Étude de l'espérance de vie dans le monde par pays en 2022

Thierno BAH et Elliot RAULT-MAISONNEUVE

2025-02-12

Introduction :

L'espérance de vie constitue l'un des indicateurs les plus représentatifs du niveau de développement d'un pays. Il reflète de manière synthétique la qualité des systèmes de santé, l'accès à l'éducation, les conditions économiques, les infrastructures de base, ainsi que la stabilité institutionnelle. Pourtant, malgré les avancées technologiques et économiques observées au cours des dernières décennies, de fortes disparités subsistent entre les pays, et l'espérance de vie connaît même un recul dans certaines régions du monde. Ces constats soulèvent de nombreuses interrogations quant aux mécanismes réels qui influencent la longévité des populations.

Dans ce contexte, ce projet a pour objectif d'analyser les principaux facteurs susceptibles d'influencer l'espérance de vie à l'échelle mondiale, en s'appuyant sur des données récentes de l'année 2022. L'ambition de ce travail ne se limite pas à l'identification de corrélations entre les variables explicatives et l'espérance de vie. Il s'agit également d'explorer les limites des approches économétriques traditionnelles face à des problématiques complexes telles que l'endogénéité, la multicolinéarité, ou encore la sélection de variables dans des contextes de haute dimension.

Ce projet s'inscrit ainsi dans une démarche d'apprentissage et de mise en pratique des outils d'économétrie avancée, incluant les méthodes d'estimation par variables instrumentales, la technique des doubles moindres carrés, les méthodes de régularisation (Ridge, Lasso, Elastic Net), l'analyse en composantes principales (PCA), ainsi que les approches modernes comme le double machine learning.

Mais pour obtenir des résultats solides, il est nécessaire de dépasser les simples corrélations. Certains facteurs peuvent être mal mesurés, dépendre d'autres variables du modèle, ou être très corrélés entre eux, ce qui rend les estimations classiques peu fiables. C'est pourquoi nous avons structuré notre démarche autour de plusieurs étapes clés :

1 - Estimation d'un modèle de base par régression linéaire (MCO) pour repérer les relations principales entre l'espérance de vie et les variables explicatives.

2 - Analyse de l'endogénéité, pour détecter les biais potentiels liés à certaines variables, et recours, si nécessaire, à des méthodes comme les moindres carrés en deux étapes (2SLS).

3 - Traitement de la multicolinéarité, en utilisant des techniques modernes comme la PCA, le Ridge, le Lasso ou l'Elastic Net, afin d'améliorer la stabilité des estimations.

4 - Enfin, application du double machine learning, afin d'estimer de manière robuste l'effet causal de certains facteurs dans un environnement à haute dimension.

Toutes ces étapes ont pour but d'obtenir les estimations les plus fiables possibles, et ainsi de mieux comprendre comment les différents facteurs influencent l'espérance de vie dans le monde.

Présentation des variables :

Code	Nom de la variable	Définition	Unité	Source
SN.ITK.DEFC.ZS	Mal_nutrition	Prévalence de la malnutrition	% de la population	Banque Mondiale
EG.ELC.ACCS.ZS	Access_electricity	Accès à l'électricité	% de la population	Banque Mondiale
CC.EST	Control_corruption	Contrôle de la corruption	Indice (estimate)	Banque Mondiale
SE.XPD.TOTL.GD.ZS	Dep_education	Dépenses publiques en éducation	% du PIB	Banque Mondiale
GE.EST	Government._Effectiveness	Efficacité du gouvernement	Indice (estimate)	Banque Mondiale
SP.DYN.LE00.IN	Life_expectancy	Espérance de vie à la naissance	Total (années)	Banque Mondiale
SP.DYN.IMRT.IN	Mortality_infant	Taux de mortalité infantile	Pour 1 000 naissances vivantes	Banque Mondiale
SH.H2O.BASW.ZS	Access_water	Accès à l'eau potable de base	% de la population	Banque Mondiale
NY.GDP.PCAP.KD	GDP_per_habitant	PIB par habitant	$ US constants 2015	Banque Mondiale
RL.EST	Rule_of_Law	État de droit	Indice (estimate)	Banque Mondiale
PV.EST	Political_Stability	Stabilité politique	Indice (estimate)	Banque Mondiale

Explications du choix des variables explicatives et l'effet attendu :

Pour ce projet, nous avons choisi les variables que nous estimons les plus pertinentes pour expliquer les différences d'espérance de vie entre les pays. Ces variables couvrent plusieurs aspects importants : la santé, l'économie, l'éducation, les institutions et les conditions de vie de base. Chacune de ces dimensions peut avoir un impact direct ou indirect sur la durée de vie des populations.

Variables sanitaires :

La malnutrition reflète un manque d'accès à une alimentation suffisante et équilibrée, ce qui peut affaiblir la santé et augmenter les risques de maladies. La mortalité infantile est un indicateur de la qualité du système de santé, surtout pour les enfants. Nous nous attendons donc à ce que ces deux variables aient un effet négatif sur l'espérance de vie.

Variable économique :

Le PIB par habitant donne une idée du niveau de richesse dans un pays. Plus ce revenu est élevé, plus les populations ont accès à de meilleurs services, comme la santé ou l'alimentation. Nous nous attendons donc à un effet positif.

Variable éducative :

Les dépenses publiques en éducation montrent l'importance accordée à l'éducation dans le pays. L'éducation permet souvent de mieux comprendre les enjeux de santé et d'adopter de meilleurs comportements. Nous nous attendons donc à un effet positif.

Variables institutionnelles :

Le contrôle de la corruption, l'efficacité du gouvernement, l'état de droit et la stabilité politique représentent la qualité des institutions dans un pays. Des institutions solides permettent de mieux gérer les ressources et d'assurer un bon accès aux services de base. Ces variables devraient avoir un effet positif sur l'espérance de vie.

Variables environnementales et d'infrastructures de base :

L'accès à l'eau potable et à l'électricité est fondamental pour vivre dans de bonnes conditions. Cela permet de prévenir certaines maladies et d'améliorer le confort de vie. Ces deux variables sont attendues avec un effet positif.

Préparation et nettoyage des données :

Avant de passer à l'analyse, nous avons d'abord nettoyé les données pour pouvoir travailler sur une base cohérente et exploitable. Nous avons commencé par supprimer les variables qui avaient plus de 60% de

valeurs manquantes, car elles étaient trop incomplètes pour être utilisées. Par exemple, cela a été le cas de la variable sur les dépenses en santé et celle sur l'indice de Gini, que nous avons dû retirer de notre base.

Ensuite, nous avons aussi supprimé les pays qui avaient plus de 50% de données manquantes, pour éviter d'avoir trop de valeurs reconstruites artificiellement.

Pour les données manquantes restantes, nous avons utilisé une méthode d'imputation avec les 5 plus proches voisins (KNN). L'idée est simple : pour chaque valeur manquante, nous avons cherché les 5 pays les plus proches (en se basant sur les autres variables disponibles), puis nous avons pris la moyenne de leurs valeurs pour compléter.

Grâce à ces étapes, nous avons obtenu une base de données plus propre, plus complète et prête pour faire nos analyses.

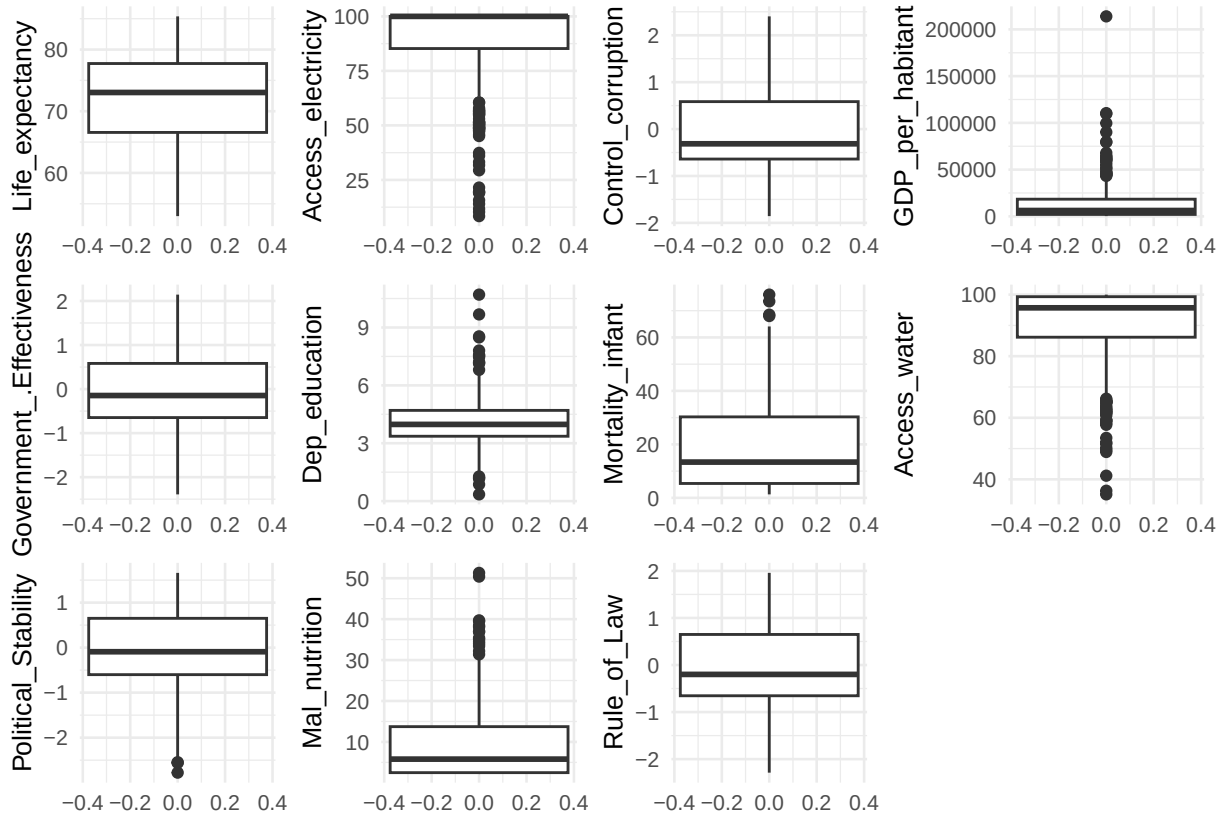
Statistiques descriptives :

```
##                            vars   n     mean       sd  median  trimmed     mad
## Pays*                         1 256   128.50    74.05  128.50   128.50   94.89
## Access_electricity            2 256    86.80    22.69  100.00    91.69    0.00
## Control_corruption            3 256    -0.05     0.96   -0.31    -0.10    0.92
## GDP_per_habitant              4 256 15490.14 23494.07 6260.00 10624.33 7311.33
## Government_.Effectiveness     5 256    -0.04     0.93   -0.15    -0.05    0.88
## Dep_education                 6 256     4.08     1.40    3.98     3.98    1.06
## Life_expectancy               7 256    72.05     7.48   73.04    72.38    7.72
## Mortality_infant              8 256    19.69    17.48   13.37    17.23   14.38
## Access_water                  9 256    89.36    14.40   95.72    92.10    6.20
## Political_Stability          10 256    -0.07     0.92   -0.09     0.01    0.92
## Mal_nutrition                11 256    10.07     9.88    5.80     8.16    4.89
## Rule_of_Law                  12 256    -0.05     0.95   -0.20    -0.07    0.93
##                               min       max     range  skew kurtosis      se
## Pays*                        1.00    256.00    255.00  0.00    -1.21    4.63
## Access_electricity           8.40    100.00     91.60 -1.72     1.92    1.42
## Control_corruption          -1.86      2.40      4.26  0.52    -0.48    0.06
## GDP_per_habitant           253.69 213937.01 213683.31  3.66    21.18 1468.38
## Government_.Effectiveness   -2.39      2.14      4.53  0.16    -0.37    0.06
## Dep_education                0.35     10.70     10.35  0.99     2.99    0.09
## Life_expectancy             53.00     85.38     32.38 -0.38    -0.64    0.47
## Mortality_infant             1.30     76.00     74.70  1.09     0.31    1.09
## Access_water                35.12    100.00     64.88 -1.58     1.61    0.90
## Political_Stability         -2.78      1.66      4.44 -0.60     0.11    0.06
## Mal_nutrition                2.50     51.30     48.80  1.64     2.39    0.62
## Rule_of_Law                 -2.29      1.96      4.25  0.20    -0.60    0.06
```

Nous constatons qu'il existe de fortes inégalités entre les pays. Par exemple, le PIB par habitant varie énormément, allant de moins de 300 dollars à plus de 200 000 dollars. Nous retrouvons la même chose pour la malnutrition, qui peut atteindre plus de 50% dans certains pays, ou encore pour la mortalité infantile, qui peut aller jusqu'à 76 décès pour 1 000 naissances. D'autres variables comme l'accès à l'électricité ou à l'eau potable ont des médianes très élevées (autour de 100%), mais certains pays restent encore en retard. Ces chiffres traduisent des écarts importants en matière de développement, ce qui est logique vu que nous travaillons à l'échelle mondiale. Pour mieux comprendre la répartition de chaque variable et détecter les éventuelles valeurs extrêmes, nous allons maintenant visualiser la distribution de chaque indicateur à l'aide de boxplots.

Statistiques univariées :

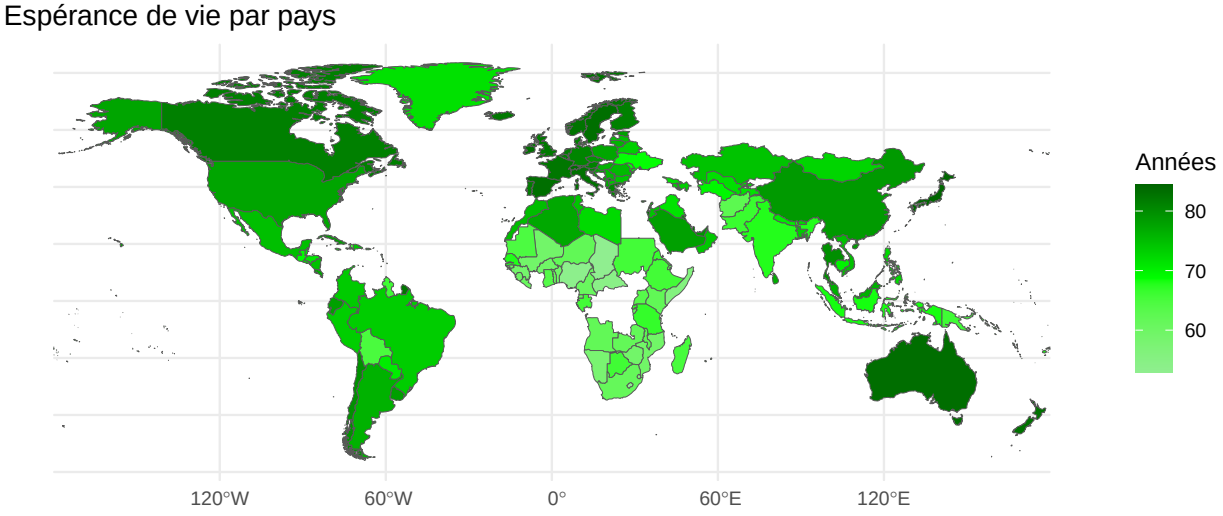
Visualisation de la distribution de chaque variable :



En observant la distribution de chacune des varibles, nous retrouvons ce que nous avions déjà constaté dans les statistiques descriptives : plusieurs variables présentent des valeurs extrêmes, comme l'accès à l'électricité, le PIB par habitant, l'accès à l'eau potable ou encore la malnutrition. Ces points très éloignés du reste des données correspondent sûrement à des pays en situation particulière, soit très développés, soit en grande difficulté. Ces visualisations confirment l'existence de fortes inégalités entre les pays pour ces indicateurs, ce qui est cohérent avec les disparités économiques, sanitaires et d'infrastructure que nous observons dans le monde.

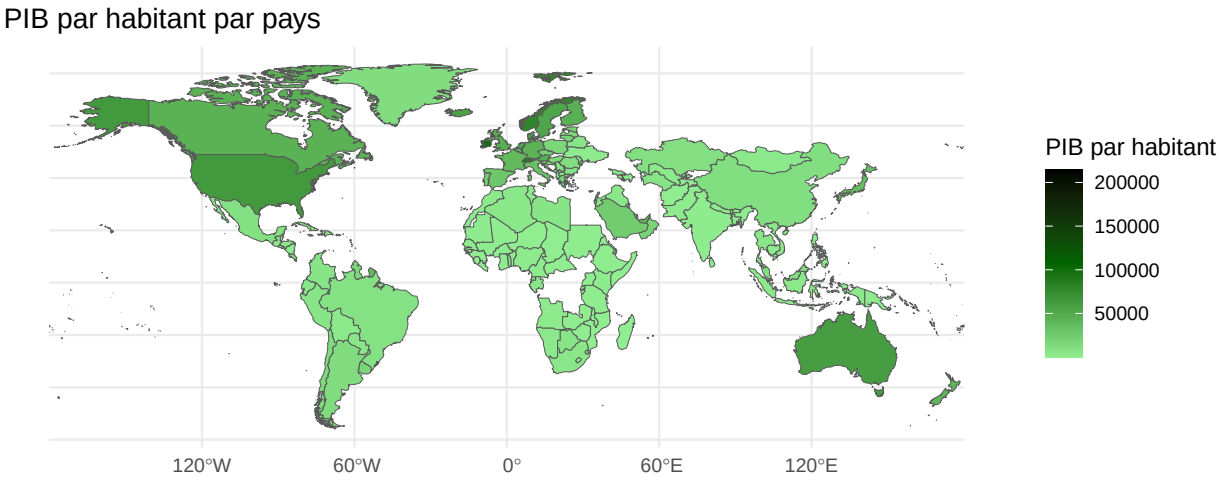
Cartographies :

Variable endogène : Espérance de vie



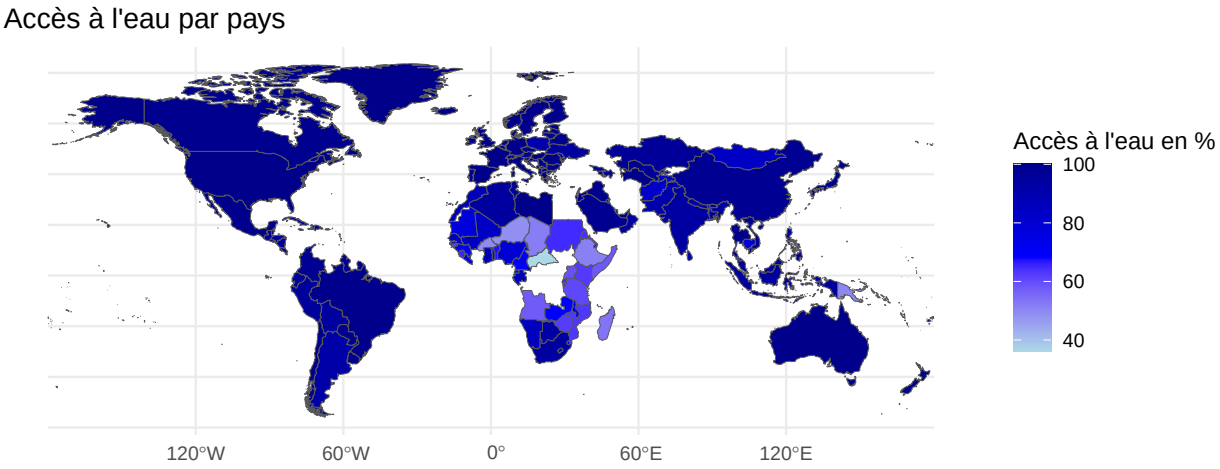
Nous remarquons en premier lieu que les pays d'Afrique ont une espérance de vie très basse comparé aux pays développés, comme le Japon par exemple. Nous avons ici un écart de 15 à 20 ans approximativement entre les pays en voie de développement et les pays développés. Nous verrons ensuite quels pourraient être les facteurs qui explique ces écarts entre ces pays.

PIB par habitant :



Nous pouvons constater que les pays développés ont en effet un PIB par habitant plus élevé que les pays en voie de développement et émergents. Nous remarquons également, du moins dans la base de données que Monaco est une valeur aberrante avec un PIB par habitant de 213 937$ en dollars constants 2015. Cela s'explique car c'est une principauté qui regroupe des individus très riches, près d'un habitant sur deux est millionnaire.

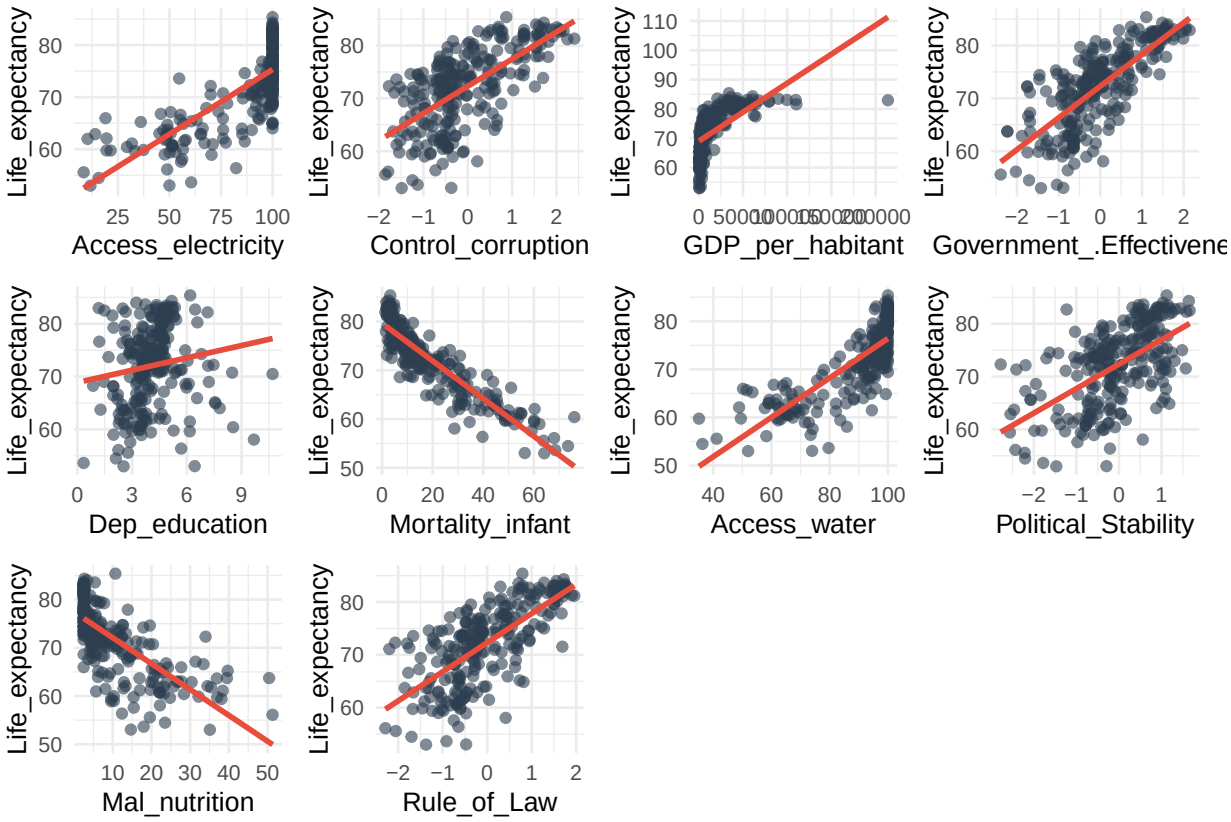
Accès à l'eau :



Cette cartographie illustre les problèmes d'accès à l'eau de la plupart des pays d'Afrique. En effet, ceci est dû au climat qui est très aride, des problèmes liés à la gestion de l'eau comme le manque d'infrastructures et d'autres facteurs. Ceci aura de fortes chances d'impacter le niveau de vie, surtout dans les pays d'Afrique.

Statistiques bivariées :

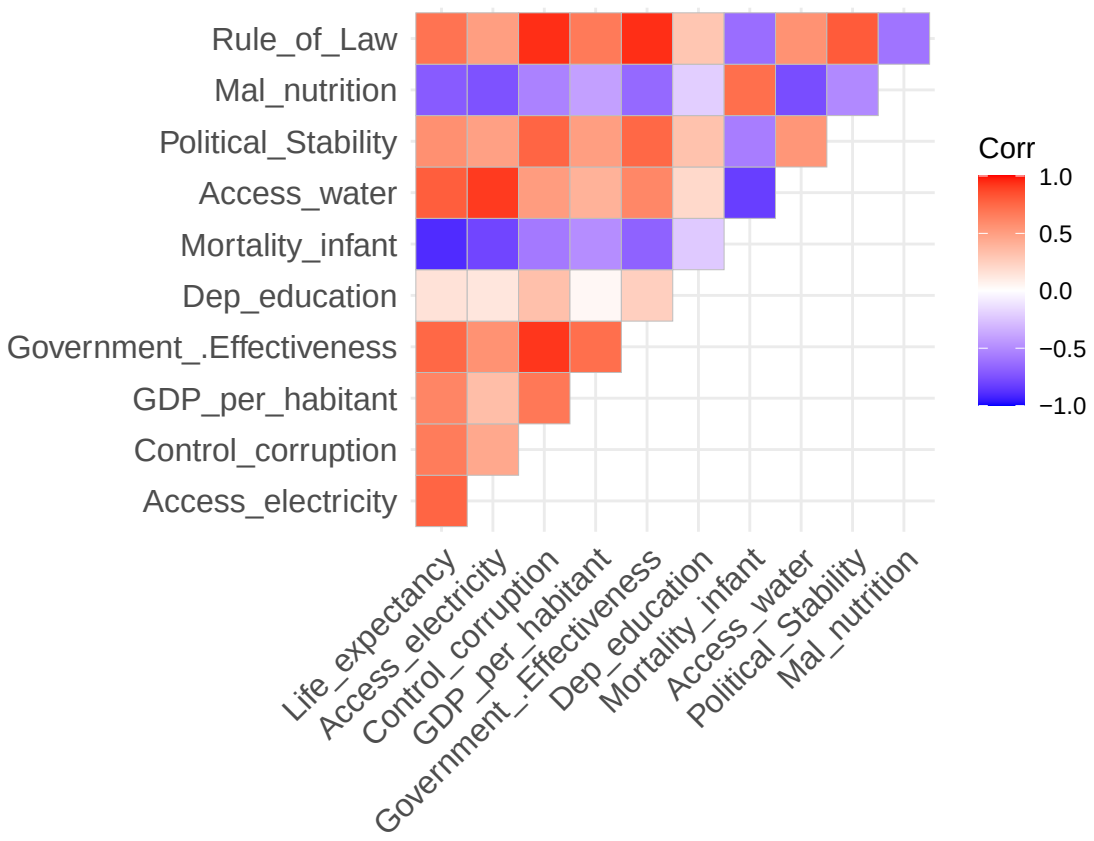
Nuages de points pour visualiser la relation entre chaque variable et l'espérance de vie :



En analysant les nuages de points entre l'espérance de vie et chaque variable explicative, nous retrouvons globalement les effets auxquels nous nous attendions dès le départ. Certaines variables semblent avoir une relation positive avec l'espérance de vie, comme le PIB par habitant, l'accès à l'eau potable, l'accès à l'électricité, les dépenses en éducation, ou encore des indicateurs institutionnels comme l'efficacité du gouvernement. Cela correspond bien à notre intuition : plus un pays est développé, mieux ses citoyens vivent longtemps.

À l'inverse, nous observons une relation négative pour des variables comme la malnutrition ou la mortalité infantile, ce qui est aussi logique, car elles reflètent des problèmes graves de santé publique.

Matrice de corrélation des variables :



En observant la matrice de corrélation, nous constatons qu'il existe une très forte corrélation positive entre certaines variables, notamment entre le contrôle de la corruption et l'état de droit, ou encore entre l'efficacité du gouvernement et le contrôle de la corruption. Ces corrélations sont logiques car ces variables mesurent toutes des aspects institutionnels liés à la qualité de la gouvernance. Nous remarquons aussi que l'espérance de vie est fortement corrélée négativement avec la malnutrition et la mortalité infantile, ce qui est cohérent avec notre intuition : plus ces indicateurs sont élevés, plus la qualité de vie est faible, donc l'espérance de vie diminue. À l'inverse, des variables comme le PIB par habitant, l'accès à l'eau ou l'électricité ont une corrélation positive avec l'espérance de vie, ce qui confirme l'importance des infrastructures de base et du niveau de développement dans la santé des populations.

Partie 1 : Endogénéité

Dans cette partie, nous allons nous intéresser au problème d'endogénéité, qui peut fausser les résultats d'une régression si elle n'est pas prise en compte. En regardant la matrice de corrélation, nous remarquons que la variable "contrôle de la corruption" est fortement corrélée à d'autres indicateurs institutionnels comme l'efficacité du gouvernement, l'état de droit et la stabilité politique. Cette forte corrélation peut révéler un problème d'erreur de mesure, ou même une relation causale inversée, ce qui rend cette variable potentiellement endogène.

Pour tester cette hypothèse, nous commencons par estimer le modèle restreint, en excluant les variables qui risquent de causer de l'endogénéité (efficacité du gouvernement, état de droit et stabilité politique). Ensuite, nous utilisons le test de Hausman pour vérifier si le "contrôle de la corruption" est réellement endogène ou non. Si le test confirme l'endogénéité, nous corrigerons le biais en utilisant des méthodes adaptées comme les variables instrumentales ou la méthode des doubles moindres carrés (2SLS).

Estimation du modèle restreint :

```
##
## Call:
## lm(formula = Life_expectancy ~ Access_electricity + Control_corruption +
##     GDP_per_habitant + Dep_education + Mortality_infant + Access_water +
##     Mal_nutrition, data = data_etude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2452 -1.5632  0.1026  1.7048  9.3954
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.463e+01  2.629e+00  28.393  < 2e-16 ***
## Access_electricity  1.586e-02  1.910e-02   0.830  0.40726
## Control_corruption  8.462e-01  2.896e-01   2.922  0.00380 **
## GDP_per_habitant    5.502e-05  1.056e-05   5.211 3.95e-07 ***
## Dep_education      -4.061e-01  1.364e-01  -2.978  0.00319 **
## Mortality_infant   -2.797e-01  1.967e-02 -14.224  < 2e-16 ***
## Access_water        3.044e-02  3.273e-02   0.930  0.35333
## Mal_nutrition      -3.274e-02  2.900e-02  -1.129  0.25991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.727 on 248 degrees of freedom
## Multiple R-squared:  0.8707, Adjusted R-squared:  0.867
## F-statistic: 238.5 on 7 and 248 DF,  p-value: < 2.2e-16
```

```
## Access_electricity Control_corruption   GDP_per_habitant      Dep_education
##           6.439531           2.627250          2.109540           1.250736
##   Mortality_infant       Access_water      Mal_nutrition
##           4.052070           7.612436          2.817055
```

Ici, nous voyons que le modèle est globalement bon, avec un R^2 d'environ 87%, donc 87% des variations des données sont expliquées par le modèle, toutes choses égales par ailleurs. Par contre, certaines variables comme l'accès à l'électricité, l'accès à l'eau et la malnutrition ne sont pas significatives. Une raison possible, c'est qu'elles sont très corrélées avec d'autres variables du modèle. Par exemple, l'accès à l'eau est fortement lié à l'accès à l'électricité, et la malnutrition est très liée à la mortalité infantile. Du coup, c'est compliqué de voir leur effet individuel. Les VIF confirment ça : nous voyons que l'accès à l'électricité (6.4) et l'accès à l'eau (7.6) ont une multicolinéarité modérée, ce qui peut cacher leur impact sur la variable endogène.

Test d'endogéneité de Haussman :

```
##
## Call:
## ivreg(formula = Life_expectancy ~ Control_corruption + GDP_per_habitant +
##     Dep_education + Mortality_infant + Mal_nutrition + Access_water +
##     Access_electricity | Government_.Effectiveness + Rule_of_Law +
##     Political_Stability + GDP_per_habitant + Dep_education +
##     Mortality_infant + Mal_nutrition + Access_water + Access_electricity,
##     data = data_etude)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -8.38334 -1.59173  0.09457  1.71316  9.42556
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         7.472e+01  2.632e+00  28.392  < 2e-16 ***
## Control_corruption  1.016e+00  3.348e-01   3.033  0.00267 **
## GDP_per_habitant    5.137e-05  1.116e-05   4.602 6.69e-06 ***
## Dep_education      -4.333e-01  1.391e-01  -3.115  0.00205 **
## Mortality_infant   -2.782e-01  1.973e-02 -14.099  < 2e-16 ***
## Mal_nutrition      -2.957e-02  2.919e-02  -1.013  0.31203
## Access_water        3.043e-02  3.275e-02   0.929  0.35372
## Access_electricity  1.614e-02  1.912e-02   0.844  0.39936
##
## Diagnostic tests:
##                  df1 df2 statistic p-value
## Weak instruments   3 246   244.678  <2e-16 ***
## Wu-Hausman         1 247     1.022   0.313
## Sargan             2  NA     3.500   0.174
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.729 on 248 degrees of freedom
## Multiple R-Squared: 0.8705,  Adjusted R-squared: 0.8668
## Wald test: 238.3 on 7 and 248 DF,  p-value: < 2.2e-16
```

Ici, nous avons utilisé la méthode des variables instrumentales pour voir si le contrôle de la corruption était endogène, c'est-à-dire s'il pouvait fausser notre modèle. Pour ça, nous avons utilisé comme instruments l'efficacité du gouvernement, l'état de droit et la stabilité politique, car ces variables sont liées à la corruption mais pas directement à l'espérance de vie.

Ensuite, nous avons fait le test de Hausman pour vérifier si le contrôle de la corruption est vraiment endogène. Le test donne un p-value de 0.31, donc nous ne rejetons pas l'hypothèse d'exogénéité. Ça veut dire que nous n'avons pas de preuve que cette variable pose problème. Du coup, le modèle MCO reste le plus adapté ici.

Estimation du modèle global :

```
##
## Call:
## lm(formula = Life_expectancy ~ ., data = data_etude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1113 -1.5139  0.1064  1.6845  9.6285
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 7.403e+01  2.651e+00  27.923  < 2e-16 ***
## Access_electricity          1.564e-02  1.924e-02   0.813  0.41708
## Control_corruption          3.405e-01  5.763e-01   0.591  0.55519
## GDP_per_habitant            4.916e-05  1.107e-05   4.442 1.35e-05 ***
## Government_.Effectiveness   8.041e-01  6.503e-01   1.236  0.21748
## Dep_education              -3.642e-01  1.375e-01  -2.648  0.00862 **
```

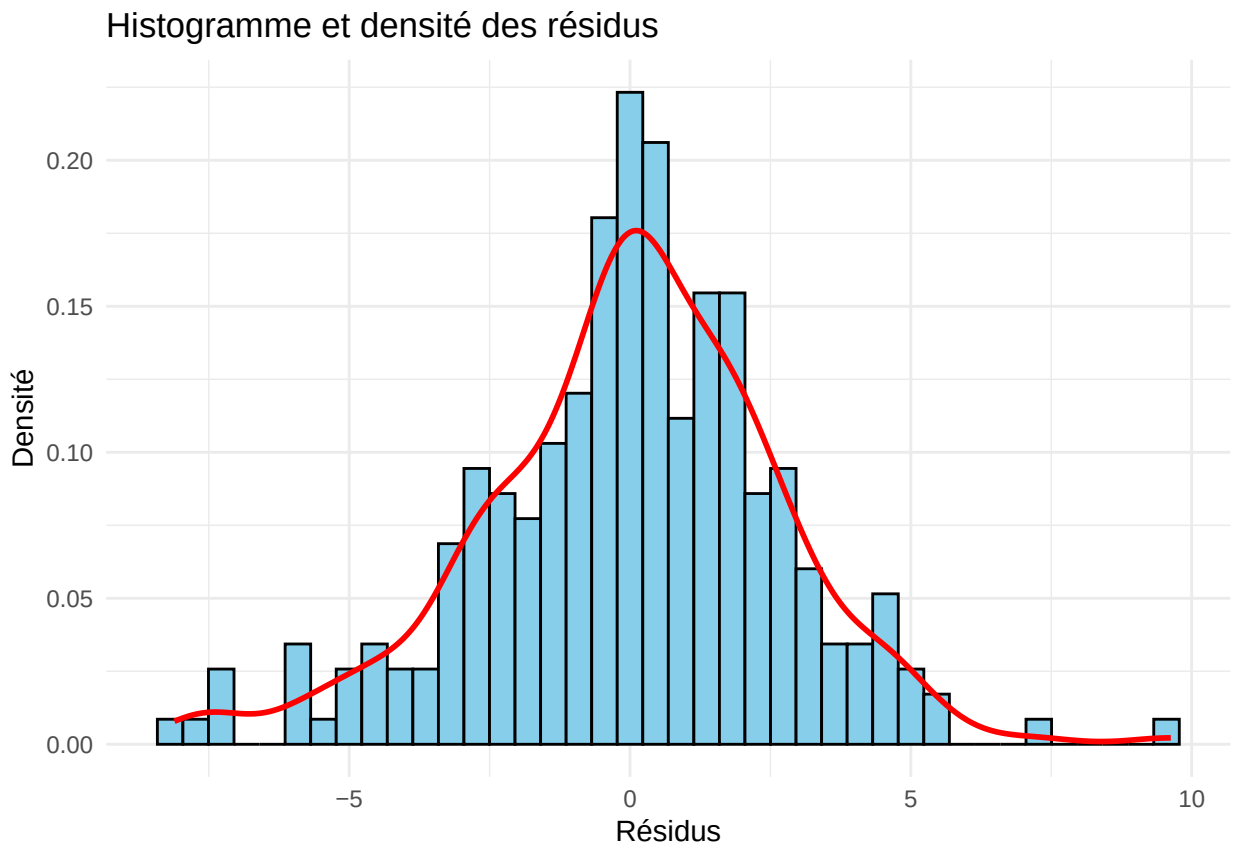
```
## Mortality_infant          -2.757e-01  1.977e-02 -13.943  < 2e-16 ***
## Access_water               3.418e-02  3.284e-02   1.041  0.29899
## Political_Stability       -4.779e-01  3.179e-01  -1.503  0.13408
## Mal_nutrition             -2.022e-02  2.972e-02  -0.680  0.49701
## Rule_of_Law                3.213e-01  7.089e-01   0.453  0.65076
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.719 on 245 degrees of freedom
## Multiple R-squared:  0.873,  Adjusted R-squared:  0.8678
## F-statistic: 168.4 on 10 and 245 DF,  p-value: < 2.2e-16
```

Comme précédemment, le test de Hausman a montré qu'il n'y avait pas de problème d'endogénéité, ce qui nous a permis d'estimer le modèle complet avec l'ensemble des variables.

Cependant, plusieurs variables ne ressortent pas significatives, comme l'accès à l'eau, l'électricité, la corruption, ou encore l'état de droit. Cela peut s'expliquer par un problème de multicolinéarité, car certaines de ces variables sont très corrélées entre elles, ce qui rend leur effet individuel difficile à isoler. Nous allons donc traiter ce problème dans la prochaine partie, en utilisant des méthodes comme la PCA, le Ridge ou le Lasso, pour mieux gérer les variables corrélées.

Test de normalité des residus :

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.9849, p-value = 0.008309
```



Pour vérifier si les résidus du modèle global suivent une loi normale, nous avons utilisé le test de Shapiro-Wilk. La p-value obtenue est de 0.008, ce qui est inférieur à 0.05, donc nous rejettons l'hypothèse de normalité. Cela signifie que les résidus ne suivent pas parfaitement une loi normale. Pourtant, visuellement, l'histogramme avec la courbe de densité montre une forme assez proche d'une distribution normale. Cette légère non-normalité ne remet pas en cause la validité globale du modèle, mais nous gardons ça en tête pour interpréter correctement les résultats.

Partie 2 : Problème de multicolinéarité

Dans cette partie, nous nous intéressons au problème de multicolinéarité, c'est-à-dire quand plusieurs variables explicatives sont fortement corrélées entre elles. Cela peut poser un problème dans une régression MCO, car ça rend difficile pour savoir l'effet réel de chaque variable et cela peut fausser les résultats.

Pour corriger ce souci, nous allons utiliser deux types de méthodes :

Nous utiliserons d'abord, les méthodes de réduction de dimension, comme la PCA (Analyse en Composantes Principales) et la PLS (Moindres Carrés Partiels), qui permettent de résumer l'information des variables corrélées en un petit nombre de nouvelles variables.

Ensuite, nous utiliserons les méthodes de régularisation, comme le Ridge, le Lasso et l'Elastic Net, qui ajoutent une pénalité dans le modèle pour réduire l'impact des variables trop proches les unes des autres et rendre la régression plus fiable.

Méthode de réduction de dimension :

Analyse en composantes principales :

```
## Data:    X dimension: 256 10
##  Y dimension: 256 1
## Fit method: svdpc
## Number of components considered: 10
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           7.494    3.855    3.548    3.420    3.324    3.274    3.145
## adjCV        7.494    3.852    3.544    3.416    3.314    3.264    3.132
##        7 comps  8 comps  9 comps  10 comps
## CV       2.862    2.866    2.887     2.895
## adjCV    2.851    2.854    2.875     2.882
##
## TRAINING: % variance explained
##                   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                   62.54    76.42    86.31     90.7    93.90    96.23    98.04
## Life_expectancy     73.97    78.08    79.96     82.0    83.04    84.18    87.29
##                   8 comps  9 comps  10 comps
## X                   98.89    99.56     100.0
## Life_expectancy     87.30    87.30      87.3
```

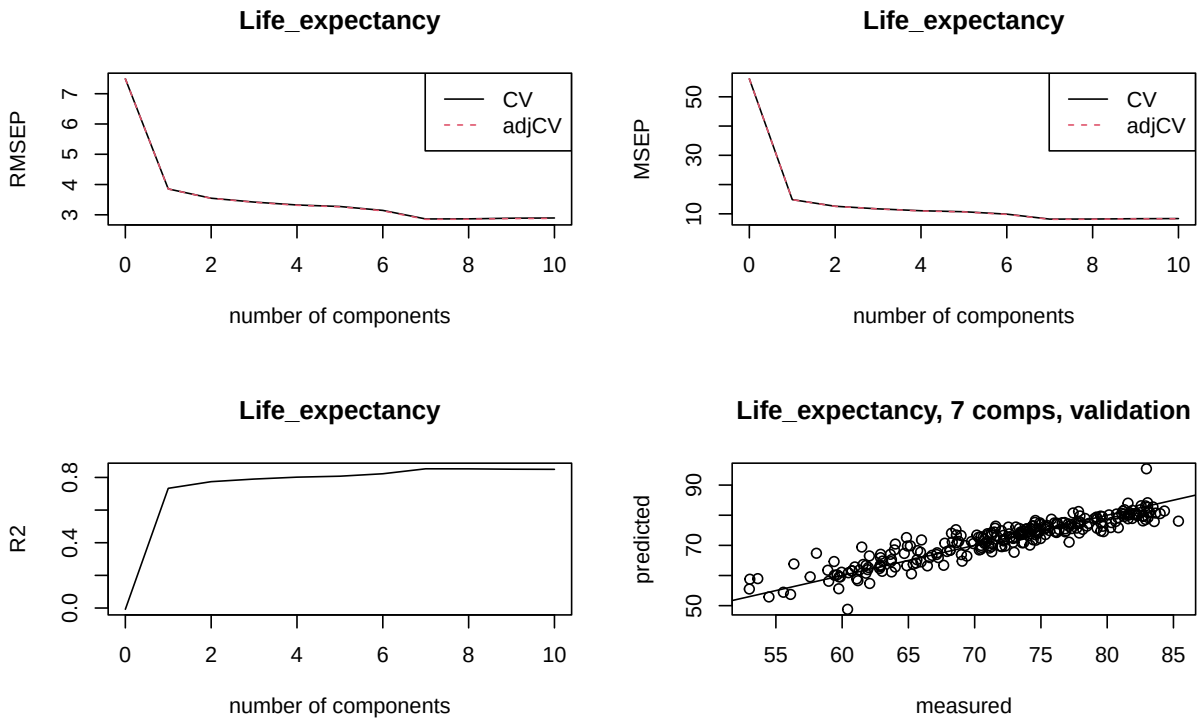
Nous constatons que l'erreur de validation croisée (CV) diminue progressivement à mesure que nous ajoutons des composantes, et atteint son minimum autour de la 7e composante. À ce niveau-là, l'erreur est la plus faible, ce qui veut dire que le modèle est le plus précis avec 7 composantes. En plus, nous voyons que les 7 premières composantes expliquent environ 98% de la variance des variables explicatives et plus de 87% de la variance de l'espérance de vie. Nous pouvons donc dire que 7 composantes suffisent largement pour capturer l'essentiel de l'information.

Pourcentage de variance expliquée par chaque composante :

```
##     Comp 1     Comp 2     Comp 3     Comp 4     Comp 5     Comp 6     Comp 7
## 62.5360761 13.8888182  9.8900039  4.3814648  3.1990406  2.3379141  1.8063975
##     Comp 8     Comp 9    Comp 10
##  0.8525506  0.6695981  0.4381361
```

Ici, nous voyons que la 1ère composante explique à elle seule plus de 62% de la variance des variables explicatives, la 2ème en ajoute environ 14%, et la 3ème presque 10%. À elles trois, elles expliquent donc plus de 85% de la variance totale, ce qui montre qu'elles résument déjà très bien les données. Lorsque nous allons jusqu'à la 7e composante, nous atteignons environ 98% de variance expliquée, ce qui veut dire que les 7 premières composantes suffisent largement pour capturer presque toute l'information.

Graphiques de validation pour le choix du nombre de composantes :



D'après les graphiques, nous observons que l'erreur de prédiction (RMSE) diminue rapidement jusqu'à la 7e composante, puis reste presque constante. Cela veut dire que la 7e composante est celle qui donne les meilleures performances en minimisant l'erreur. En parallèle, le R^2 augmente fortement jusqu'à cette composante et se stabilise ensuite, ce qui montre que 7 composantes suffisent à expliquer la majeure partie de la variance. Enfin, le graphique de validation montre que les valeurs prédites sont très proches des valeurs observées, ce qui confirme la bonne qualité du modèle basé sur les 7 composantes.

Coefficients des composantes :

```
## , , 7 comps
##
##                           Life_expectancy
## Access_electricity              0.2375861
## Control_corruption              0.2605328
## GDP_per_habitant                1.1576012
## Government_.Effectiveness       0.6726864
## Dep_education                  -0.5182464
## Mortality_infant               -4.8021630
## Access_water                    0.6383380
## Political_Stability            -0.4616103
## Mal_nutrition                  -0.1940582
## Rule_of_Law                     0.4551141
```

Nous remarquons que les coefficients sont beaucoup plus petits que ceux du modèle MCO. Cela montre que la PCR réduit fortement l'impact de chaque variable individuelle. C'est normal : le modèle ne travaille plus directement sur les variables d'origine, mais sur des combinaisons linéaires de celles-ci à travers les composantes principales. Du coup, les coefficients sont plus faibles, ce qui permet aussi de réduire les effets de la multicolinéarité.

Visualisation des coefficients :

Nous retrouvons ici les signes des coefficents que nous avons observés précédemment : par exemple, la mortalité infantile a un effet très négatif sur l'espérance de vie, tandis que l'efficacité du gouvernement et l'état de droit ont des effets positifs. Ce qui est intéressant aussi, c'est de regarder les intervalles de confiance autour de chaque coefficient. Pour certaines variables comme l'accès à l'électricité, le contrôle de la corruption ou encore la malnutrition, nous voyons que les barres d'erreur croisent la ligne zéro. Cela veut dire que leurs effets ne sont pas significatifs statistiquement.

Prédiction :

Définition de la base d'appretissage et de test : Ici nous avons séparé nos données : 80% vont servir à l'apprentissage du modèle, et les 20% restants seront utilisés pour tester sa capacité à bien prédire sur de nouvelles données.

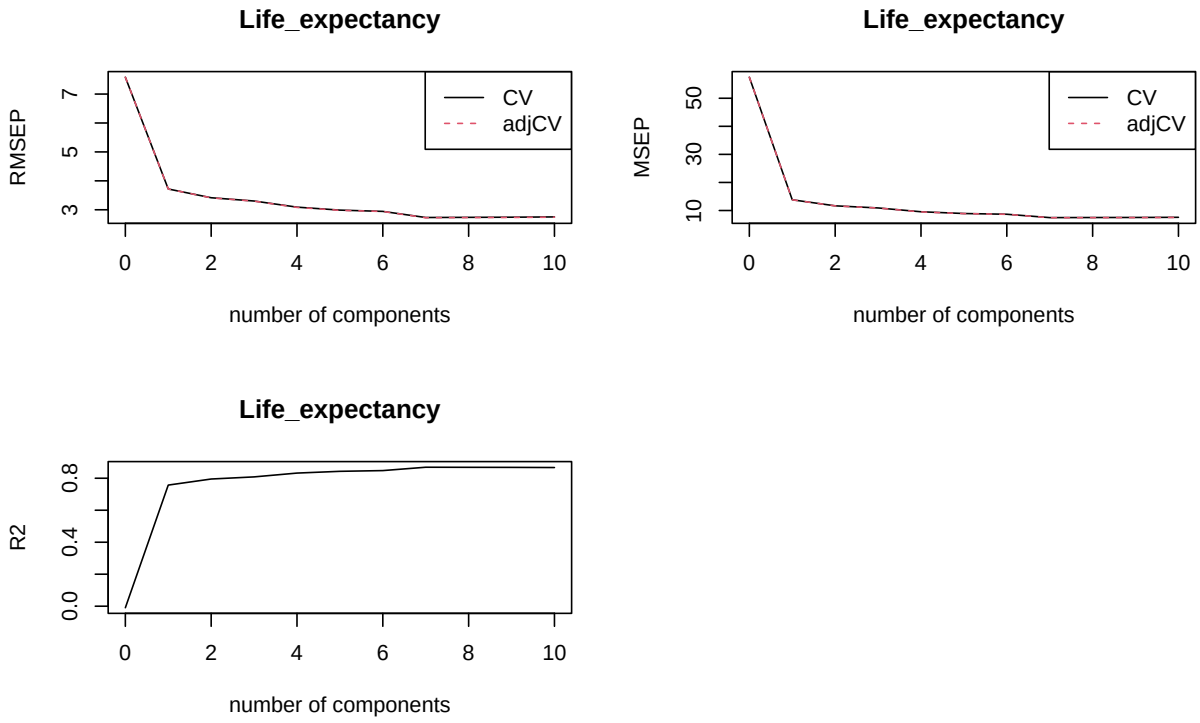
Estimation sur la base d'apprentissage :

```
## Data:    X dimension: 208 10
##  Y dimension: 208 1
## Fit method: svdpc
## Number of components considered: 10
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           7.582    3.718    3.416    3.303    3.091    2.989    2.944
## adjCV        7.582    3.717    3.412    3.299    3.082    2.981    2.942
```

```
##        7 comps  8 comps  9 comps  10 comps
## CV       2.733    2.738    2.743     2.753
## adjCV    2.725    2.730    2.735     2.744
##
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                  64.38    77.80    87.14    91.46    94.52    96.53    98.20
## Life_expectancy    76.13    80.31    81.72    84.10    85.72    86.13    88.24
##                  8 comps  9 comps  10 comps
## X                  98.97    99.59    100.00
## Life_expectancy    88.25    88.28     88.29
```

Cette estimation du modèle sur les données d'apprentissage montre que l'erreur diminue jusqu'à la 8ème composante. Même si, précédemment, le minimum avait été atteint à la 7ème composante sur l'ensemble des données, cette petite différence peut s'expliquer par le fait que nous travaillons ici uniquement sur un sous-échantillon. Dans tous les cas, à partir de la 7e ou 8e composante, l'amélioration devient très faible. Nous pouvons donc dire que 7 ou 8 composantes suffisent largement pour bien expliquer l'espérance de vie.

Graphiques de validation pour le choix du nombre de composantes :



En regardant les graphiques, nous remarquons que l'erreur de prédiction (RMSEP) diminue jusqu'à la 8e composante, où elle atteint son minimum. Mais à partir de la 7e composante, la courbe devient presque plate, ce qui veut dire que l'amélioration devient très faible. Le R^2 aussi reste stable à partir de là. Nous pouvons donc dire que 7 ou 8 composantes suffisent largement pour bien expliquer la variable "espérance de vie".

Erreur moyenne sur les données de test pour chaque composante :

```
## (Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps
##       7.105        4.327        3.892        3.801        3.829        3.980
##     6 comps      7 comps      8 comps      9 comps     10 comps
##       3.631        3.169        3.169        3.189        3.213
```

Sur les données de test, nous constatons que l'erreur RMSEP diminue jusqu'à la 7e composante, où elle atteint son minimum (environ 3.169). Cela montre que le modèle prédit le mieux à ce stade. Après la 7e composante, l'erreur remonte légèrement, ce qui confirme que trop de composantes peuvent réduire la performance sur des données nouvelles. Donc, même sur le test, 7 composantes reste le meilleur choix.

L'erreur moyenne sur la base de test pour les 7 composantes :

```
## [1] 3.168954
```

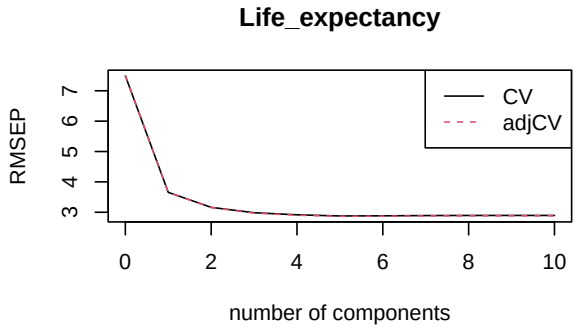
En utilisant les 7 composantes pour faire la prédiction sur les données de test, nous obtenons une erreur RMSE de 3.169, ce qui confirme les résultats précédents : le modèle est le plus performant avec 7 composantes.

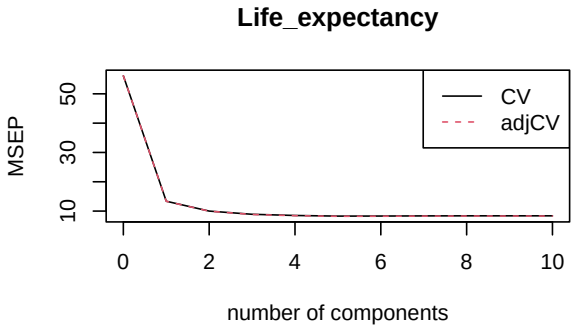
PLS (Estimation avec la méthode des moindres carrés partiels) :

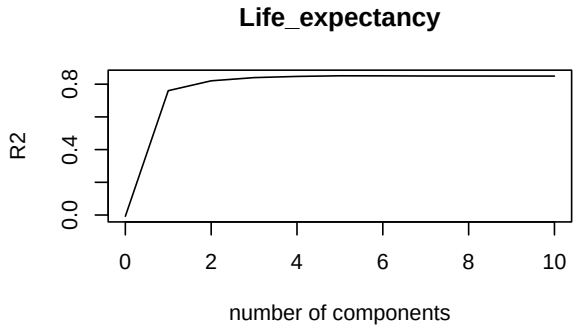
```
## Data:    X dimension: 256 10
##  Y dimension: 256 1
## Fit method: kernelpls
## Number of components considered: 10
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           7.494    3.655    3.162    2.985    2.916    2.878    2.882
## adjCV        7.494    3.652    3.157    2.972    2.902    2.866    2.870
##        7 comps  8 comps  9 comps  10 comps
## CV       2.891    2.894    2.895     2.895
## adjCV    2.878    2.881    2.882     2.882
##
## TRAINING: % variance explained
##                   1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                   62.36    74.74    80.98    88.79    92.54    95.49    97.77
## Life_expectancy     76.80    83.34    86.31    87.04    87.28    87.30    87.30
##                   8 comps  9 comps  10 comps
## X                   98.75    99.33     100.0
## Life_expectancy     87.30    87.30      87.3
```

Contrairement à la PCR, nous remarquons qu'avec la méthode PLS, l'erreur de prédiction (RMSE) diminue rapidement jusqu'à la 5e composante, puis se stabilise. De plus, le pourcentage de variance expliquée des variables explicatives atteint environ 92% dès la 5e composante. Cela montre que le modèle PLS arrive à capter presque toute l'information utile avec seulement 5 composantes.

Graphiques de validation pour le choix du nombre de composantes :







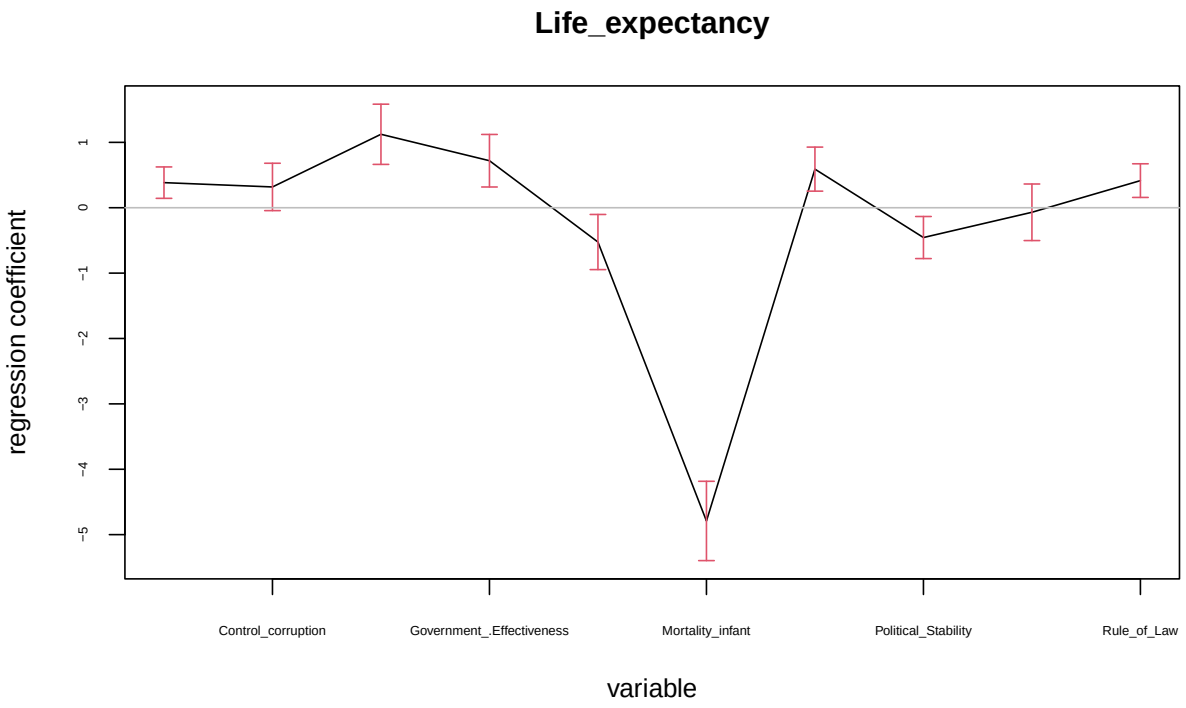
D'après les graphiques, nous voyons que l'erreur de prédiction (RMSEP) diminue rapidement jusqu'à la 5e composante, puis se stabilise. Le R^2 devient aussi presque constant après la 5e composante, ce qui confirme que 5 composantes suffisent pour bien modéliser l'espérance de vie avec la méthode PLS.

Coefficients des composantes :

```
## , , 5 comps
##
##                            Life_expectancy
## Access_electricity              0.38380412
## Control_corruption              0.31795347
## GDP_per_habitant                1.12335124
## Government_.Effectiveness       0.71909389
## Dep_education                  -0.52464503
## Mortality_infant               -4.79079036
## Access_water                    0.59042362
## Political_Stability            -0.45593662
## Mal_nutrition                  -0.06969595
## Rule_of_Law                     0.41477647
```

Nous remarquons que les coefficients sont beaucoup plus petits que ceux du modèle MCO. Cela montre que la PLS réduit aussi fortement l'impact de chaque variable individuelle.

Visualisation des coefficients pour chaque composante :



Comme pour la PCR, nous remarquons ici avec la méthode PLS que certains coefficients sont assez faibles ou sont non significatifs (comme par exemple le contrôle de la corruption ou la malnutrition). Globalement, les effets des variables restent cohérents, mais la méthode réduit leur intensité à cause de la régularisation.

Prédiction :

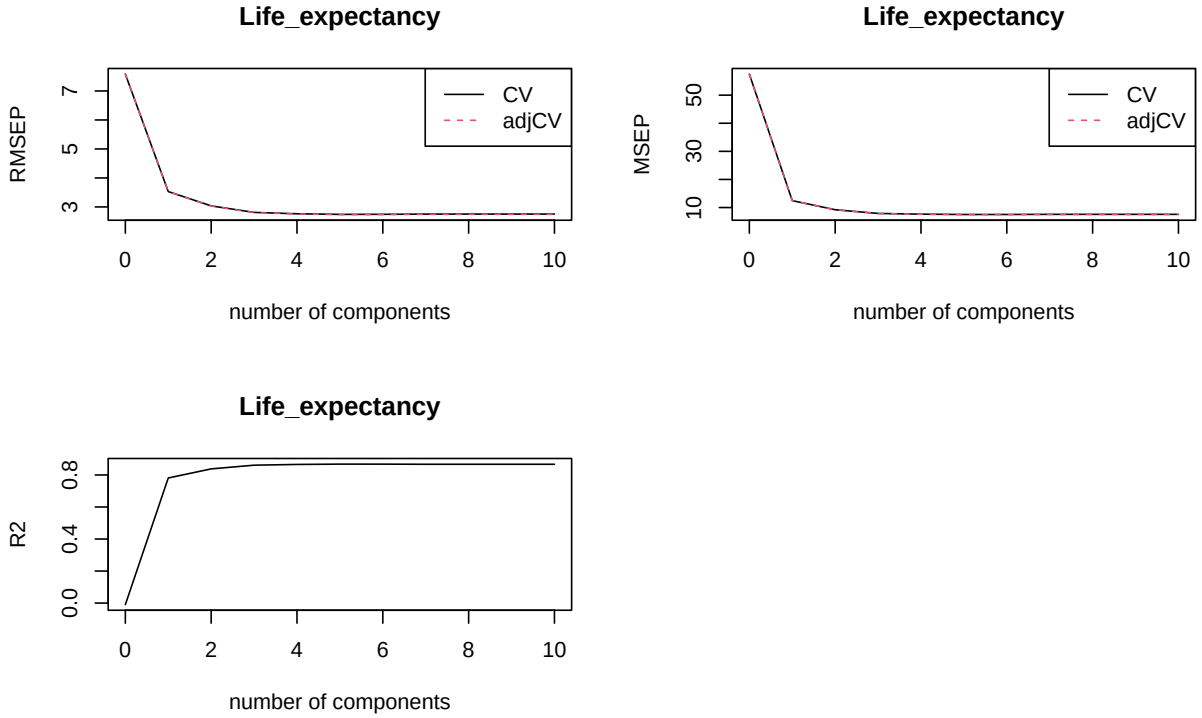
Estimation sur la base d'apprentissage :

```
## Data:    X dimension: 208 10
##  Y dimension: 208 1
## Fit method: kernelpls
## Number of components considered: 10
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           7.582    3.530    3.036    2.811    2.764    2.743    2.745
## adjCV        7.582    3.528    3.031    2.804    2.756    2.734    2.736
##        7 comps  8 comps  9 comps  10 comps
## CV       2.753    2.753    2.753     2.753
## adjCV    2.743    2.744    2.744     2.744
##
## TRAINING: % variance explained
##                  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X                  64.22    76.28    82.28    89.71    93.22    96.09    97.12
## Life_expectancy    78.68    84.93    87.49    88.02    88.26    88.28    88.29
##                  8 comps  9 comps  10 comps
## X                  98.73    99.32    100.00
```

```
## Life_expectancy    88.29    88.29     88.29
```

Sur les données d'apprentissage, comme pour la PCR, nous remarquons un petit décalage. L'erreur diminue jusqu'à la 5e composante, alors que sur l'ensemble des données, elle atteignait son minimum dès la 6e. Cela peut s'expliquer par le fait que nous travaillons ici sur un échantillon, donc les résultats peuvent légèrement varier.

Graphiques de validation pour le choix du nombre de composantes :



Sur les graphiques de validation croisée de la PLS, nous remarquons que l'erreur de prédiction (RMSEP et MSEP) diminue rapidement jusqu'à la 6e composante, puis devient quasiment stable. Le R^2 augmente également jusqu'à cette composante. Cela confirme que 6 composantes suffisent à bien expliquer l'espérance de vie dans notre échantillon d'apprentissage.

Erreur moyenne sur les données de test pour chaque composante :

```
## (Intercept)      1 comps      2 comps      3 comps      4 comps      5 comps
##       7.105        4.132        3.512        3.642        3.363        3.198
##     6 comps      7 comps      8 comps      9 comps     10 comps
##       3.178        3.220        3.210        3.213        3.213
```

Sur les données test, nous observons que l'erreur est minimale à la 6e composante, ce qui confirme les résultats précédents de façon approximative : le modèle PLS atteint une bonne performance prédictive avec seulement 6 composantes.

L'erreur moyenne sur la base de test pour les 6 composantes :

```
## [1] 3.178372
```

En calculant l'erreur moyenne de prédiction sur les données test avec 6 composantes, nous obtenons une valeur d'environ 3.18. Cela correspond bien à l'erreur minimale observée précédemment pour la 6e composante, ce qui confirme qu'elle est suffisante pour bien prédire l'espérance de vie avec le modèle PLS.

Méthode de régularisation :

Ridge / Lasso / Elastic Net :

Centrage et réduction des données :

```
##  Life_expectancy   Access_electricity Control_corruption GDP_per_habitant
##  Min.   :-2.5472   Min.   :-3.45546   Min.   :-1.8935    Min.   :-0.6485
##  1st Qu.:-0.7336   1st Qu.:-0.06844   1st Qu.:-0.6208    1st Qu.:-0.5723
##  Median : 0.1324   Median : 0.58164   Median :-0.2805    Median :-0.3929
##  Mean   : 0.0000   Mean   : 0.00000   Mean   : 0.0000    Mean   : 0.0000
##  3rd Qu.: 0.7599   3rd Qu.: 0.58164   3rd Qu.: 0.6597    3rd Qu.: 0.1231
##  Max.   : 1.7823   Max.   : 0.58164   Max.   : 2.5611    Max.   : 8.4467
##  Government_.Effectiveness Dep_education      Mortality_infant
##  Min.   :-2.5150           Min.   :-2.66307   Min.   :-1.0523
##  1st Qu.:-0.6479           1st Qu.:-0.51176   1st Qu.:-0.8192
##  Median :-0.1117           Median :-0.07319   Median :-0.3615
##  Mean   : 0.0000           Mean   : 0.00000   Mean   : 0.0000
##  3rd Qu.: 0.6692           3rd Qu.: 0.44548   3rd Qu.: 0.6053
##  Max.   : 2.3420           Max.   : 4.73020   Max.   : 3.2210
##   Access_water     Political_Stability Mal_nutrition      Rule_of_Law
##  Min.   :-3.7676   Min.   :-2.95026    Min.   :-0.7658   Min.   :-2.3695
##  1st Qu.:-0.2217   1st Qu.:-0.58153    1st Qu.:-0.7658   1st Qu.:-0.6410
##  Median : 0.4418   Median :-0.02829    Median :-0.4321   Median :-0.1609
##  Mean   : 0.0000   Mean   : 0.00000    Mean   : 0.0000   Mean   : 0.0000
##  3rd Qu.: 0.6925   3rd Qu.: 0.77960    3rd Qu.: 0.3711   3rd Qu.: 0.7370
##  Max.   : 0.7393   Max.   : 1.87771    Max.   : 4.1711   Max.   : 2.1215
```

Ici, nous centrons et nous réduisons les données pour éviter les problèmes d'échelle entre les variables. Cela permet à toutes les variables d'avoir le même poids dans les méthodes de régularisation comme le Ridge, le Lasso ou l'Elastic Net, qui sont sensibles aux différences de grandeur entre les variables.

Séparation des donnees en données d'apprentissage et de test : Ici nous avons séparé encore une fois nos données : 80% vont servir à l'apprentissage du modèle, et les 20% restants seront utilisés pour tester sa capacité à bien prédire sur de nouvelles données.

Estimation Ridge :

```
## [1] 0.0869749
```

En appliquant la régression Ridge avec une validation croisée, nous obtenons une valeur optimale de lambda = 0.0869749. Ce qui signifie que le modèle Ridge est le plus performant avec ce niveau de régularisation.

Prédiction (Ridge) :

```
##        RMSE        R2
## 1 0.4001038 0.8209834
```

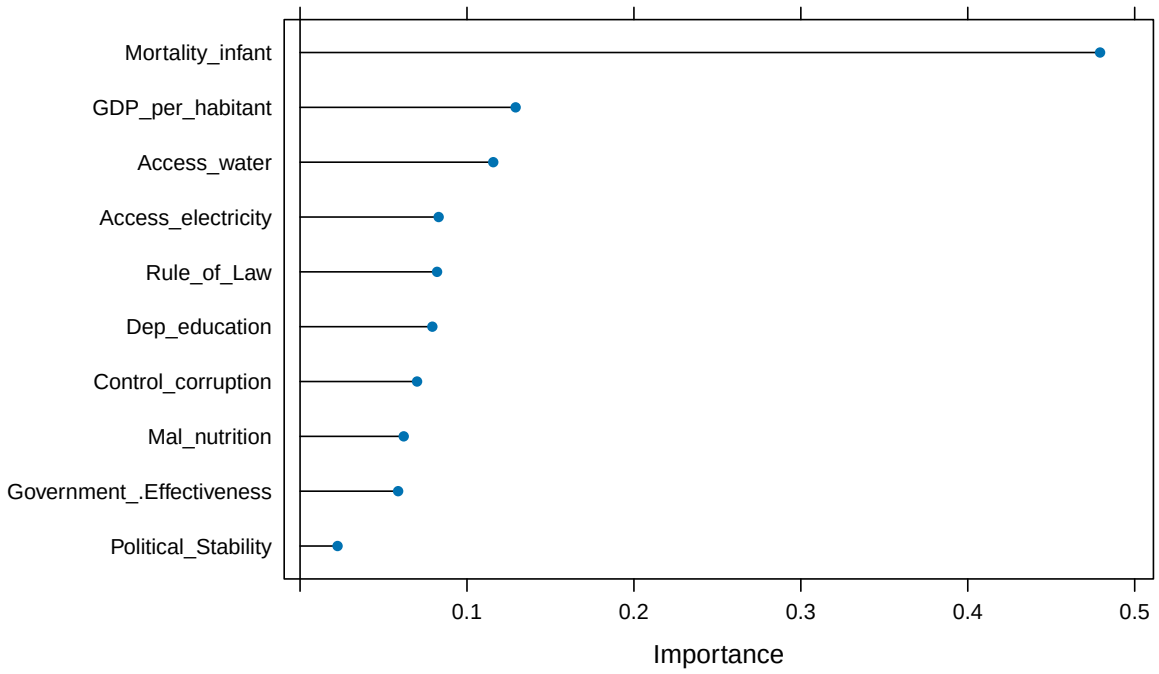
Ici, après avoir appliqué la régression Ridge sur les données de test, nous obtenons un R^2 d'environ 0.78, ce qui veut dire que le modèle explique 78% de la variation de l'espérance de vie. L'erreur quadratique moyenne (RMSE) est de 0.444, ce qui montre que les prédictions sont assez proches des valeurs réelles.

Coefficients :

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                      s1
## (Intercept)                -0.01016013
## Access_electricity          0.08301009
## Control_corruption          0.07008602
## GDP_per_habitant            0.12913064
## Government_.Effectiveness   0.05873798
## Dep_education              -0.07923391
## Mortality_infant           -0.47924824
## Access_water                0.11572078
## Political_Stability        -0.02245647
## Mal_nutrition              -0.06209018
## Rule_of_Law                 0.08206771
```

Nous remarquons que les coefficients estimés avec la régression Ridge sont beaucoup plus petits que ceux obtenus avec la régression MCO. C'est normal car Ridge applique une pénalisation qui réduit l'amplitude des coefficients pour limiter les effets de la multicolinéarité et éviter le surapprentissage. Même si les variables restent dans le modèle, leur impact individuel est modéré.

Importance des variables :



Nous observons que la variable la plus importante dans le modèle Ridge est la mortalité infantile, suivie par le PIB par habitant, l'accès à l'eau et l'accès à l'électricité. Ces variables semblent jouer un rôle central dans l'explication de l'espérance de vie.

Lasso :

Estimation lasso :

```
##    alpha     lambda
## 18     1 0.01072267
```

Nous avons estimé une régression Lasso avec validation croisée, et le meilleur lambda obtenu est d'environ 0.0123, avec alpha = 1 (car nous sommes en Lasso). Cela représente le niveau de pénalisation optimal : il permet de réduire certains coefficients sans trop impacter la qualité des prédictions du modèle.

Prédiction :

```
##        RMSE        R2
## 1 0.3918117 0.8283267
```

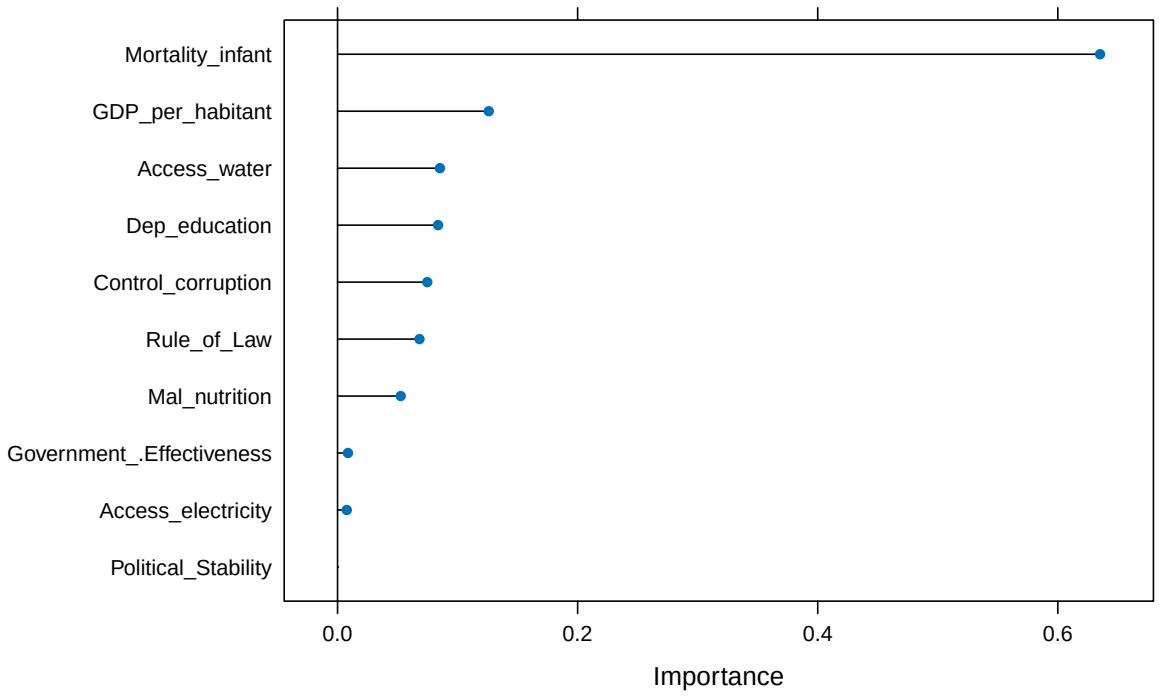
Après avoir appliqué la régression Lasso sur les données de test, Nous obtenons un R^2 d'environ 0.79, ce qui signifie que le modèle explique 79% de la variation de l'espérance de vie. L'erreur quadratique moyenne (RMSE) est de 0.431, ce qui montre que les prédictions sont proches des valeurs observées. Nous notons aussi une légère amélioration de la performance par rapport au modèle Ridge estimé précédemment.

Coefficients :

```
## 11 x 1 sparse Matrix of class "dgCMatrix"
##                                      s1
## (Intercept)                -0.01162232
## Access_electricity          .
## Control_corruption          .
## GDP_per_habitant            0.10669901
## Government_.Effectiveness   0.09951823
## Dep_education               .
## Mortality_infant           -0.61768740
## Access_water                0.06623348
## Political_Stability         .
## Mal_nutrition              -0.01632082
## Rule_of_Law                 0.01282332
```

Ici, nous voyons bien l'effet de la régression Lasso : plusieurs coefficients ont été réduits exactement à zéro (comme Control_corruption, dep_education, Political_Stability, Mal_nutrition, Rule_of_Law, etc.). Cela signifie que ces variables sont considérées comme non pertinentes pour le modèle final. Le Lasso a donc permis de faire une vraie sélection de variables en ne gardant que celles qui ont le plus d'impact sur l'espérance de vie.

Importance des variables :



Le graphique montre que seules quelques variables sont considérées comme importantes par le modèle Lasso, notamment Mortality_infant, GDP_per_habitant, et Access_electricity, acces_water. Cela confirme que Lasso fait une sélection automatique en gardant uniquement les variables les plus utiles pour prédire l'espérance de vie.

Elastic net :

Estimation Elastic net :

```
##   alpha     lambda
## 6   0.1 0.06415936
```

Dans cette estimation par la méthode Elastic Net, le meilleur modèle a été obtenu avec un alpha égal à 0.4 et un lambda d'environ 0.0277. Cela signifie que le modèle combine 40% de pénalisation Lasso (qui sélectionne les variables) et 60% de Ridge (qui régularise les coefficients), ce qui permet de profiter des avantages des deux méthodes.

Prédiction :

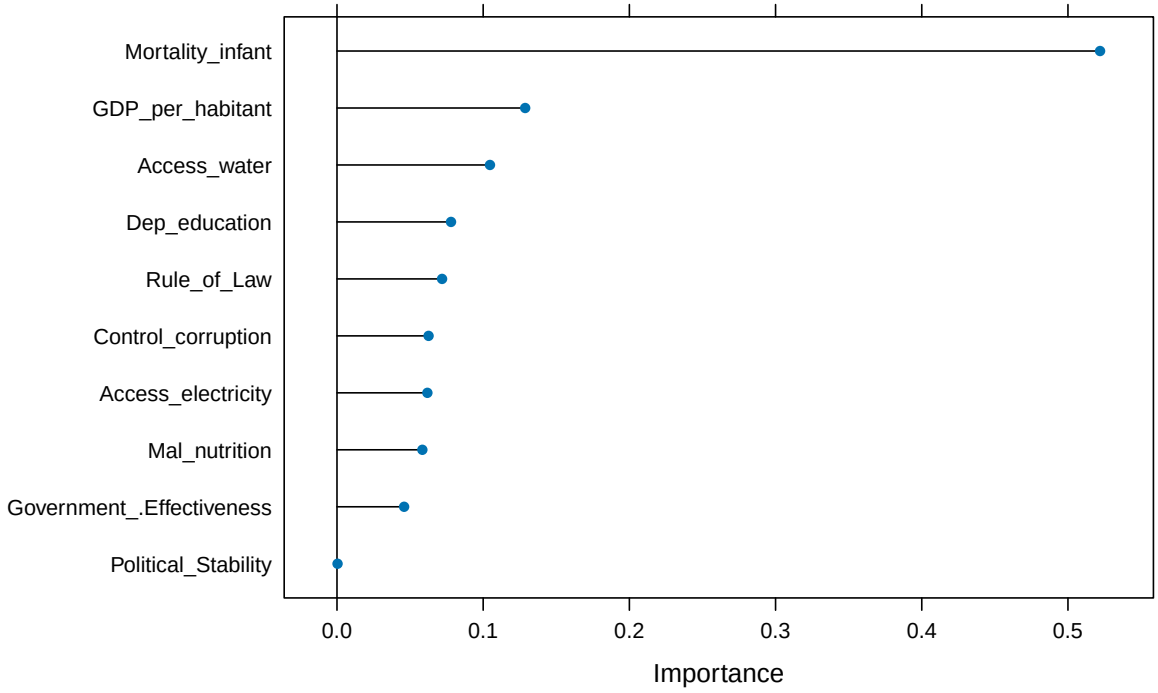
```
##        RMSE        R2
## 1 0.3970225 0.8237301
```

Ici, nous obtenons un R^2 d'environ 0.79 et un RMSE de 0.434. Ces résultats sont légèrement moins bons que ceux obtenus avec le modèle Lasso mais meilleur que le modèle Ridge, ce qui signifie que l'Elastic Net, dans ce cas précis, est entre les deux pour prédire l'espérance de vie.

Importance des variables :

```
## glmnet variable importance
##
##                              Overall
## Mortality_infant           0.5219595
## GDP_per_habitant           0.1287324
## Access_water               0.1046525
## Dep_education              0.0780021
## Rule_of_Law                0.0718825
## Control_corruption         0.0626486
## Access_electricity         0.0618392
## Mal_nutrition              0.0583944
## Government_.Effectiveness 0.0459110
## Political_Stability        0.0003923
```

Nous voyons que les variables les plus influentes pour prédire l'espérance de vie sont la mortalité infantile, le PIB par habitant, l'efficacité du gouvernement et l'accès à l'eau. En revanche, certaines variables comme la corruption, la stabilité politique, la malnutrition ou l'état de droit ont une importance de 0, ce qui veut dire que le modèle ne les a pas retenues car elles n'apportaient pas suffisamment d'information utile pour la prédiction.



Visuellement aussi, nous voyons clairement que la mortalité infantile est de loin la variable la plus importante dans le modèle Elastic Net, avec une influence beaucoup plus grande que les autres. Ensuite viennent le PIB par habitant, l'accès à l'électricité et l'accès à l'eau, qui ont une importance modérée. Les autres variables comme la stabilité politique ou encore la mal nutrition ont une importance proche de 0, ce qui confirme qu'elles ont été peu ou pas retenues par le modèle.

Comparaison des trois modèles :

```
##
## Call:
## summary.resamples(object = ., metric = "RMSE")
##
## Models: ridge, lasso, elastic
## Number of resamples: 10
##
## RMSE
##              Min.   1st Qu.    Median      Mean   3rd Qu.      Max. NA's
## ridge   0.2524763 0.3185922 0.4112579 0.3875408 0.4431279 0.4854907    0
## lasso   0.2313924 0.2684217 0.3917606 0.3731046 0.4492924 0.5120333    0
## elastic 0.2323136 0.2654303 0.3974614 0.3713942 0.4542200 0.5152779    0
```

En comparant les trois modèles avec la validation croisée, nous remarquons que l'Elastic Net présente le plus faible RMSE moyen (0.3591), juste devant le Lasso (0.3594) et le Ridge (0.3667). Même si les écarts sont faibles, l'Elastic Net montre une légère supériorité en termes de précision de prédiction.

Estimation d'un modèle MCO avec les variables retenues par l'Elastic Net:

```
##
```

```
## Call:
## lm(formula = Life_expectancy ~ . - Political_Stability, data = data_etude)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1456 -1.6297  0.1978  1.6767  9.3941
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                7.461e+01  2.629e+00  28.377  < 2e-16 ***
## Access_electricity         1.373e-02  1.924e-02   0.713  0.47638
## Control_corruption         3.033e-01  5.772e-01   0.525  0.59981
## GDP_per_habitant           5.021e-05  1.107e-05   4.534 9.04e-06 ***
## Government_.Effectiveness  8.376e-01  6.516e-01   1.285  0.19987
## Dep_education             -3.850e-01  1.372e-01  -2.807  0.00541 **
## Mortality_infant          -2.763e-01  1.982e-02 -13.944  < 2e-16 ***
## Access_water               3.089e-02  3.285e-02   0.940  0.34798
## Mal_nutrition             -2.293e-02  2.974e-02  -0.771  0.44159
## Rule_of_Law               -2.685e-02  6.717e-01  -0.040  0.96815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.726 on 246 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8671
## F-statistic: 185.9 on 9 and 246 DF,  p-value: < 2.2e-16
```

Ici, nous avons estimé un modèle MCO en retenant uniquement les variables considérées comme les plus importantes par l'Elastic Net, à l'exception de Political_Stability qui avait un poids nul. Ce modèle présente un bon ajustement global (R^2 ajusté environ égal à 0.87), ce qui indique qu'il explique bien l'espérance de vie. Parmi les variables, Mortality_infant et GDP_per_habitant ressortent comme fortement significatives. Certaines variables restent non significatives, ce qui peut s'expliquer par une colinéarité résiduelle ou un lien plus faible avec l'espérance de vie.

Partie 3: Le double machine learning

Dans cette partie, nous cherchons à estimer l'effet de la mortalité infantile sur l'espérance de vie à l'aide de la méthode Double Machine Learning (DML). Cette approche nous permet de mieux isoler l'effet causal d'une variable, tout en prenant en compte les autres variables explicatives comme facteurs de confusion. Nous avons choisi d'utiliser la variable "Mortality_infant" car, d'après les résultats obtenus avec les modèles Ridge, Lasso et Elastic Net, elle ressort comme la plus importante dans la prédiction de l'espérance de vie. Cela nous semble donc pertinent de la considérer comme variable explicative principale dans notre estimation DML.

Estimation avec Lasso en utilisant glmnet pour le DML :

Estimation du coefficient mortalité infantile et de son intervalle de confiance :

```
## INFO  [15:15:50.203] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 1/5)
## INFO  [15:15:50.359] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 2/5)
## INFO  [15:15:50.416] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 3/5)
## INFO  [15:15:50.487] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 4/5)
## INFO  [15:15:50.534] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 5/5)
```

```
## INFO  [15:15:50.716] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 1/5)
## INFO  [15:15:50.807] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 2/5)
## INFO  [15:15:50.858] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 3/5)
## INFO  [15:15:50.933] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 4/5)
## INFO  [15:15:51.023] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 5/5)
```

Ici, nous avons estimé l'effet de la mortalité infantile sur l'espérance de vie à l'aide de la méthode Double Machine Learning (DML). Le modèle utilise par défaut une validation croisée à 5 blocs, et s'appuie sur une régression Lasso avec validation croisée intégrée. Le paramètre de régularisation (lambda) est automatiquement sélectionné par le modèle à l'aide de cette validation croisée interne. Cela permet d'optimiser la précision de l'estimation tout en réduisant les risques de sur-apprentissage.

Résultat de l'estimation :

```
## Estimates and significance testing of the effect of target variables
##                   Estimate. Std. Error t value Pr(>|t|)
## Mortality_infant  -0.27585    0.02429  -11.36   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le coefficient estimé est de -0.27586, ce qui signifie qu'une augmentation de la mortalité infantile d'une unité est associée à une baisse d'environ 0.27586 unités de l'espérance de vie. De plus, la p_value < 0.001, ce qui montre que cet effet est négatif et statistiquement significatif.

Utilisation de boostrap pour la détermination des IC :

```
##                        2.5 %     97.5 %
## Mortality_infant -0.3217366 -0.2299667
```

L'intervalle de confiance bootstrap à 95% pour l'effet de la mortalité infantile est compris entre -0.323 et -0.229. Comme cet intervalle est entièrement négatif et ne contient pas zéro, cela confirme que l'effet estimé est significatif et négatif. Nous pouvons donc dire avec un bon niveau de certitude que la mortalité infantile a un impact défavorable sur l'espérance de vie.

Estimation avec détermination des paramètres de Tuning par l'algorithme : Ici, nous cherchons à améliorer l'estimation du modèle DML en déterminant le meilleur paramètre lambda grâce à une procédure de tuning automatique. L'algorithme va tester différentes valeurs de lambda (entre 0.05 et 0.1) en utilisant une validation croisée à 5 blocs et sélectionnera celle qui minimise l'erreur quadratique moyenne (MSE). Cela permet d'avoir un modèle plus précis et mieux régularisé.

Résultat de l'estimation

```
## INFO  [15:15:56.321] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_l' (iter 1/5)
## INFO  [15:15:56.335] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_l' (iter 2/5)
## INFO  [15:15:56.349] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_l' (iter 3/5)
## INFO  [15:15:56.363] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_l' (iter 4/5)
## INFO  [15:15:56.375] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_l' (iter 5/5)
## INFO  [15:15:56.450] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_m' (iter 1/5)
## INFO  [15:15:56.463] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_m' (iter 2/5)
## INFO  [15:15:56.481] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_m' (iter 3/5)
## INFO  [15:15:56.498] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_m' (iter 4/5)
## INFO  [15:15:56.511] [mlr3] Applying learner 'regr.glmnet' on task 'nuis_m' (iter 5/5)
```

```
## Estimates and significance testing of the effect of target variables
##                   Estimate. Std. Error t value Pr(>|t|)
## Mortality_infant  -0.27848    0.02509   -11.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##                       2.5 %    97.5 %
## Mortality_infant -0.325234 -0.231717
```

Nous constatons que l'estimation avec tuning a légèrement réduit le coefficient de la mortalité infantile par rapport au modèle initial. De plus, l'intervalle de confiance est un peu plus resserré, ce qui indique une meilleure précision. Le tuning semble donc améliorer un peu la qualité de l'estimation, même si la différence reste faible.

Résultat du paramètre choisi :

```
## $ml_l
## $ml_l$Mortality_infant
## $ml_l$Mortality_infant$family
## [1] "gaussian"
##
## $ml_l$Mortality_infant$use_pred_offset
## [1] TRUE
##
## $ml_l$Mortality_infant$lambda
## [1] 0.1
##
##
##
## $ml_m
## $ml_m$Mortality_infant
## $ml_m$Mortality_infant$family
## [1] "gaussian"
##
## $ml_m$Mortality_infant$use_pred_offset
## [1] TRUE
##
## $ml_m$Mortality_infant$lambda
## [1] 0.1
```

Le tuning a permis de déterminer une valeur optimale de lambda égale à 0.1. Même si la valeur utilisée automatiquement dans l'estimation initiale n'était pas directement observable, la comparaison des résultats montre que les deux modèles sont très proches en termes de coefficients, ce qui indique que le niveau de régularisation était déjà bien calibré. Le tuning a néanmoins permis de préciser davantage les intervalles de confiance.

Conclusion du DML:

Cette méthode nous a permis d'estimer l'effet de la mortalité infantile sur l'espérance de vie tout en contrôlant les autres variables explicatives pouvant agir comme facteurs de confusion. En utilisant des modèles Lasso avec validation croisée, nous avons pu obtenir une estimation robuste de cet effet. Les résultats montrent un impact négatif clair et précis de la mortalité infantile sur la l'espérance de vie. Enfin, le tuning des paramètres a permis de renforcer légèrement la qualité de l'estimation, en améliorant la précision du modèle.

Regroupement des pays selon leur profil d'espérance de vie :

Dans cette partie, nous utilisons l'algorithme de classification non supervisée k-means pour regrouper les pays en trois groupes (clusters) selon leur espérance de vie. Le but est de repérer des pays qui ont un profil similaire, et mieux comprendre les écarts de santé dans le monde. Cette approche permet de simplifier l'analyse en classant les pays par grands niveaux d'espérance de vie (faible, intermédiaire, élevée).

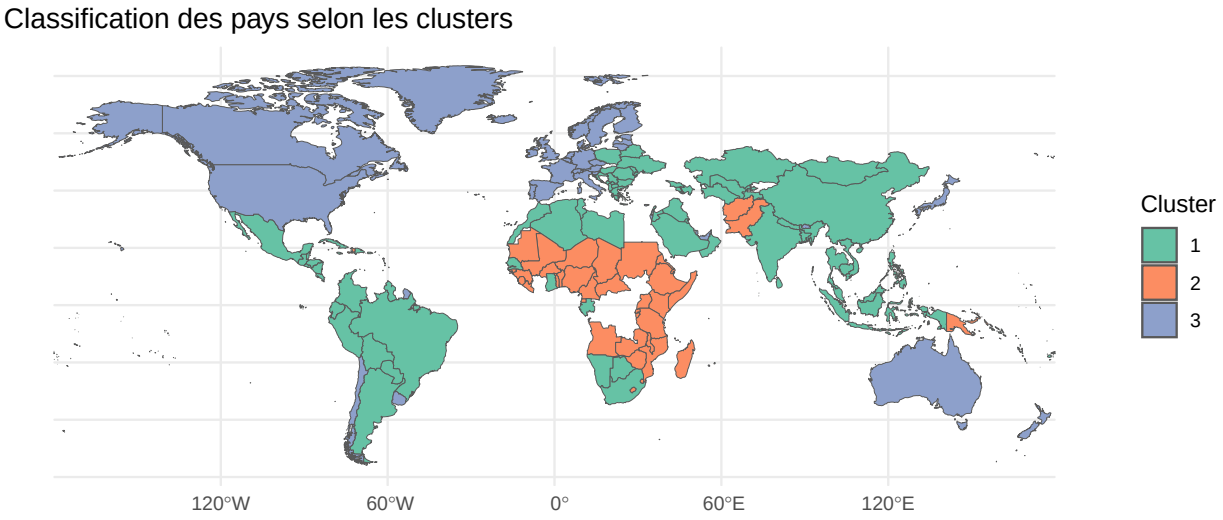
```
## # A tibble: 3 x 7
##   cluster Effectif Moyenne Médiane_LifeExp   Min   Max Écart_type
##   <fct>      <int>   <dbl>           <dbl> <dbl> <dbl>      <dbl>
## 1 1            135    72.7            73.1  58.1  80.8       4.04
## 2 2             59    61.7            61.9  53.0  72.3       3.73
## 3 3             62    80.4            81.6  71.5  85.4       3.18
```

Cluster 2 : Ce groupe correspond aux pays avec la meilleure espérance de vie (moyenne 80,4 ans). Ces pays affichent aussi une faible dispersion (écart-type 3,18), ce qui signifie que leur niveau de vie est assez homogène.

Cluster 1 : Le groupe le plus nombreux (135 pays) représente un niveau intermédiaire d'espérance de vie, autour de 72,7 ans. Ce cluster présente une plus grande diversité entre les pays, comme le montre un écart-type plus élevé.

Cluster 3 : Ce cluster regroupe les pays à faible espérance de vie (moyenne 61,7 ans). Il est caractérisé par un niveau de santé ou de développement plus préoccupant.

Représentation cartographique des groupes de pays selon l'espérance de vie :



Sur la carte, nous voyons que les pays sont regroupés en trois clusters selon leur espérance de vie. Le cluster 2 (en orange) regroupe surtout les pays développés comme ceux d'Europe, d'Amérique du Nord, d'Océanie et certains d'Asie de l'Est. Ces pays ont une espérance de vie élevée. Le cluster 3 (en bleu) est principalement composé de pays africains et quelques pays d'Asie du Sud, avec une espérance de vie plus faible. Enfin, le cluster 1 (en vert) représente une situation intermédiaire, avec des pays d'Amérique Latine, d'Europe de l'Est ou d'Asie centrale. Cela montre bien les inégalités géographiques en terme d'ésperance de vie.

Conclusion :

Ce projet a permis d'explorer en profondeur les déterminants de l'espérance de vie dans le monde en 2022, à travers une démarche économétrique rigoureuse alliant méthodes classiques (MCO, 2SLS) et approches modernes (régularisation, double machine learning). Les résultats mettent en évidence l'importance majeure de certains facteurs, notamment la mortalité infantile, le PIB par habitant, ainsi que l'accès aux infrastructures de base.

Malgré la complexité des interactions entre variables et les défis méthodologiques (multicolinéarité, endogénéité), l'utilisation de techniques avancées comme la PCA, le Lasso ou le DML a permis de dégager des effets robustes, tout en soulignant les limites des approches purement corrélationnelles.

Au-delà des chiffres, cette étude rappelle que l'espérance de vie est le fruit de politiques publiques cohérentes, d'institutions solides, et d'un accès équitable aux ressources essentielles. Dans un monde marqué par de profondes inégalités, ces résultats plaident pour des investissements ciblés dans les domaines de la santé, de l'éducation, et de la gouvernance, afin d'améliorer durablement la qualité de vie des populations