 78.05% est expliquée par les 3 premières composantes. En utilisant la règle du coude nous décidons de ne retenir que les deux premiers axes. Ainsi nous retenons 69.06% de la variance, un pourcentage qui n'est pas très élevé. Nous utiliserons d'autres méthodes lors de la création du modèle pour confirmer ou non ce résultat.

De plus ce choix est confirmé par la règle de Kaiser qui nous indique de garder les composantes ayant une valeur propre strictement supérieure à 1, soit les deux premières composantes.

Figure 5 : Histogramme des pourcentage des variances expliquées



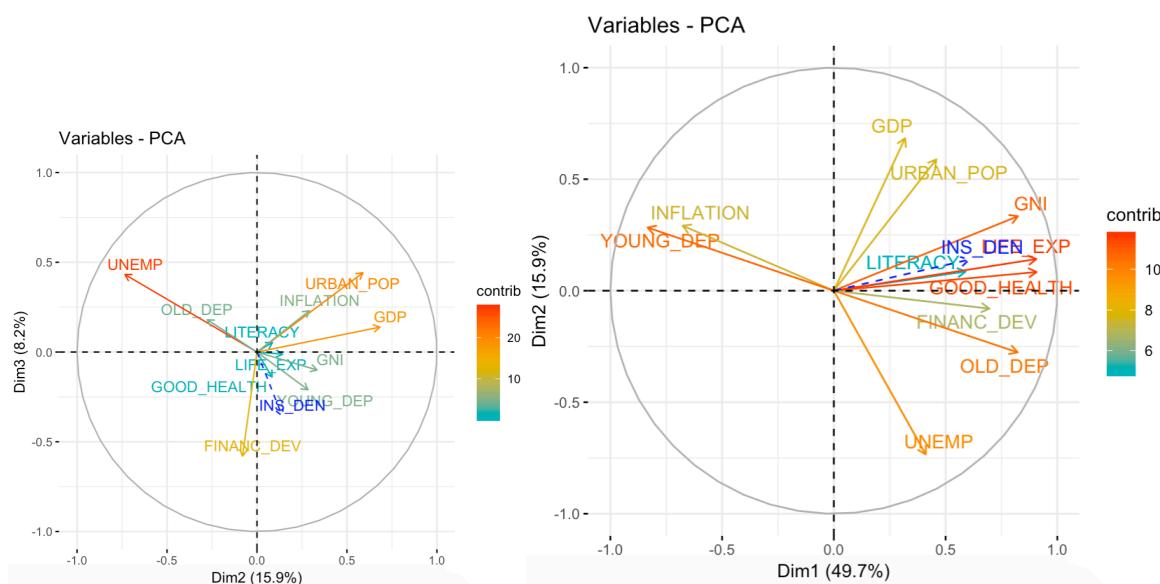
Source : Sortie R

L'analyse en composantes principales (ACP) nous donne ainsi 3 possibilités de cercles de corrélation (Figure 6). Certaines variables sont très distinctement regroupées entre elles, nous aidant à confirmer leurs corrélations. La variable INS_DEN est ‘moyennement’ représentée sur le plan (1,2) suggérant que, bien que cette dernière ait la plus grande inertie projetée, elle n'affiche pas forcément les variables impactant le plus notre variable expliquée.

Les variables explicatives apportant le plus de contributions à l'axe 1 (Figure 6, plan factoriel 1-2) sont OLD_DEPENDANT, LIFE_EXP, GNI, GOOD_HEALTH, FINANC_DEV, YOUNG_DEPENDANT, INFLATION. Tandis que les cinq premières variables sont positivement corrélées à l'axe 1, les deux dernières, soient YOUNG_DEPENDANT et INFLATION font opposition à l'axe. De plus, l'ensemble de ces variables sont bien représentées dans ce plan, étant loin du centre du cercle. URBAN_POP contribue ‘moyennement’ à l'axe 1, nous l'utiliserons plutôt dans l'interprétation de l'axe 2. Finalement, la variable LITERACY ne sera pas utilisée dans notre interprétation car elle est la plus mal représentée sur le plan factoriel 1-2, étant plus proche du centre, et elle contribue très peu aux axes 1 ou 2.

Les variables les plus contributives à l'axe 2 sont GDP, URBAN_POP et UNEMP. Tandis que ces deux premières sont positivement corrélées à cet axe, UNEMP y fait opposition.

Figure 6: ACP – cercle de corrélation



Source : Sortie R

Nous voyons que OLD_DEP, LIFE_EXP, GNI, GOOD_HEALTH, FINANC_DEV, GDP, FINANC_DEV, URBAN_POP sont toutes non seulement positivement corrélées entre elles, mais également avec la variable illustrative INS_DEN. Ceci confirme les résultats du test de Spearman et du tableau de corrélation. Plus un pays est développé, plus la population urbaine est grande, plus le niveau de santé est élevé, et plus la proportion des personnes âgées est grande, nous remarquons qu'il y a une hausse dans la demande des produits d'assurances. Les habitants de ces pays ont tendance à investir à long terme. À l'inverse, le taux d'inflation et le plus grand pourcentage de moins de 18 ans auraient un impact négatif sur le développement du marché de l'assurance des pays. Effectivement, ces deux variables impliquent une diminution du pouvoir d'achat, d'où une grande réticence à investir.

Ces observations nous permettent de mieux interpréter les variables synthétiques.

La première serait **le niveau de développement d'un pays**. Effectivement, plus le pays est développé, plus son PIB, les revenus, le niveau de santé ainsi que le pouvoir d'achat des habitants sont élevés. Les individus ont les moyens de prendre soin d'eux et d'investir pour leurs futurs et ceux de leurs bénéficiaires.

La deuxième variable synthétique est **le développement des villes**. Cette interprétation est basée sur le fait que plus la ville est grosse, plus le pourcentage de la population urbaine sera grand et plus le taux d'emploi est grand. Les villes les plus développées, avec plus d'infrastructures et de plus grande taille, offrent plus d'emplois que les zones rurales et aident à faire augmenter le PIB du pays.

HIERARCHICAL AGGLOMERATIVE CLUSTERING

Nous voulions pour mieux visualiser les individus les classer dans différents groupes, une méthode le permettant est la méthode de classification ascendante hiérarchique.

Nous avons commencé à classer les individus selon les groupes avec des méthodes de classification supervisée tels que la régression logistique ou l'analyse factorielle discriminante. Cependant ces méthodes n'ayant pas donné de résultats satisfaisants (dû à la structure des données), nous avons considéré cette méthode de classification hiérarchique.

Cette méthode de classification non supervisée consiste à placer les individus dans un certain nombre de clusters préalablement choisi. Ces clusters sont calculés en fonction des distances entre individus.

Ces regroupements successifs produisent un arbre binaire de classification.

Ainsi si nous choisissons un nombre de cluster égal à 4, l'utilisation de 5 clusters ne donnant pas de résultats satisfaisants, nous obtenons l'arbre et les groupes suivants :

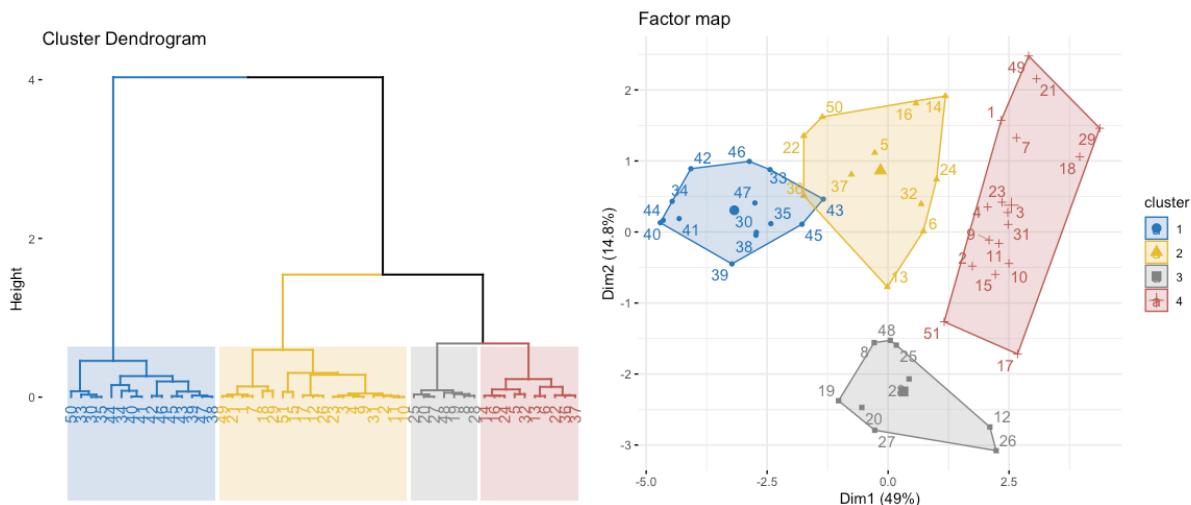


Figure 15: Arbre de classification (à gauche) et visualisation des 4 clusters (à droite)

Voici les 4 clusters représentés sur les axes F1-F2. En annexe 14 nous retrouvons un tableau avec les pays représentés pour chaque cluster.

Le 1^{er} cluster (en bleu) représente les pays avec un taux d'inflation plus élevé et une population jeune. Ce cluster regroupe 23.5% des pays issus de la base de données, par exemple : Bolivie, Malaisie, Indonésie. Ce sont ces pays où la densité d'assurance est la plus faible.

Le 2nd cluster (en jaune) représenterait le développement des villes et la richesse produite par le pays. Les pays représentés par ce cluster ont des villes de plus grosses tailles et un PIB plus important. Ce sont des pays comme le Brésil, l'Uruguay, ou l'Argentine où la population urbaine y est plus importante

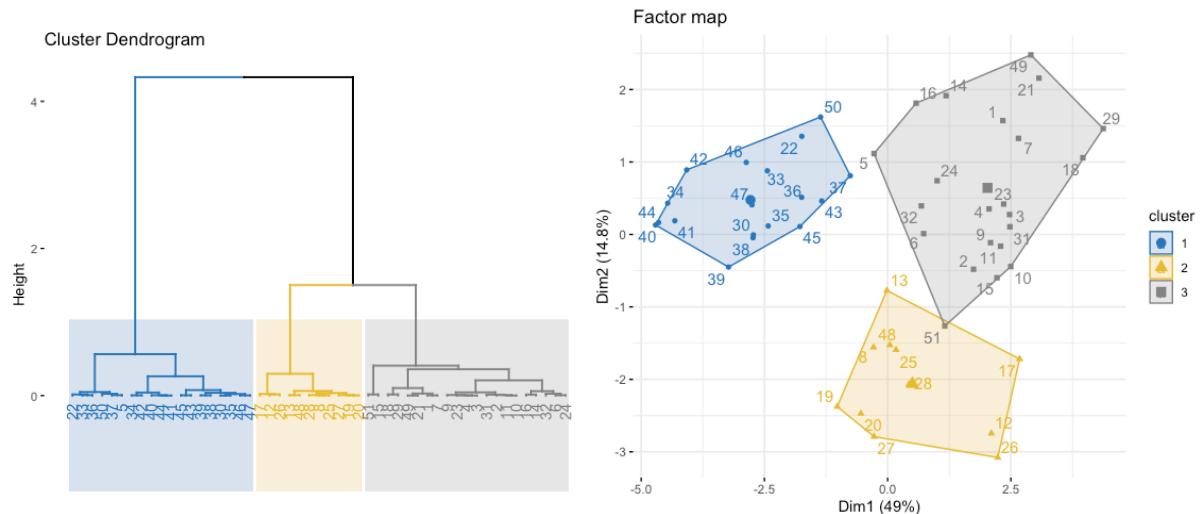
Ce cluster regroupe 13.7% des pays issus de la base de données.

Le 3^{ème} cluster (en gris) représenterait les pays avec un taux de chômage plus élevé : comme la Lituanie, l'Estonie. Ce cluster regroupe 21.56% des pays de la base de données.

Le 4^{ème} cluster (en rouge) représenterait les pays les plus développés avec une population en meilleure santé et vivant plus longtemps, un PIB plus important. Ce cluster regroupe 41.17% des pays de la base de donnée, par exemple : la France, l'Allemagne , l'Australie, le Canada

Ce sont les pays où il y a une plus grande densité d'assurance

Nous pouvons également classer les pays dans 3 clusters :



Voici les 3 clusters représentés sur les axes F1-F2. On peut voir en annexe 15 un tableau avec les pays représentés par chaque cluster.

Le 1^{er} cluster (en bleu) regroupe 32.29 % des pays de cette base de données, il pourrait de nouveau représenter les pays avec un taux d'inflation plus élevé et une population jeune : on retrouve par exemple un certain nombre de pays d'Amérique du sud (Colombie, Brésil, Argentine). Ce sont des pays où la densité d'assurance est la plus faible.

Le 2nd cluster (en jaune) regroupe 21.57% des pays de cette base de données, représenterait les pays avec un taux de chômage plus élevé. On retrouve l'Italie, la Pologne, l'Estonie ou encore le Portugal.

Le 3^{ème} cluster (en gris) regroupe 43.13 % des pays de cette base de données, représenterait les pays les plus développés avec un PIB plus élevé, une population urbaine plus importante, un revenu national et un développement du secteur financier là aussi important. Ce sont des pays où la densité d'assurance est élevée. La population est en meilleure santé et a une plus grande espérance de vie. Ce sont des pays comme les Etats Unis, l'Angleterre, le Japon, la France ou encore l'Allemagne.

Les habitants des pays représentés par ce troisième cluster pourraient être ceux investissant le plus pour leurs assurances.

V- ANALYSE ECONOMÉTRIQUE

ESTIMATION DU MODELE

Ayant fait une première analyse des données nous cherchons maintenant à trouver un modèle adapté. L'objectif est de créer différents modèles en utilisant plusieurs techniques, de comparer les résultats, et finalement de choisir le modèle qui représente le mieux la relation entre nos variables. Les différentes méthodes utilisées dans notre étude sont :

1. La régression multiple en utilisant AIC
2. La régression par composantes principales
3. Méthode 'Partial Least Squares'

REGRESSION LINEAIRE – AIC

Afin de réduire les risques de colinéarité entre les variables explicatives, nous utilisons la fonction 'step' de R qui utilise l'AIC, le critère d'information d'Akaike qui donne une mesure de la qualité du modèle. Cela nous permet de réduire les variables colinéaires entre elles, tout en faisant attention de prioriser celles étant les plus significatives à INS_DEN.

Nous obtenons le modèle suivant :

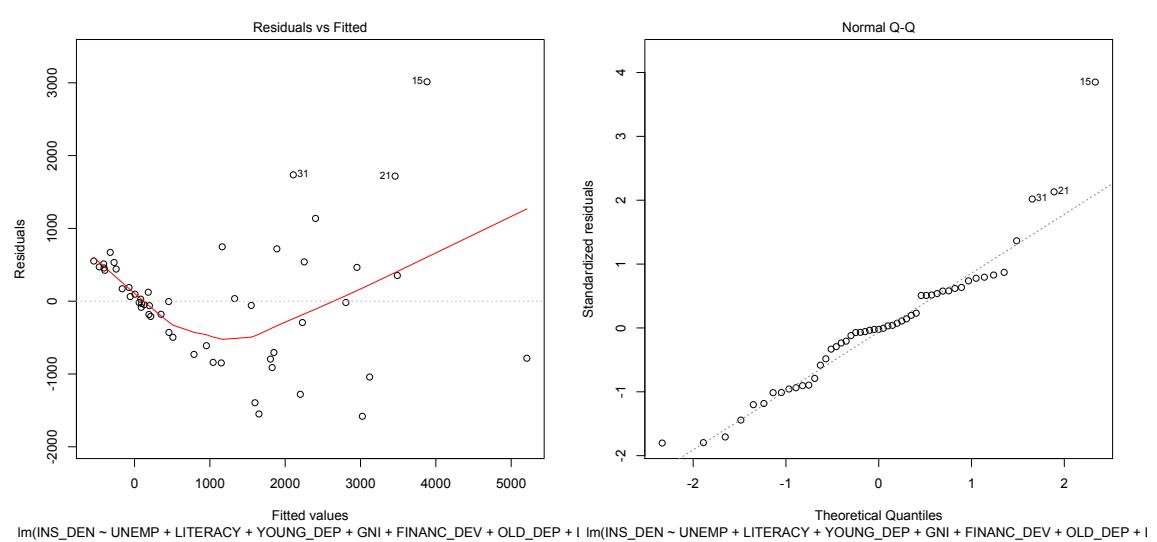
$$Y = \alpha + \beta_1 (\text{LITERACY}) + \beta_2 (\text{GOOD_HEALTH}) + \beta_3 (\text{LIFE_EXP}) + \beta_4 (\text{UNEMP}) + \beta_5 (\text{OLD_DEP}) + \beta_6 (\text{FINANC_DEV}) + \beta_7 (\text{YOUNG_DEP}) + \beta_8 (\text{GNI}) + \varepsilon$$

Selon l'Annexe 16, le modèle ayant le plus petit AIC, soit 703.93, est celui qui est retenu.

Avant d'interpréter le modèle nous allons tout d'abord effectuer des tests afin de déterminer si la méthode de moindres carrés ordinaires (MCO) pourrait être appliquée.

En premier lieu, en effectuant les résultats du test de Cooks (Annexe 7) nous voyons que la distance est plus petite que 1. En effet selon l'annexe 7 nous voyons qu'il existe un seul résidu (le numéro 15) qui se trouve entre un intervalle de 0.5 et 1.0 d'intervalle de confiance. Ceci implique qu'aucune observation influence l'estimation du modèle. Le graphique de 'Residual versus Fitted' (Figure 7, à gauche) nous montre que les résidus 31, 12 et 15 sont dispersés bien plus haut que la ligne, suggérant que plus la variance augmente plus on va à droite. Ceci nous montre qu'il y a un possiblement un problème d'hétéroscédasticité. De plus, le 'Normal Q-Q plot' (Figure 7, droite) nous affiche que ces mêmes résidus sont également dispersés plus haut que la ligne théorique. Cette fois-ci cela nous montre que les résidus ne suivraient peut-être pas une loi gaussienne. Une transformation des variables serait peut-être nécessaire afin de résoudre à ce problème.

Figure 7 : 'Residual vs Fitted Graph' (gauche), et 'Normal Q-Q' (droite)



Afin de valider ces résultats graphiques nous cherchons à effectuer les tests nécessaires. Tout d'abord nous faisons le test de RESET qui nous permet de vérifier la forme linéaire du modèle. Nous obtenons une p-valeur de 0.001802 (Figure 8), qui est plus bas que 0.05 au seuil de 5% de tolérance. Ainsi, nous rejetons l'hypothèse nulle H0 : « la forme fonctionnelle est linéaire ». Le modèle n'ayant donc pas de forme linéaire, nous confirmons le besoin de transformer les variables. Puis le test VIF, qui nous permet de tester la colinéarité entre les variables, nous affiche que les variables LIFE_EXP et GOOD_HEALTH sont bel et bien colinéaires ($VIF > 10$), comme nous l'avions supposé lors de l'analyse exploratoire. Nous savons d'ores et déjà qu'une de ces deux variables devra être enlevée.

Nous poursuivons avec le test de Breusch-Pagan (Figure 8) qui cette fois-ci affiche une p-valeur de 0.01188, qui est inférieur à 0.05 au seuil de 5%. Notre modèle présente effectivement un problème d'hétérosécédasticité, comme indiqué par les graphiques (Figure 7). Finalement, le dernier test effectué est le Shapiro test (Figure 8). Celui-ci nous affiche que les résidus ne suivent pas une loi gaussienne, car la p-valeur de 0.01168 est inférieur à 0.05 au seuil de 5%, nous poussant à rejeter H0. Nous espérons qu'un changement dans la forme fonctionnelle du modèle pourrait nous aider à régler ce problème.

Figure 8 : Résultat des tests (Reset, Breusch-Pagan, Shapiro-wilk)

RESET test

```
data: modele1
RESET = 7.431, df1 = 2, df2 = 40, p-value = 0.001802
```

```
vif(modele1)
UNEMP    LITERACY   YOUNG_DEPEND      GNI   FINANC_DEV     OLD_DEPEND    LIFE_EXP  GOOD_HEALTH
2.382986 1.441003  4.231563  4.040221  2.125332  3.397998  10.328202 10.130680
```

studentized Breusch-Pagan test

```
data: modele1
BP = 19.618, df = 8, p-value = 0.01188
```

Shapiro-Wilk normality test

```
data: residus
W = 0.93945, p-value = 0.01168
```

Afin de résoudre ces problèmes nous faisons plusieurs transformations dans les variables. En premier lieu, nous enlevons la variable LIFE_EXP car elle est colinéaire à GOOD_HEALTH. Comme les résultats des tests (Annexe 9) du modèle ne sont pas tous satisfaisants pour appliquer la méthode de MCO (bien que le test VIF soit maintenant satisfait), nous transformons la variable INS_DEN en logarithme. Toutefois, les résultats des nouveaux tests ne sont pas encore satisfaisants (Annexe 10). Il faudra transformer en logarithme les variables INS_DEN et GNI pour que le modèle final satisfasse toutes les conditions du MCO. Effectivement, comme nous le montre Figure 9, les tests de RESET, Breusch-Pagan et Shapiro-Wilk ont maintenant tous des p-valeurs supérieur à 0.05 au seuil de 5%, et toutes les variables ont des VIF inférieur à 10 ou même 5.

Figure 9: Résultat des tests du modèle final

```

RESET test

data: modele4
RESET = 1.2349, df1 = 2, df2 = 41, p-value = 0.3015

> vif(modele4)
    UNEMP    log(GNI)  FINANC_DEV      OLD_DEP    YOUNG_DEP  GOOD_HEALTH    LITERACY
    1.903974   4.579377   2.082096    3.493468    4.086295   4.270486    1.468461
> bptest(modele4)

studentized Breusch-Pagan test

data: modele4
BP = 7.3537, df = 7, p-value = 0.393

> residus<-residuals(modele4)
> shapiro.test(residus)

Shapiro-Wilk normality test

data: residus
W = 0.97741, p-value = 0.4354

```

Source : Sortie R

Suite aux résultats nous pouvons garder ce modèle. Toutefois, lors de l'analyse économique nous avons suspecté un problème d'endogénéité entre notre variable expliquée, INS_DEN, et les variables FINANC_DEV et GOOD_HEALTH. Comme ces deux variables explicatives ne sont pas corrélées (confirmé par le test VIF), nous pouvons effectuer un test d'endogénéité sur les

deux en même temps. Afin de choisir le bon instrument, nous cherchons parmi les variables déjà choisies dans notre modèle. Le but est de choisir une variable ‘remplaçante’ qui serait corrélée et qui serait même une causalité de notre variable suspectée d’endogénéité. Nous voulons effectuer trois tests, le ‘Weak Instrument’, le test de Sargan et le test de Wu-Hausman. Le test de Sargan teste la qualité des instruments utilisés, et cela en vérifiant si la corrélation entre les résidus de l’instrument et de la variable à être possiblement remplacée. Toutefois, comme cela ne requiert l’utilisation d’un seul instrument pour remplacer nos deux variables FINANC_DEV et GOOD_HEALTH, nous ne sommes pas en mesure de l’effectuer. Effectivement, dans notre cas, nous cherchons à utiliser un instrument par variable potentiellement endogène.

Lors de l’analyse économique et exploratoire, nous avions vu que la variable FINANC_DEV était corrélée avec la variable URBAN_POP. Ceci est parce que plus le pourcentage de la population urbaine est grand, plus il y a des banques et des institutions financières qui sont d’ailleurs centralisées dans les villes. Nous avions ainsi pensé qu’une population urbaine en croissance causerait l’expansion du secteur financier. C’est pour cela que la population urbaine serait un instrument dans ce cas.

Nous avions également vu que les variables LIFE_EXP et GOOD_HEALTH étaient très corrélées. De plus, un individu en bonne santé aura plus de chance d’avoir une plus grande espérance de vie. Ainsi, la variable LIFE_EXP sera utilisée comme instrument.

Figure 10: Résultat des tests d’endogénéité

```

Call:
ivreg(formula = log(INS_DEN) ~ UNEMP + log(GNI) + FINANC_DEV +
      OLD_DEP + YOUNG_DEP + GOOD_HEALTH + LITERACY | UNEMP + log(GNI) +
      URBAN_POP + OLD_DEP + YOUNG_DEP + LITERACY + LIFE_EXP, data = baseFinal)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.5057 -0.5242 -0.1474  0.9074  2.1617 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -4.165212  10.530843 -0.396 0.694411    
UNEMP        0.100741   0.083814  1.202 0.235955    
log(GNI)      1.549578   0.394584  3.927 0.000306 ***  
FINANC_DEV    0.037140   0.019650  1.890 0.065501 .    
OLD_DEP       -0.045544   0.036339 -1.253 0.216864    
YOUNG_DEP     0.002876   0.033810  0.085 0.932600    
GOOD_HEALTH   -0.118505   0.192143 -0.617 0.540650    
LITERACY      0.002069   0.045375  0.046 0.963846   

Diagnostic tests:
                    df1 df2 statistic p-value    
Weak instruments (FINANC_DEV) 2   43      5.308  0.0087 **  
Weak instruments (GOOD_HEALTH) 2   43      36.071 6.35e-10 ***  
Wu-Hausman                  2   41      2.208  0.1228    
Sargan                      0   NA      NA      NA      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.099 on 43 degrees of freedom
Multiple R-Squared: 0.7818,   Adjusted R-squared: 0.7462 
Wald test: 21.58 on 7 and 43 DF,  p-value: 3.675e-12

```

La Figure 10 nous affiche les résultats des tests effectués. Le test de Weak Instrument pour la variable de FINANC_DEV est satisfait au seuil de 5% avec une p-valeur de 0.0087, tandis que celui de GOOD_HEALTH est satisfait au seuil de 1%. Ces résultats confirment que URBAN_POP et LIFE_EXP sont de bons instruments. En faisant le test de Wu-Haussman toutefois nous voyons que nos variables FINANC_DEV et GOOD_HEALTH qui étaient suspectées d'endogénéité sont en fait exogènes. En effet, la p-valeur du test (0.1228) est supérieure à 0.05 ou 0.1 à leurs seuils respectifs.

Après ces tests, avec la confirmation que toutes nos variables sont bel et bien exogènes, nous pouvons interpréter le modèle final soit :

$$\begin{aligned}
\log(\text{INS_DEN}) \sim & 0.061925 * (\text{UNEMP})_t + 1.194863 * (\log(\text{GNI}))_t + 0.018071 * (\text{FINANC_DEV})_t - \\
& 0.032030 * (\text{OLD_DEP})_t - 0.004088 * (\text{YOUNG_DEP})_t + 0.123975 * (\text{FINANC_DEV})_t + \\
& 0.004079 * (\text{LITERACY})_t + \epsilon,
\end{aligned}$$

Ce modèle est obtenu du sommaire des résultats (Figure 11), nous voyons que le coefficient de détermination est 0.8494 (le coefficient ajusté étant de 0.8249). Ce résultat nous semble très satisfaisant. Avec uniquement 7 variables nous sommes capables d'expliquer 84.94% de

la variation de l'INS_DEN. La p-valeur du test de Fisher est 1.116 e-15, une valeur étant plus petite que 0.01 à 1% de tolérance. Ceci suggère qu'il existe au moins une variable significative à la densité de l'assurance-vie, soit la prime d'assurance par habitant. Effectivement, si nous regardons plus en détails les variables, nous voyons que log(GNI), et FINANC_DEV sont les plus significatives (au seuil de 1%) par rapport à notre variable expliquée. Les autres variables ne sont toutefois pas significatives, et cela même au seuil de 10%.

Figure 11 : Sommaire des résultats du modèle final

```

Call:
lm(formula = log(INS_DEN) ~ UNEMP + log(GNI) + FINANC_DEV + OLD_DEP +
    YOUNG_DEP + GOOD_HEALTH + LITERACY, data = baseFinal)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3433 -0.4109  0.0153  0.6323  1.5512 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -16.070853   5.437535 -2.956 0.005049 ** 
UNEMP        0.061925   0.057487  1.077 0.287392    
log(GNI)      1.194863   0.271204  4.406 6.89e-05 ***  
FINANC_DEV    0.018071   0.004487  4.028 0.000225 ***  
OLD_DEP       -0.032030   0.027274 -1.174 0.246706    
YOUNG_DEP     -0.004088   0.025351 -0.161 0.872663    
GOOD_HEALTH    0.123975   0.078042  1.589 0.119483    
LITERACY      0.004079   0.037380  0.109 0.913604    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.913 on 43 degrees of freedom
Multiple R-squared:  0.8494,    Adjusted R-squared:  0.8249 
F-statistic: 34.65 on 7 and 43 DF,  p-value: 1.116e-15

```

Source : Sortie R

Les deux seules variables significatives de notre modèle ont un impact positif sur la demande en assurance-vie d'un pays. Ainsi, lorsque le revenu national brut par habitant augmente, la prime d'assurance-vie par habitant augmente significativement. Plus exactement, une augmentation d'un pourcent du revenu national brut par habitant fait augmenter de 1.19% la prime d'assurance par habitant, toutes choses étant égales par ailleurs. En effet, la consommation de produits d'assurance vie augmente car les individus ont un plus grand pouvoir d'achat. Ce résultat concorde avec les hypothèses émises lors de l'analyse économique. De plus, nous pouvons dire que le coefficient du GNI est l'élasticité de celui-ci par rapport à la demande d'assurance.

En ce qui concerne la variable FINANC_DEV (représentée par le ratio d'offre de la monnaie et le PIB), nous voyons qu'une augmentation d'un dollar de cette variable fait augmenter de 1.8% la densité d'assurance-vie par habitant, toutes choses étant égales par ailleurs. Le coefficient de cette variable mesure la semi-élasticité de INS_DEN par rapport à elle. La relation positive entre les deux variables n'est pas étonnante car plus le secteur bancaire et financier est développé, plus la quantité de produits d'investissements comme l'assurance-vie est grande, ainsi engendrant une plus grande demande dans ce milieu.

Regardons maintenant quelques variables non-significatives à la variation de la demande en assurance-vie. Lors de l'analyse économique nous avions supposé que plus la population de jeunes ayant moins de 18 ans était grande, plus la demande en assurance-vie dans ce pays était petite. Cette hypothèse est confirmée dans notre étude (Figure 11) bien que ce résultat ne soit pas significatif. Nous ne pouvons pas en dire autant pour la variable UNEMP. Bien que nous supposions que le taux de chômage diminuerait la demande en assurance-vie, la Figure 11 nous affiche une relation positive. Finalement, nous voyons que le pourcentage de personnes en bonne santé dans un pays ou encore le taux d'alphabétisation ont un impact positif et non significatif par rapport à la demande de l'assurance – vie, tandis que le pourcentage de personnes âgées fait baisser la demande en assurance -vie (contrairement à notre hypothèse).

REGRESSION SUR COMPOSANTES PRINCIPALES

La régression sur les composantes principales consiste à faire une régression sur des dimensions préalablement réduites par l'Analyse en Composantes Principales. C'est une technique qui peut se faire automatiquement en utilisant R ou encore manuellement, soit en appliquant manuellement la régression sur un modèle réduit. En faisant ce type de régression nous gardons toutefois en tête que le modèle créé déformerai le moins possible la variance, car il ne gardera que les plus grandes valeurs propres. Toutefois, il ne prend pas en considération les variables les plus significatives par rapport à la variable INS_DEN. En effet, le principe d'ajustement repose sur la maximisation de la variance des variables explicatives indépendamment de la variable dépendante.

À la lumière des constatations faites lors de la régression multiple, nous commençons d'abord par chercher le modèle qui vérifie les hypothèses de linéarité de la forme fonctionnelle et de

normalité des résidus. Après avoir effectué les tests nécessaires, nous aboutissons à la même conclusion que précédemment : une transformation logarithmique des variables INS_DEN et GNI conduit au bon modèle. On débute ensuite la méthode automatique du PCR en sélectionnant les 3 premières composantes principales. Ce résultat est obtenu en utilisant la méthode de validation croisée (Figure 12), qui consiste à trouver le nombre de composantes qui minimise le critère de PRESS, soit la moyenne des carrés des écarts résiduels. Effectivement la Figure 12 ainsi que l'Annexe 6 nous montrent un CV ajusté de 1.004 à 3 composantes, ce qui est bien le résultat minimum (en comparant avec les autres nombres de composantes). De plus, la Figure 12 nous montre aussi qu'en retenant 3 composantes, nous expliquons 74,87% de la variance des variables explicatives et 81.13% de la variabilité du logarithme de la densité d'assurance-vie.

Figure 12: Résultat du PCR

```
Data: X dimension: 51 11
      Y dimension: 51 1
Fit method: svdpc
Number of components considered: 11

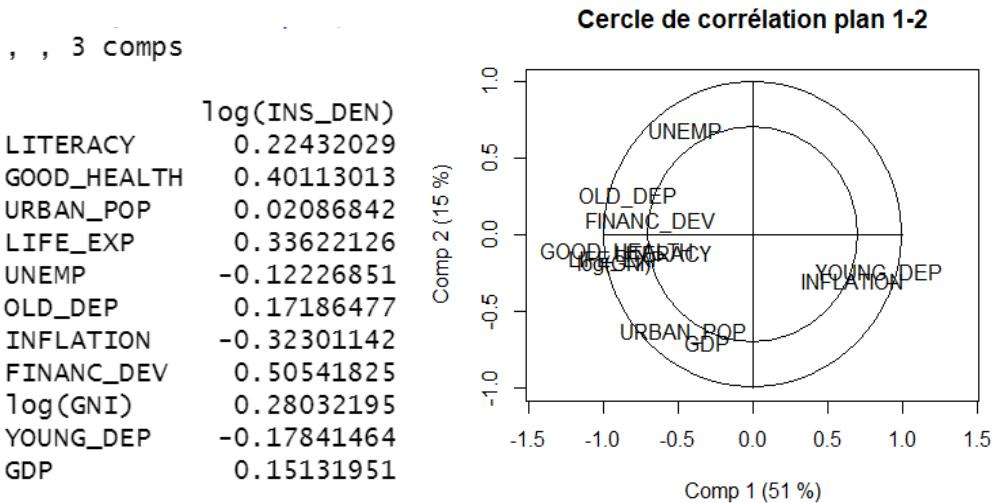
VALIDATION: RMSEP
Cross-validated using 51 leave-one-out segments.
              (Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps
CV            2.204   1.079   1.090   1.006   1.026   1.064   1.075   1.125   1.130
adjCV         2.204   1.078   1.089   1.004   1.025   1.062   1.074   1.123   1.128
              9 comps 10 comps 11 comps
CV            1.169   1.164   1.015
adjCV         1.167   1.161   1.012

TRAINING: % variance explained
              1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps 9 comps
X             51.46    66.68   74.87   81.81   87.60   91.67   94.57   96.94   98.53
log(INS_DEN) 76.82    77.29   81.13   81.21   81.27   81.88   82.01   82.28   82.29
              10 comps 11 comps
X             99.61    100.00
log(INS_DEN) 84.92    87.99
```

Source : Sortie R

La variable qui impacte le plus (et dans notre cas positivement) notre variable INS_DEN est FINANC_DEV selon Figure 13 (gauche). Ceci n'est pas surprenant car le développement du secteur financier implique l'augmentation de produits d'investissements, d'où d'assurance. Figure 13 (droite) confirme nos observations lors de l'analyse en composantes principales (plan 1-2) durant lequel nous avons trouvé que les deux premières composantes principales sont le niveau de développement d'un pays et le développement des villes.

Figure 13: Correlation entre variables sur le plan 1-2 (droite) et impact des variables sur INS_DEN pour 3 composantes (gauche)



Source : Sortie R

Analysons maintenant la 3^{ème} composante. Pour cela, nous utilisons les ‘loadings’ des variables par rapport à cette composante, telle qu’affichée par Figure 14, à gauche (qui nous montre les loadings sur chacune des 3 composantes). Nous utilisons comme seuil la valeur absolue de 0.25, c’est-à-dire des variables explicatives affichant un ‘loading’ de plus de la valeur absolue de ce nombre seraient les variables les plus contributives et de meilleure qualité par rapport à la 3^{ème} composante. Nous pouvons ainsi voir facilement que UNEMP, FINANC_DEV, URBAN_POP, et finalement INFLATION (bien que sa contribution ne soit pas si grande) sont les variables représentant le plus notre 3^{ème} variable synthétique. De plus, URBAN_POP, UNEMP et INFLATION ont un impact positif alors que FINANC_DEV a un impact négatif. Nous pouvons ainsi interpréter cette variable synthétique comme étant **l’insécurité financière**. En effet, une instabilité financière entraîne le chômage, et une hausse de l’inflation, ce qui diminue le pouvoir d’achat. Les habitants peuvent plus difficilement investir, ce qui crée une baisse de la demande des produits d’investissements, et entraîne une baisse dans le développement du secteur financier. Finalement, comme les zones urbaines ont plus d’instituts financiers et d’investissement tels que les banques, les pays ayant les plus grandes populations urbaines sont plus impactés par une instabilité financière.

Nous notons que ces résultats sont équivalents aux résultats obtenus en faisant une ACP sur les variables (et cela avant leur transformation en logarithme). Les résultats de l’ACP sont affichés dans la Figure 14 à droite (les contributions des variables par rapport à l’axe 3).

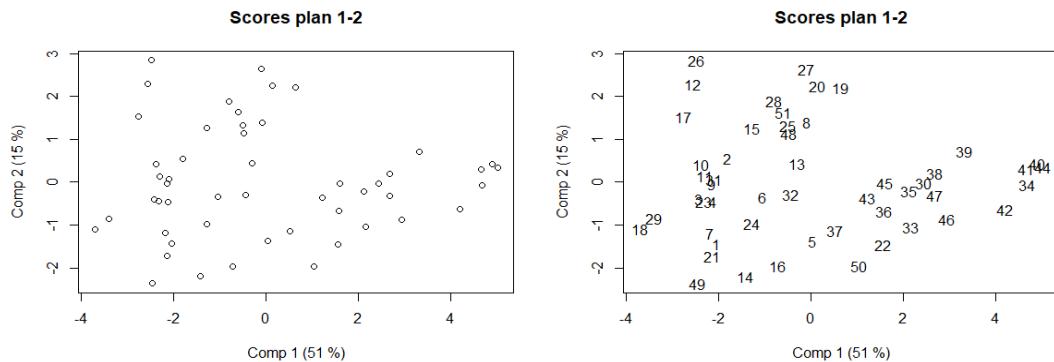
Figure 14: Loadings sur les 3 composantes (gauche), Contributions des variables lors de l'ACP (droite)

	Comp 1	Comp 2	Comp 3		Dim.1	Dim.2	Dim.3
LITERACY	-0.25	-0.09	-0.03	LITERACY	6.357406	0.4240928	0.32618146
GOOD_HEALTH	-0.38	-0.08	-0.19	GOOD_HEALTH	15.070906	0.4016966	2.15677976
URBAN_POP	-0.19	-0.49	0.43	URBAN_POP	3.847357	19.8299736	21.66873896
LIFE_EXP	-0.38	-0.13	-0.04	LIFE_EXP	15.024808	1.1554258	0.01621902
UNEMP	-0.19	0.53	0.47	UNEMP	3.095052	30.8391824	20.76098051
OLD_DEPEND	-0.35	0.19	0.19	OLD_DEPEND	12.373972	4.3702040	3.59491228
INFLATION	0.28	-0.23	0.27	INFLATION	8.375102	4.8652379	5.90551040
FINANC_DEV	-0.29	0.07	-0.63	FINANC_DEV	8.866579	0.3696709	37.39704287
log(GNI)	-0.39	-0.14	0.11	GNI	12.418663	6.3880525	1.11406138
YOUNG_DEPEND	0.36	-0.19	-0.19	YOUNG_DEPEND	12.711848	4.5975383	4.91445657
GDP	-0.13	-0.54	0.03	GDP	1.858307	26.7589251	2.14511677

Source : Sortie R

Nous cherchons maintenant à regarder la qualité de représentation des individus par rapport aux composantes principales. Nous utilisons les scores pour nous aider (Annexe 12).

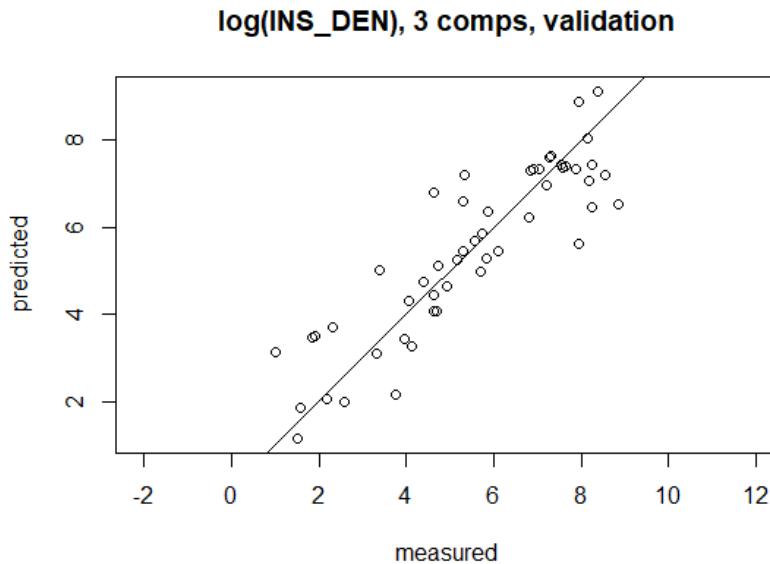
Figure 15 : plot des scores



Sur le graphique ci-dessus nous représentons les scores donnés en annexe 12 pour une meilleure visualisation. Il apparaît que les individus sont globalement bien représentés. Nous observons aussi une tendance de séparation en deux nuages mais que nous ne regarderons pas davantage ici.

Nous nous intéressons enfin à la qualité prédictive du modèle. Les prédictions se retrouvent en annexe 12 et nous en faisons une représentation ci-dessous pour les deux premiers coefficients.

Figure 16 : prédictions par PCR pour 3 composantes



Sur la base de cette figure, la qualité prédictive du modèle semble satisfaisante. Rappelons qu'avec les 3 composantes retenues, nous avons expliqué 81,13% de la variabilité de notre variable dépendante. La performance est donc bien meilleure par rapport au modèle final retenu pour la régression multiple où nous avons obtenu un R² ajusté de 82,49% mais avec 7 variables.

METHODE PARTIAL LEAST SQUARES (PLS)

Comme pour la régression sur composantes principales (PCR), la régression des moindres carrés partiels (PLS) ne se fait pas directement sur les variables explicatives de départ mais plutôt sur des variables latentes obtenues par réduction de dimension. En construisant des composantes orthogonales entre elles, elle permet de résoudre les problèmes éventuels de multi colinéarité des variables rencontrés dans la régression multiple standard (les moindres carrés ordinaires). De même, elle permet aussi de résoudre le problème de degrés de liberté dans le cas où on aurait moins d'individus que de variables explicatives.

Cependant, la PLS est différente de la PCR sur certains aspects. En effet, on distingue deux types de PLS : la PLS simple dite PLS1 lorsqu'il n'y a qu'une seule variable à expliquer et la PLS2 utilisée lorsqu'il y a au moins deux variables à expliquer. La PCR ne peut pas être appliquée pour plus d'une variable à expliquer. L'autre différence d'envergure entre la PLS et la PCR porte sur le critère d'ajustement. Contrairement à la PCR, la PLS prend également en compte

la variabilité de la variable à prédire. Ainsi, les composantes retenues sont celles qui maximisent la covariance entre les variables explicatives et les variables à prédire. Elles ne sont donc pas nécessairement les mêmes que celles obtenues par la PCR (et donc par l'ACP). Elles sont construites de manière à expliquer le mieux possible les variables « réponses ».

Dans notre cas, nous n'avons qu'une seule variable à prédire et nous appliquons donc la PLS simple. Les variables n'étant pas exprimées dans les mêmes unités, nous faisons une PLS sur les données réduites afin de ne pas surestimer l'importance de certaines variables lors de la construction des composantes.

➤ **Détermination du nombre de composantes à retenir**

Comme pour la PCR, nous procédons par validation croisée. Cependant, nous ne nous contentons pas uniquement de choisir le nombre de composantes qui minimise globalement le PRESS car ce choix ne paraît pas être le meilleur ici. En effet, la Figure 17 ci-dessous montre que le minimum est atteint à partir de 10 composantes avec un CV ajusté de 1.012. Mais prendre 10 composantes sur 11 composantes possibles semble très maladroit. D'autant plus que la variance de la variable réponse expliquée par ces 10 composantes (88% qui est la totalité de la variance pouvant être expliquée par notre modèle si nous considérons toutes les 11 composantes) n'est pas très différente de la variance expliquée par le modèle à 2 composantes par exemple (environ 83%).

Il apparaît donc qu'il nous faut un critère de sélection plus robuste et plus pertinent. Nous testons alors deux stratégies implémentées dans la fonction « selectNcomp » de la librairie pls de R. La première stratégie est basée sur le critère « one-sigma heuristic » qui consiste à choisir le modèle avec le moins de composantes possible dont l'erreur d'écart au meilleur modèle global est inférieure à un écart-type (d'où le nom de la méthode). La seconde stratégie repose sur une approche par permutation et teste (au risque 1%) si ajouter une nouvelle composante est bénéfique ou non. Partant du modèle réalisant le minimum global dans la validation croisée, elle utilise la méthode « backwards » pour réduire progressivement le

nombre de composantes tant qu'il n'y a pas de détérioration dans les performances du modèle.³⁰

Figure 17 : Résultat de la PLS

```
Data: X dimension: 51 11
Y dimension: 51 1
Fit method: kernelpls
Number of components considered: 11

VALIDATION: RMSEP
Cross-validated using 51 leave-one-out segments.
(Intercept) 1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps
CV          2.204   1.062   1.067   1.099   1.137   1.051   1.055   1.029
adjCV       2.204   1.061   1.066   1.098   1.133   1.048   1.052   1.026
               8 comps 9 comps 10 comps 11 comps
CV          1.034   1.027   1.015   1.015
adjCV       1.031   1.024   1.012   1.012

TRAINING: % variance explained
           1 comps 2 comps 3 comps 4 comps 5 comps 6 comps 7 comps 8 comps
X            51.38    60.84   72.55   76.90   80.07   84.44   87.88   91.12
log(INSP_DEN) 78.51    82.88   84.11   86.23   87.24   87.56   87.84   87.97
               9 comps 10 comps 11 comps
X            95.60    98.33   100.00
log(INSP_DEN) 87.99    87.99   87.99
```

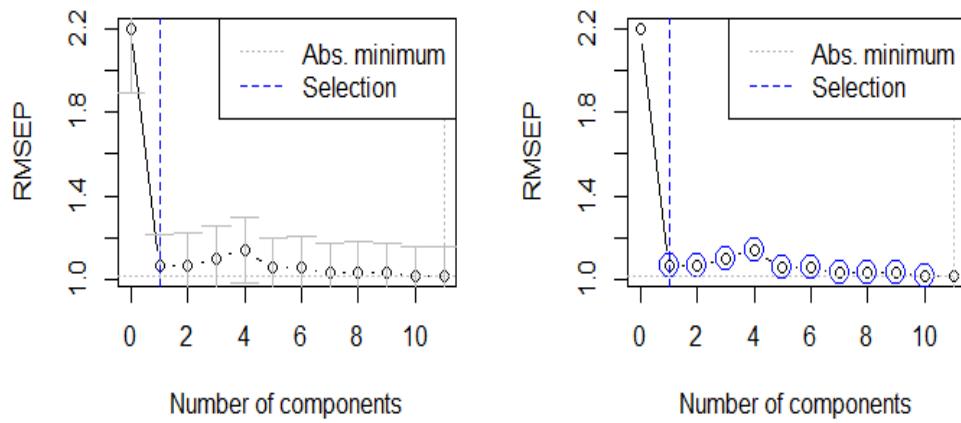
Source : Sortie R

Sur la Figure 18 ci-dessous, à gauche nous avons les résultats de la méthode « one-sigma » et à droite ceux de la deuxième stratégie. Les deux méthodes nous proposent de garder juste une composante qui explique à elle seule 51.38% de la variance des variables explicatives et 78.51% de la variance de la variable réponse (cf. Figure 17 ci-dessus). Par la suite, nous nous focaliserons donc sur cette composante mais nous pourrons regarder aussi certains résultats pour les deuxième et troisième composantes qui apportent une légère amélioration aux variances expliquées.

³⁰ Introduction to the pls Package.

<https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf>. 2020/04/01

Figure 18 : Méthodes one-sigma et permutation



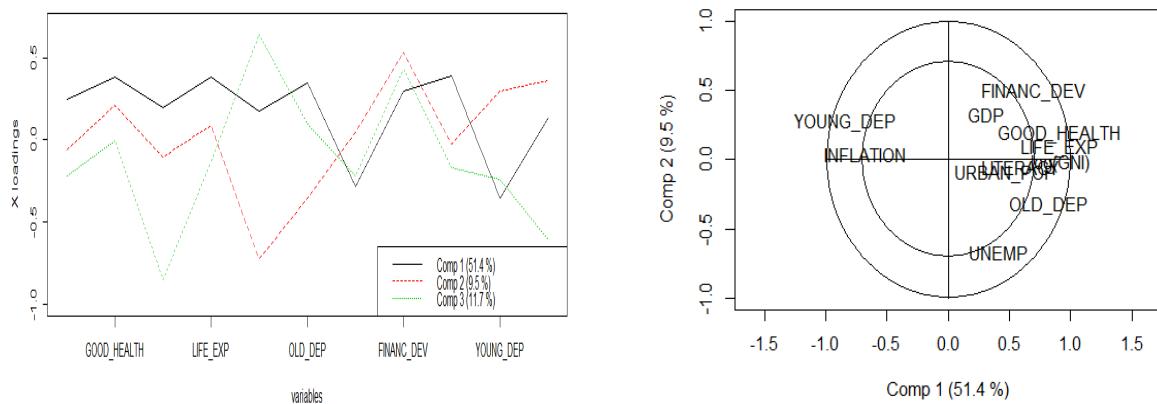
Source : Sortie R

➤ **Signification des axes : corrélation entre variables latentes et variables explicatives**

L'étude de la corrélation entre un axe et les variables explicatives permet de donner une signification à cet axe. Une manière de le faire est d'examiner les « loadings » des dites variables. Les loadings ne sont pas des corrélations mais reflètent la corrélation avec les axes. Ce sont les coordonnées des variables sur les axes. Elevés au carré, ils correspondent au cos² des variables sur les axes. Ainsi, ils indiquent la qualité de représentation des variables explicatives. Plus une variable est bien représentée sur un axe, plus elle est corrélée avec cet axe et plus elle a contribué à la constitution de l'axe. Ainsi, les loadings apportent une information équivalente à celle donnée par les poids des variables.

Nous considérons comme bien représentées sur un axe les variables dont les loadings sont supérieurs à 0.25 en valeur absolue.

Figure 19 : loadings et cercle de corrélation dans le plan 1-2



Source : sortie R

Sur la Figure 19 ci-dessus et les graphiques de l'annexe 13, nous remarquons ce qui suit :

- L'axe 1 sature les variables GOOD_HEALTH (0.39), log(GNI) (0.39), LIFE_EXP (0.38), OLD_DEP (0.35), FINANC_DEV (0.30), LITERACY (0.25) qui lui sont corrélées positivement et les variables YOUNG_DEP (-0.35), INFLATION (-0.28) qui lui sont corrélées négativement.
- L'axe 2 sature positivement les variables FINANC_DEV, GDP et YOUNG_DEP et négativement les variables UNEMP et OLD_DEP.
- L'axe 3 sature positivement les variables UNEMP et FINANC_DEV et négativement les variables URBAN_POP et et GDP.

L'axe 1 représente bien la plupart des variables (8 sur 11), ce qui conforte encore le choix d'une seule composante pertinente pour la méthode. Avec autant de variables, il serait difficile de donner une interprétation précise à l'axe mais une tendance s'observe ici : l'axe 1 présente le contraste entre les habitants des pays riches bénéficiant d'une bonne formation, qui sont généralement en bonne santé et vivent plus longtemps et les habitants des pays pauvres globalement jeunes avec un pouvoir d'achat moins élevé. Les premiers seront plus susceptibles d'investir sur le long terme engendrant ainsi un développement du marché financier et de l'assurance à l'inverse des seconds.

L'axe 2 pourrait représenter le **développement du marché du travail** avec une opposition entre les pays ayant une main d'œuvre jeune favorisant une production conséquente des richesses à l'intérieur du pays avec des répercussions positives sur le secteur financier et les pays confrontés aux problèmes de chômage et d'importants départs en retraite.

L'axe 3 représenterait le **développement urbain** avec une opposition entre les pays avec des villes dynamiques et productives et les pays avec des villes caractérisées par des problèmes d'emploi à l'exception du secteur financier.

➤ **Comprendre la contribution des axes à la réponse : corrélation entre variables latentes et variable réponse**

L'étude des corrélations entre les réponses et un axe permet de comprendre, cerner ce qui y est expliqué. Dans notre cas, il n'y a pas de suspens possible car nous avons une seule variable réponse.

Figure 20 : Y loadings

Comp 1	Comp 2	Comp 3
0.82	0.49	0.29

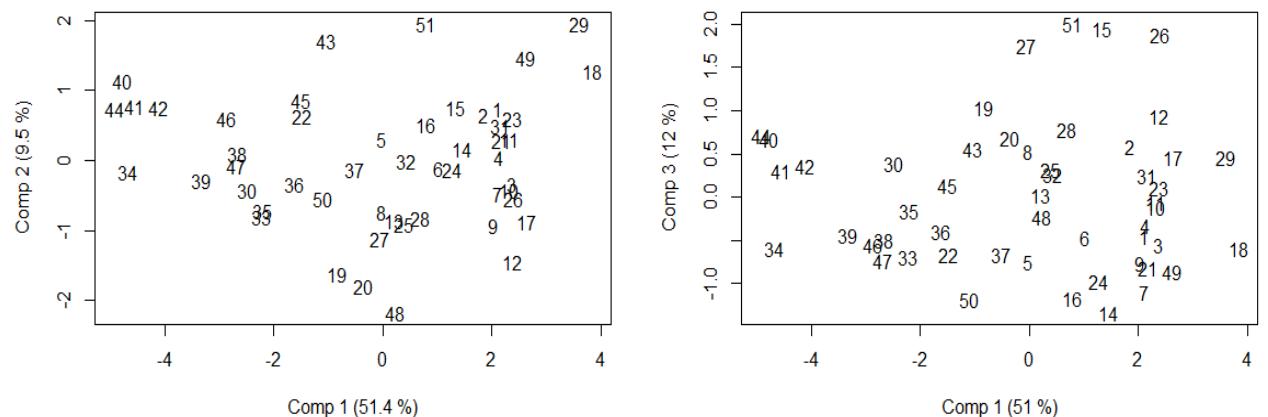
Sur la figure ci-dessus, nous constatons que notre variable réponse est fortement corrélée avec la première composante. Du fait de la signification de l'axe 1, nous pouvons en déduire que plus un pays est riche avec une population bien instruite vivant assez longtemps, plus la demande de produits d'assurance sera importante.

La grande partie de la variance de la variable réponse étant expliquée par la première composante (celle que nous avons également retenue lors de la sélection), il ne nous semble pas utile de nous attarder sur les liaisons avec les autres composantes.

➤ Scores

Les scores permettent d'évaluer la qualité de représentation des individus.

Figure 21 : Scores dans les plans 1-2 et 1-3



Sur cette Figure 21, nous remarquons qu'il n'y a pas de regroupement particulier d'individus suivant les axes considérés. Ceci soutient davantage la place prépondérante des variables dans notre étude et non des individus particuliers.

➤ Coefficients de la régression

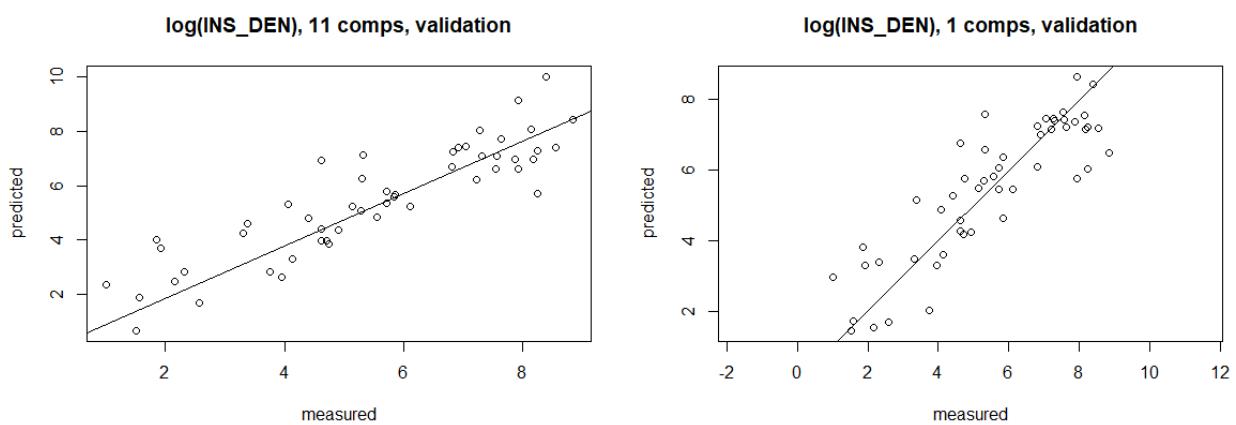
Figure 22 : Coefficients de régression

	, , 1 comps	, , 2 comps	, , 3 comps
	$\log(\text{INS_DEN})$	$\log(\text{INS_DEN})$	$\log(\text{INS_DEN})$
LITERACY	0.1926419	LITERACY	0.132676608
GOOD_HEALTH	0.3284373	GOOD_HEALTH	0.435124986
URBAN_POP	0.1376777	URBAN_POP	0.001168004
LIFE_EXP	0.3092747	LIFE_EXP	0.299684382
UNEMP	0.1131936	UNEMP	-0.079662422
OLD_DEP	0.2515181	OLD_DEP	0.062638857
INFLATION	-0.2251487	INFLATION	-0.200247520
FINANC_DEV	0.2872601	FINANC_DEV	0.606864940
$\log(\text{GNI})$	0.3298830	$\log(\text{GNI})$	0.424730419
YOUNG_DEP	-0.2697750	YOUNG_DEP	-0.160878903
GDP	0.1276697	GDP	0.257607806

Les coefficients de régression donnent l'impact de chaque variable sur la réponse. Sur la figure ci-dessus, nous constatons qu'en gardant la première composante uniquement, les variables GOOD_HEALTH, LIFE_EXP FINANC_DEV et $\log(\text{GNI})$ sont celles qui impactent le plus la réponse positivement. Si on jette un coup d'œil à ce qui se passerait si on garde trois composantes par exemple, on observe l'importance grandissante des variables FINANC_DEV, $\log(\text{GNI})$ et GOOD_HEALTH. Signalons aussi l'impact négatif des variables INFLATION et YOUNG_DEP comme nous l'avions supposé au début de ce travail.

➤ Prédiction

Figure 23 : graphiques de prédictions (11 comp vs 1 comp)



Sur la figure ci-dessus des valeurs prédictes (en gardant les deux premiers coefficients), on remarque que la prédiction basée sur une seule composante n'est pas si moins bonne que celle utilisant toutes les composantes. Et elle semble meilleure que celle de la PCR en termes de dispersion. En outre, avec deux composantes, la PLS explique une variabilité de la variable dépendante légèrement supérieure à celle expliquée par 3 composantes pour la PCR comme vu précédemment.

VI- CONCLUSION

L'objectif de ce travail était d'étudier les facteurs affectant la demande de l'assurance vie. Pour ce faire, nous avons retenu la densité de l'assurance comme outil de mesure de la demande en assurance vie pour chaque pays. Nous avons ensuite considéré onze indicateurs comme variables explicatives de cette densité en nous basant sur des études déjà effectuées à ce sujet : le taux d'alphabétisation de la population, le pourcentage de la population en bonne santé, la population urbaine, l'espérance de vie, le taux de chômage, le ratio des inactifs par rapport aux actifs, le taux d'inflation, le développement du secteur financier, le PIB par habitant, le revenu national brut et le ratio des jeunes inactifs.

Dans un premier temps, nous avons mis en évidence l'existence des corrélations entre ces différentes variables au travers de l'analyse descriptive des données et l'analyse exploratoire portant essentiellement sur la réduction de dimension par l'ACP. Cette analyse s'est achevée par une classification hiérarchique ascendante qui nous a permis d'établir quelques regroupements possibles des pays en rapport avec notre étude.

Dans un second temps, nous avons fait l'analyse économétrique des données en utilisant trois méthodes de régression différentes : la régression multiple, la régression sur composantes principales (PCR) et la régression des moindres carrés partiels (PLS). Afin de garantir la linéarité de notre modèle et de satisfaire les hypothèses du modèle gaussien, les différents tests statistiques nous ont conduit à faire une transformation logarithmique de deux variables : la densité d'assurance (notre variable dépendante) et le revenu national brut.

Pour ce qui est de ces trois méthodes, elles ont été complémentaires mais avec des bien meilleurs résultats pour les deux méthodes de régression sur variables latentes. Plus précisément, le problème de multicolinéralité entre certaines variables nous a conduit à retenir finalement sept variables pour faire la régression multiple. Quant à la PCR, en retenant trois composantes uniquement, nous avons été capables d'expliquer une plus grande variabilité de notre variable réponse. La PLS faisait encore mieux avec deux composantes uniquement mais nous nous sommes limités à une seule composante qui était suffisante pour notre étude.

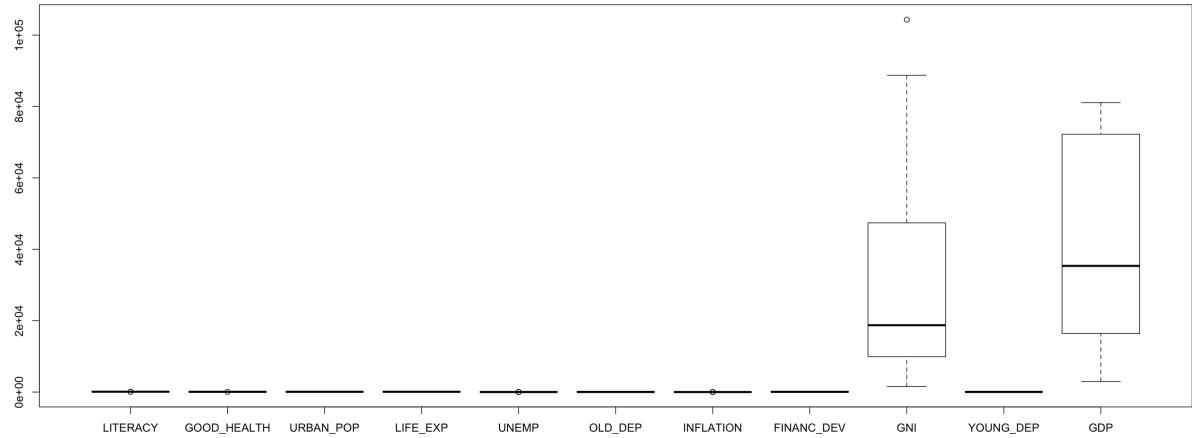
Pour finir, signalons que le fait d'avoir des variables qui ne sont pas exprimées dans les mêmes unités n'a pas permis de comparer les coefficients de régression obtenus par ces trois méthodes. En effet, la régression multiple a été effectuée sur les variables non réduites contrairement aux régressions PCR et PLS qui nécessitent une réduction préalable des variables de départ avant la construction des variables latentes dans ce cas de figure. Cependant, toutes ces méthodes nous ont permis d'identifier les variables les plus déterminantes dans l'explication de la densité de l'assurance vie comme le PIB ou le développement du secteur financier par exemple, confirmant ainsi certaines de nos hypothèses émises dans l'introduction de ce travail.

BIBLIOGRAPHIE

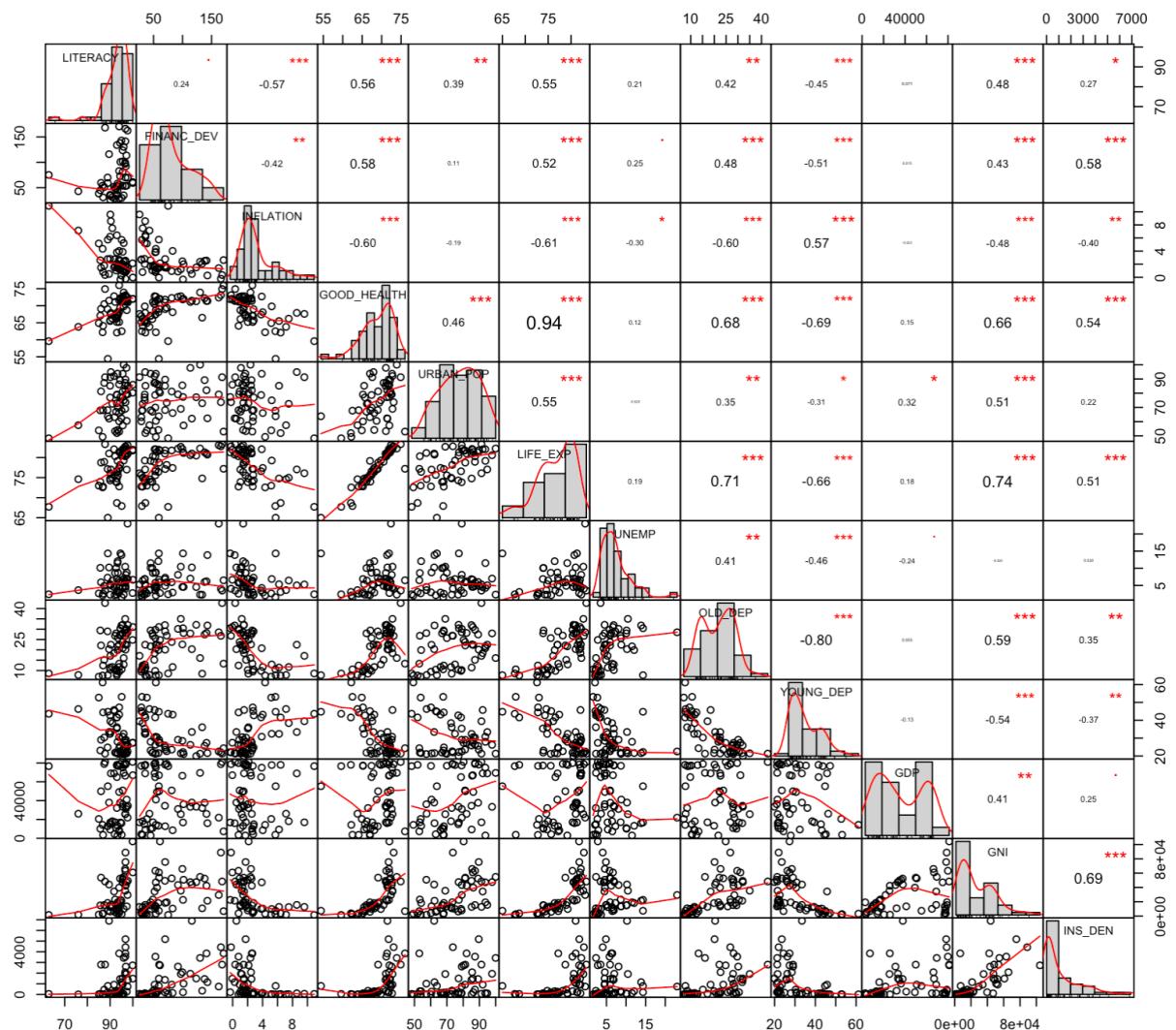
- AHLGRIM, K., & D'ARCY, S. (2012). The effect of Deflation or High Inflation on the Insurance Industry. *Casualty Actuarial Society*.
- Bonnart, J. (2012). *Les conséquences des crises financières de 2008/2009 et 2011/2012 sur l'assurance.* Récupéré sur <https://halshs.archives-ouvertes.fr/halshs-00655657/document>
- BROWN , & GUY C. (s.d.). Living too Long . *EMBO Reports*, 16(2). Récupéré sur <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4328740/pdf/embr0016-0137.pdf>
- COLLETT, R., ABKEMEIER, N., BONACH, E., & PAPASAVVAS, D. (1990). Evolution of Life Insurance Industry Throughout the World. *Record of Society of Actuaries*.
- DANIEL, C. (s.d.). *Régressions sur variables latentes : Principal Component Regression et Partial Least Square Fichier*. Université d'Angers .
- Dieng, M. S., & Fall, M. (2012). *Les déterminants de la consommation d'assurance-vie: le cas de l'UEMOA.* Sénégal.
Documentation R . (s.d.).
- Group, W. (s.d.). *WorldBank: Data*. Récupéré sur <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS>
- Healthy life years statistics*. (s.d.). Consulté le 02 10, 2020, sur Eurostat statistic explained: https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Healthy_life_years_statistics/fr
- INSEE. (s.d.). *Metadonnées - Définitions INSEE*. Consulté le 02 10, 2020, sur <https://www.insee.fr/fr/metadonnees/definition/c1374>
- MEVIK, B.-H., & WEHRENS, R. (s.d.). *Introduction to the PLS package* . Consulté le 04 01, 2020, sur <https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf>
- National Association of Insurance Commissioners. Longevity Risk*. (s.d.). Consulté le 02 10, 2020, sur https://www.naic.org/cipr_topics/topic_longevity_risk.htm
- OCDE Statistics*. (s.d.). Récupéré sur <https://stats.oecd.org>
- retraites, S. g. (2016). *Le point sur les fonds de pensions - synthèse des travaux de l'OCDE* . Récupéré sur Vieillissement, emploi et retraite: panorama international .
- Roser, M., & Ortiz-Ospina, E. (2013). *Our World in Data*. Récupéré sur <https://ourworldindata.org/literacy>
- sciences, F. (s.d.). *Définition PIB*. Consulté le 02 02, 2020, sur Futura Sciences: <https://www.futura-sciences.com/planete/definitions/developpement-durable-pib-6295/>
- THORSTEN, B., & WEBB, I. (2002). Economic, Demographic, and Institutional Determinants of Life Insurance Consumption across Countries. *World Bank and International Insurance Foundation*.
- TIENYU, H. (2003). The determinants of the demand for life insurance in an emerging economy - the case of China. 29, 82-96.
- UNESCO. (2019). *The World Bank Group*. Récupéré sur Data world bank: <https://data.worldbank.org/indicator/SE.ADT.LITR.ZS>

ANNEXE

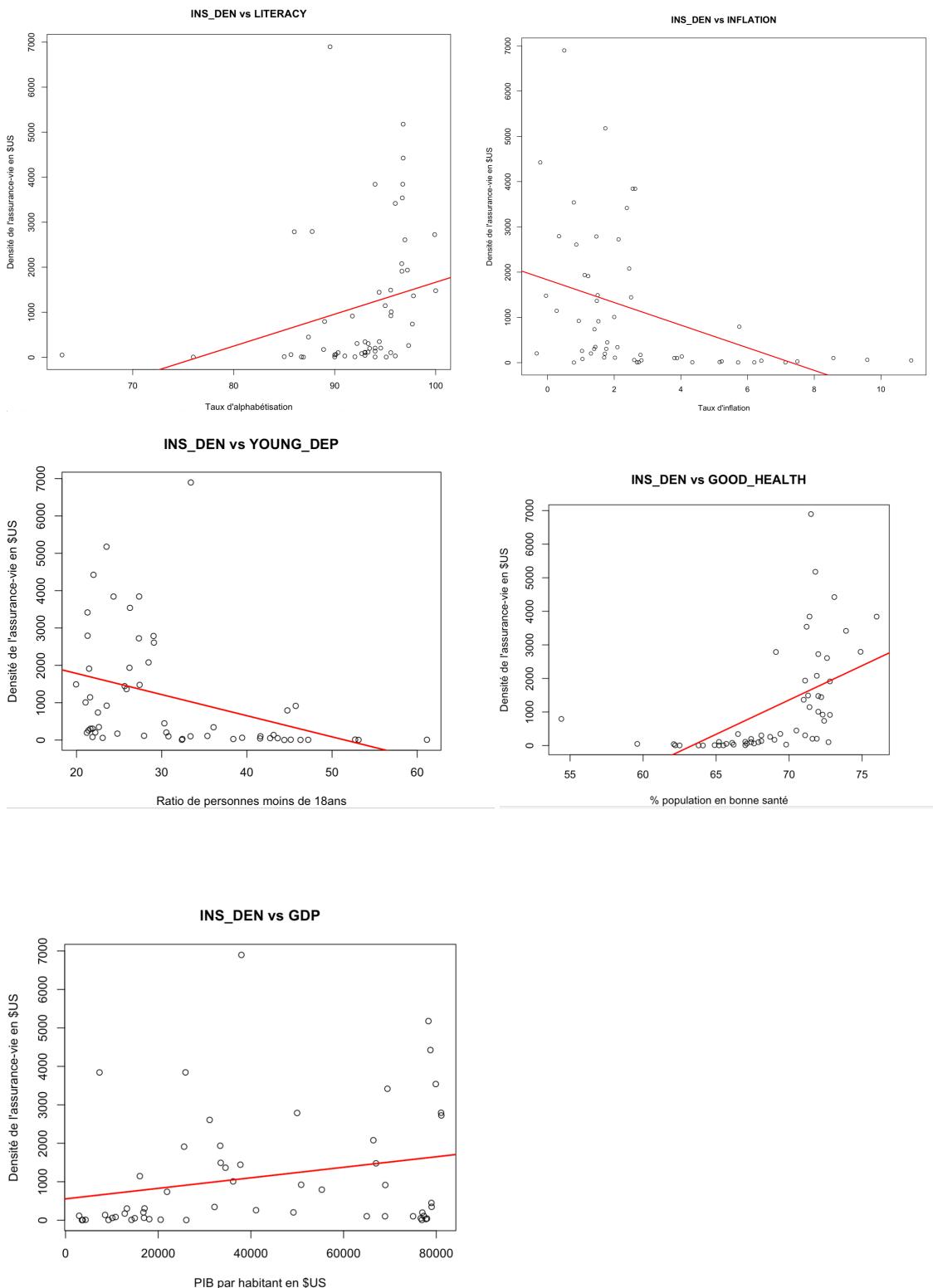
1. Boites à moustache de toutes les variables explicatives :



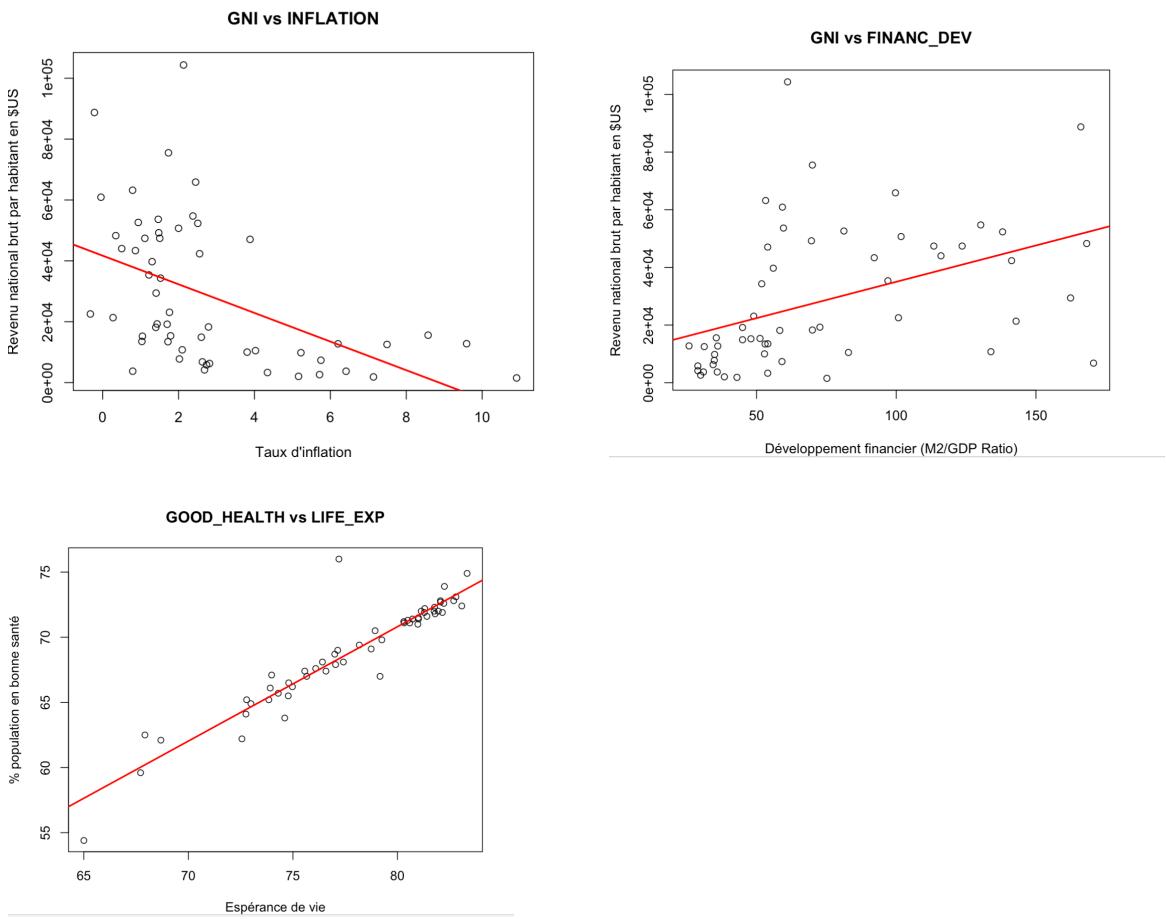
2. Test de Spearman : Tableau de corrélation



3. Nuage de points (Relations entre la variable expliquée et certaines variables explicatives) :



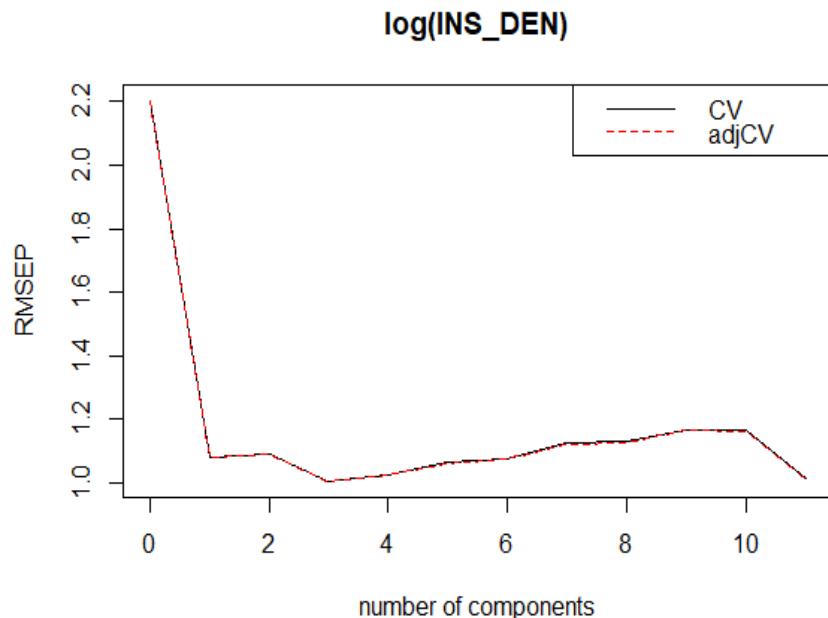
4. Nuage de points (Relations entre certaines variables explicatives) :



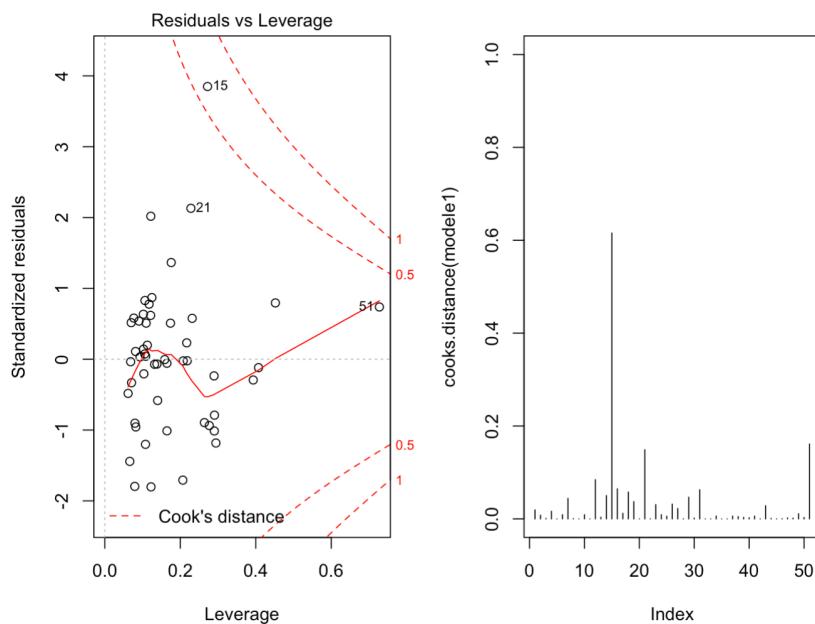
5. Tableau des valeurs propres

	eigenvalue	percentage of variance	cumulative percentage of variance	percentage of variance
comp 1	5.16	51.64	51.64	
comp 2	1.74	17.42	69.06	
comp 3	0.90	8.99	78.05	
comp 4	0.69	6.92	84.97	
comp 5	0.45	4.54	89.50	
comp 6	0.33	3.30	92.81	
comp 7	0.30	2.98	95.79	
comp 8	0.19	1.95	97.74	
comp 9	0.18	1.78	99.52	
comp 10	0.05	0.48	100.00	

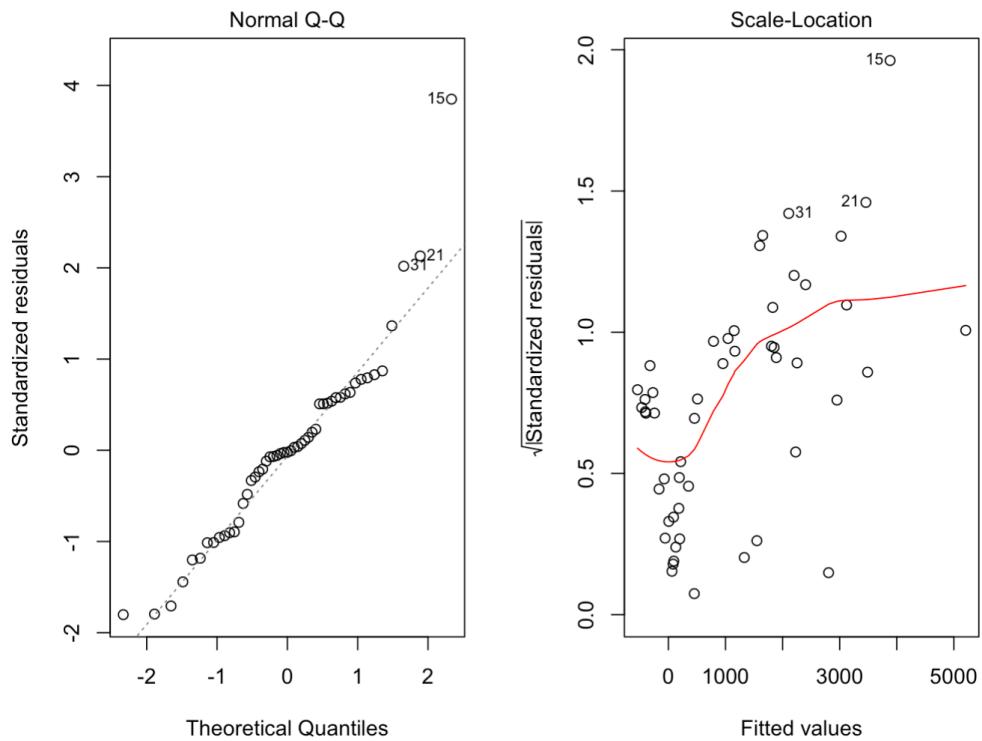
6. Graphique affichant le nombre de composantes principales. On atteint un minimum est atteint à 3 composantes.



7. Distance de Cooks



8. Analyse des résidus



9. Résultat du modèle en enlevant LIFE_EXP

RESET test

```
data: modele1a
RESET = 7.0755, df1 = 2, df2 = 41, p-value = 0.002292

> vif(modele1a)
    UNEMP        GNI   FINANC_DEV      OLD_DEP     YOUNG_DEP    LITERACY GOOD_HEALTH
    2.157910    2.807312    2.016322    3.338679    3.923592    1.415041    3.633585
> bptest(modele1a)
```

studentized Breusch-Pagan test

```
data: modele1a
BP = 19.716, df = 7, p-value = 0.006217
```

```
> residus<-residuals(modele1a)
> shapiro.test(residus)
```

Shapiro-Wilk normality test

```
data: residus
W = 0.94259, p-value = 0.01562
```

10. Résultat du modèle cette fois en utilisant le log(INS_DEN)

```

RESET test

data: modele2
RESET = 5.0016, df1 = 2, df2 = 41, p-value = 0.01138

> bptest(modele2)

studentized Breusch-Pagan test

data: modele2
BP = 8.189, df = 7, p-value = 0.3162

> residus<-residuals(modele2)
> shapiro.test(residus)

Shapiro-Wilk normality test

data: residus
W = 0.97346, p-value = 0.3063

```

11. Scores de PCR

	Comp 1	Comp 2	Comp 3
1	-2.03912313	-1.43937455	-0.26483745
2	-1.79020750	0.55124403	-1.02439051
3	-2.42306037	-0.39360329	0.25741613
4	-2.12333567	-0.45584325	-0.03334253
5	0.05195598	-1.37324423	0.33539096
6	-1.04040580	-0.34796267	0.12410000
7	-2.16737247	-1.19436991	0.73913022
8	-0.06908237	1.39461331	0.37516402
9	-2.14004494	-0.03584303	0.64480143
10	-2.36624453	0.41730901	0.01912918
11	-2.29617609	0.13388020	-0.65715798
12	-2.54893398	2.29373487	0.74674780
13	-0.30431875	0.43211038	0.87590742
14	-1.41523741	-2.20448610	0.66593608
15	-1.27820054	1.26209076	-0.74963508
16	-0.71238474	-1.96660338	0.05181382
17	-2.75792003	1.53011686	0.23317954
18	-3.70128301	-1.09924526	-1.10541842
19	0.63954588	2.21141725	1.33721893
20	0.14751529	2.24940831	1.06137084
21	-2.13797955	-1.72240390	0.34041662
22	1.57187388	-1.45451384	-0.14802968
23	-2.31855569	-0.44822079	-0.40126676
24	-1.27637984	-0.97254847	0.21474838
25	-0.48930322	1.32062608	0.26895546
26	-2.48503973	2.84459880	-0.29976016
27	-0.09353273	2.64466006	0.55530581
28	-0.79927041	1.89096037	0.17762671
29	-3.39072643	-0.85387242	-1.86871794
30	2.44751530	-0.02886870	1.04527030
31	-2.10243455	0.06834294	-0.65541685
32	-0.43222046	-0.29539585	0.50901206
33	2.15656314	-1.05144568	1.80855747
34	4.68590347	-0.07291187	-0.26818682
35	2.13118894	-0.21354568	1.19402249
36	1.59074380	-0.67783562	0.68697353
37	0.51857060	-1.13565926	0.74992029
38	2.68297011	0.20370563	-0.74148248
39	3.31248580	0.71150007	-0.54767738
40	4.90807820	0.42227823	-1.44170938
41	4.65838272	0.30166063	-0.91401865
42	4.20299941	-0.63443543	-0.24247334
43	1.22083539	-0.36448970	-1.84056110
44	5.01559669	0.34618079	-0.23837043
45	1.60649122	-0.03199663	-1.16102400
46	2.94282322	-0.87353846	-0.61775723
47	2.68748624	-0.31822133	-0.19330217
48	-0.47475083	1.13643185	1.90208290
49	-2.45345029	-2.36982544	-0.75883434
50	1.05018106	-1.96337342	1.88816609
51	-0.60273129	1.62680772	-2.63499382

12. Prédictions de PCR pour 3 composantes retenues

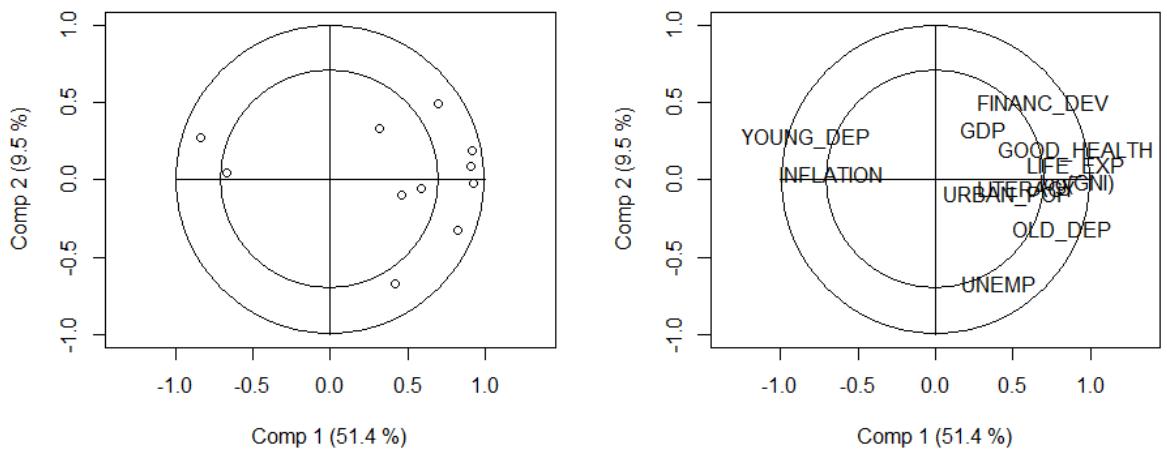
	, , 3 comps
	log(INS_DEN)
1	7.425817
2	7.336066
3	7.377491
4	7.274770
5	5.467194
6	6.320943
7	7.048287
8	5.224371
9	6.933940
10	7.344732
11	7.625942
12	6.945490
13	5.299991
14	6.594324
15	6.718087
16	6.278275
17	7.433606
18	9.100649
19	4.126539
20	4.641786
21	7.265670
22	4.472751
23	7.596461
24	6.542486
25	5.618549
26	7.301256
27	5.017392
28	5.842411
29	9.166199
30	3.065684
31	7.477073
32	5.652687
33	3.074854
34	1.863224
35	3.274436
36	3.991173
37	4.877848
38	3.653946
39	3.001612
40	2.155452
41	2.132559
42	2.305129
43	5.390150
44	1.536032
45	4.735491
46	3.514781
47	3.464226
48	4.892897
49	8.089577
50	4.034381
51	6.981721

13. Loadings et cercles de corrélations

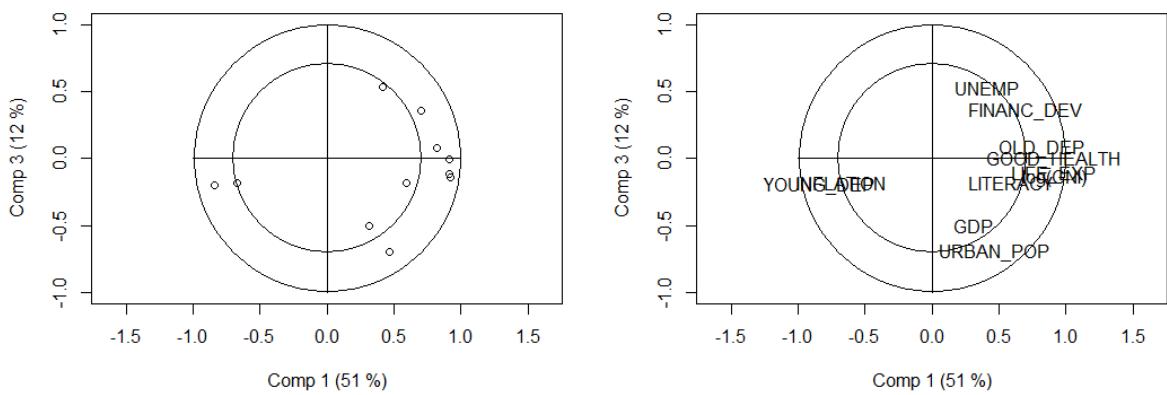
Loadings des variables explicatives pour les trois premières composantes :

	Comp 1	Comp 2	Comp 3
LITERACY	0.25	-0.06	-0.22
GOOD_HEALTH	0.39	0.21	-0.01
URBAN_POP	0.20	-0.10	-0.85
LIFE_EXP	0.38	0.09	-0.14
UNEMP	0.18	-0.73	0.64
OLD_DEPEND	0.35	-0.36	0.10
INFLATION	-0.28	0.05	-0.22
FINANC_DEV	0.30	0.54	0.43
Log(GNI)	0.39	-0.03	-0.17
YOUNG_DEPEND	-0.35	0.30	-0.24
GDP	0.13	0.36	-0.60

Cercle de corrélation dans le plan 1-2 :



Cercle de corrélation dans le plan 1-3 :



14. Classification ascendante hiérarchique avec 4 clusters

15. Classification ascendante hiérarchique avec 3 clusters

		pays	cluster		pays	cluster
1		Australia	3	1	Australia	4
2		Austria	3	2	Austria	4
3		Belgium	3	3	Belgium	4
4		Canada	3	4	Canada	4
5		Chile	3	5	Chile	2
6	Czech Republic	Czech Republic	3	6	Czech Republic	4
7		Denmark	3	7	Denmark	4
8		Estonia	2	8	Estonia	3
9		Finland	3	9	Finland	4
10		France	3	10	France	4
11		Germany	3	11	Germany	4
12		Greece	2	12	Greece	3
13		Hungary	2	13	Hungary	3
14		Iceland	3	14	Iceland	4
15		Ireland	3	15	Ireland	4
16		Israel	3	16	Israel	4
17		Italy	2	17	Italy	3
18		Japan	3	18	Japan	4
19		Latvia	2	19	Latvia	3
20		Lithuania	2	20	Lithuania	3
21		Luxembourg	3	21	Luxembourg	4
22		Mexico	1	22	Mexico	2
23		Netherlands	3	23	Netherlands	4
24		New Zealand	3	24	New Zealand	4
25		Poland	2	25	Poland	3
26		Portugal	2	26	Portugal	3
27	Slovak Republic	Slovak Republic	2	27	Slovak Republic	3
28		Slovenia	2	28	Slovenia	3
29		Switzerland	3	29	Switzerland	4
30		Turkey	1	30	Turkey	1
31	United Kingdom	United Kingdom	3	31	United Kingdom	4
32	United States	United States	3	32	United States	4
33		Argentina	1	33	Argentina	2
34		Bolivia	1	34	Bolivia	1
35		Brazil	1	35	Brazil	2
36		Colombia	1	36	Colombia	2
37		Costa Rica	1	37	Costa Rica	2
38		Ecuador	1	38	Ecuador	1
39	El Salvador	El Salvador	1	39	El Salvador	1
40		Guatemala	1	40	Guatemala	1
41		Honduras	1	41	Honduras	1
42		Indonesia	1	42	Indonesia	1
43		Malaysia	1	43	Malaysia	1
44		Nicaragua	1	44	Nicaragua	1
45		Panama	1	45	Panama	1
46		Paraguay	1	46	Paraguay	1
47		Peru	1	47	Peru	1
48	Puerto Rico	Puerto Rico	2	48	Puerto Rico	3
49		Singapore	3	49	Singapore	4
50		Uruguay	1	50	Uruguay	2
51		China	3	51	China	4

16. Résultats de la fonction step, donnant le modèle minimisant l'AIC

```

Step: AIC=703.93
baseFinal$INS_DEN ~ baseFinal$LITERACY + baseFinal$GOOD_HEALTH +
  baseFinal$LIFE_EXP + baseFinal$UNEMP + baseFinal$OLD_DEP +
  baseFinal$FINANC_DEV + baseFinal$GNI + baseFinal$YOUNG_DEP

      Df Sum of Sq    RSS    AIC
<none>            35369865 703.93
- baseFinal$LITERACY   1   2229238 37599103 705.04
- baseFinal$YOUNG_DEP   1   2715693 38085558 705.70
- baseFinal$OLD_DEP    1   3436424 38806290 706.66
- baseFinal$GOOD_HEALTH 1   6195775 41565640 710.16
- baseFinal$LIFE_EXP    1   6647866 42017731 710.71
- baseFinal$FINANC_DEV   1   7044494 42414359 711.19
- baseFinal$UNEMP       1   8022431 43392296 712.35
- baseFinal$GNI          1   31774934 67144799 734.62

Call:
lm(formula = baseFinal$INS_DEN ~ baseFinal$LITERACY + baseFinal$GOOD_HEALTH +
  baseFinal$LIFE_EXP + baseFinal$UNEMP + baseFinal$OLD_DEP +
  baseFinal$FINANC_DEV + baseFinal$GNI + baseFinal$YOUNG_DEP)

Coefficients:
(Intercept)  baseFinal$LITERACY  baseFinal$GOOD_HEALTH  baseFinal$LIFE_EXP
3673.79842      -60.55560        327.71353      -308.90463
baseFinal$UNEMP  baseFinal$OLD_DEP  baseFinal$FINANC_DEV  baseFinal$GNI
199.52152        -54.61579        13.17854        0.07277
baseFinal$YOUNG_DEP
46.56571

```