PROJET D'ÉCONOMÉTRIE AVANCÉE

Gyldano DADJEDJI, NOUCHET Kwami, TANO Marc

27-04-2024

Contents

1 Introduction et objectifs de l'Étude **2**

2 Cadre théorique et modélisation économétrique **2**

3 Collecte des données **3**

4 Partie 1: Etude de l'autocorrélation et de l'endogénéité **3**

4.1 Statistiques descriptives univariées 3

4.2 Statistiques descriptives bivariées 4

4.3 Spécification du modèle 5

4.4 Analyse des résidus 6

4.5 Etude de l'endogénéité 7

5 Partie 2: Analyse de de la multicolinéarité **10**

5.1 Détection de la multicolinéarité 10

5.2 Méthodes de réduction de dimension 11

5.3 Méthodes pénalisées 15

5.4 Régression Lasso 18

5.5 Régression Elastic Net 19

5.6 Comparaison des resultats 20

6 Partie 3: Causalité et double machine learning **21**

6.1 Mise à jour de la base et estimation par double machine learning simple 21

6.2 Estimation par double machine learning avec bootstrapping 21

7 Conclusion **22**

8 Bibliographie **23**

9 Contribution au projet **24**

1 Introduction et objectifs de l'Étude

Le Produit Intérieur Brut (PIB) par habitant est un indicateur économique clé qui reflète le niveau de richesse et de productivité d'une nation. Il est essentiel pour évaluer la performance économique et orienter les politiques de développement. Notre étude vise à identifier et à analyser les variables qui influencent le PIB par habitant au Japon, la cinquième plus grande économie mondiale. Nous cherchons à comprendre les facteurs sous-jacents qui contribuent à son évolution au fil du temps.

2 Cadre théorique et modélisation économétrique

De nombreuses théories économiques éclairent les déterminants du PIB par habitant. Parmi les facteurs les plus importants, on trouve:

L'activité économique: La Formation Brute de Capital Fixe (FBCF) est un indicateur des dépenses en capital qui reflète l'investissement dans les infrastructures et les biens de production. Selon la théorie de l'investissement, ces dépenses sont cruciales pour la croissance économique.

La performance environnementale : Les émissions de CO2 par habitant (CO2) sont intégrées pour évaluer l'impact de la performance environnementale sur l'économie. Les coûts liés à la pollution et au changement climatique peuvent affecter la santé publique et la productivité, influençant ainsi le PIB.

L'ouverture au commerce : L'ouverture au commerce international (Trade) est un facteur déterminant de l'intégration économique et de la spécialisation, ce qui peut entraîner des gains d'efficacité et influencer le PIB.

Des facteurs endogènes : Le PIB de l'année précédente est considéré comme un reflet de l'accumulation de capital et du progrès technologique, des éléments centraux de la théorie de la croissance endogène.

En combinant ces variables dans un modèle économétrique, nous pouvons examiner leur influence collective sur le PIB par habitant et fournir des insights pour des décisions politiques éclairées.

L'équation de notre modèle économétrique est donc formulée comme suit :

$$log.GDP.per.capita = b_0 + b_1 * FBCF + b_2 * log.CO2.per.capita + b_3 * log.GDP.per.capita_{t-1} + b_4 * Trade + e$$

3 Collecte des données

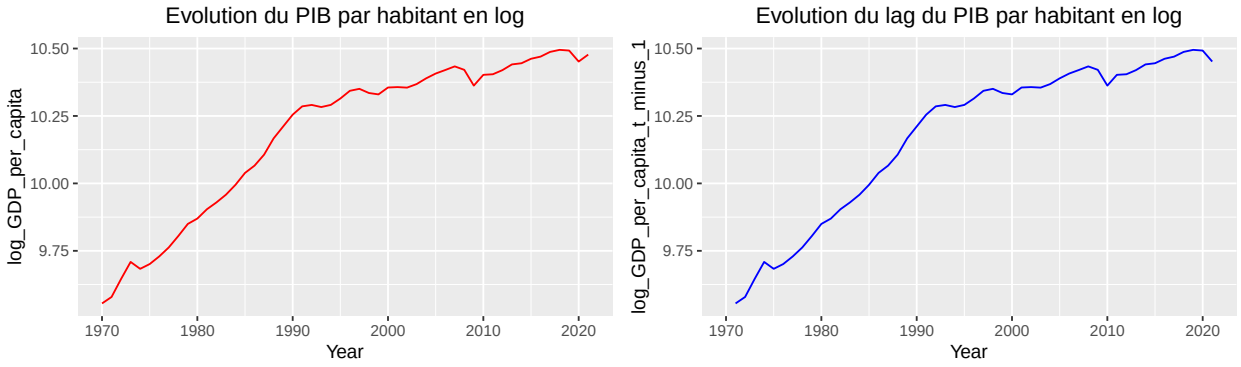
Nous avons initié notre étude en établissant une base de données à partir des informations collectées sur les sites de la Banque mondiale et de l'OCDE. Afin d'assurer une représentation plus précise des données, nous avons opté pour l'utilisation du logarithme du PIB par habitant (initialement en dollars US) ainsi que du CO2 par habitant (initialement en tonnes par habitant). L'ouverture commerciale et la formation brute de capital fixe sont exprimées en pourcentages du PIB pour une meilleure comparabilité.

```
# A tibble: 52 x 8
    Year  FBCF Trade log_GDP_per_capita log_CO2_per_capita
   <dbl> <dbl> <dbl>              <dbl>              <dbl>
 1  1970  38.9  19.2               9.55               1.94
 2  1971  37.5  19.5               9.58               1.97
 3  1972  37.4  17.8               9.65               2.00
 4  1973  39.9  18.9               9.71               2.11
 5  1974  38.1  26.3               9.68               2.08
 6  1975  35.5  24.1               9.70               2.03
 7  1976  34.2  24.8               9.73               2.05
 8  1977  33.0  23.1               9.76               2.07
 9  1978  33.3  19.3               9.80               2.05
10  1979  34.7  22.7               9.85               2.07
# i 42 more rows
# i 3 more variables: log_GDP_per_capita_t_minus_1 <dbl>,
#   log_GDP_per_capita_t_minus_2 <dbl>, log_GDP_per_capita_t_minus_3 <dbl>
```

4 Partie 1: Etude de l'autocorrélation et de l'endogénéité

4.1 Statistiques descriptives univariées

Pour approfondir notre compréhension du PIB par habitant, nous allons procéder à une analyse statistique desctriptive. Celle-ci consistera en la représentation graphique du PIB par habitant actuel et celui de l'année précédente. Cette démarche vise à identifier les tendances et les dynamiques temporelles qui caractérisent notre variable d'intérêt.

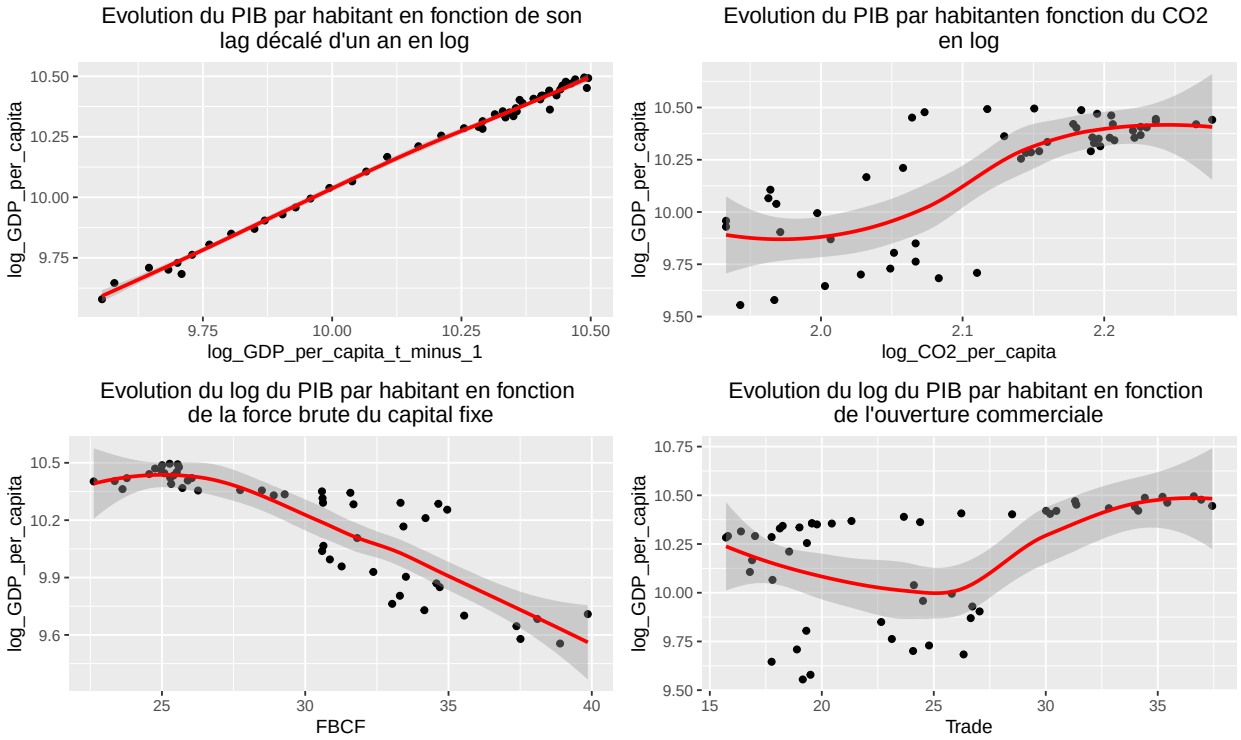


L'anlyse de l'évolution du PIB par habitant montre une relation forte et systématique entre ces deux variables, ce qui peut suggérer la présence d'une endogénéité dans notre modèle économétrique due à la simultanéité de ces deux variables. Cependant, pour établir l'endogénéité avec certitude, il est nécessaire d'effectuer des tests statistiques plus rigoureux, que nous aborderons plus tard dans le projet.

4.2 Statistiques descriptives bivariées

4.2.1 Nuages des variables

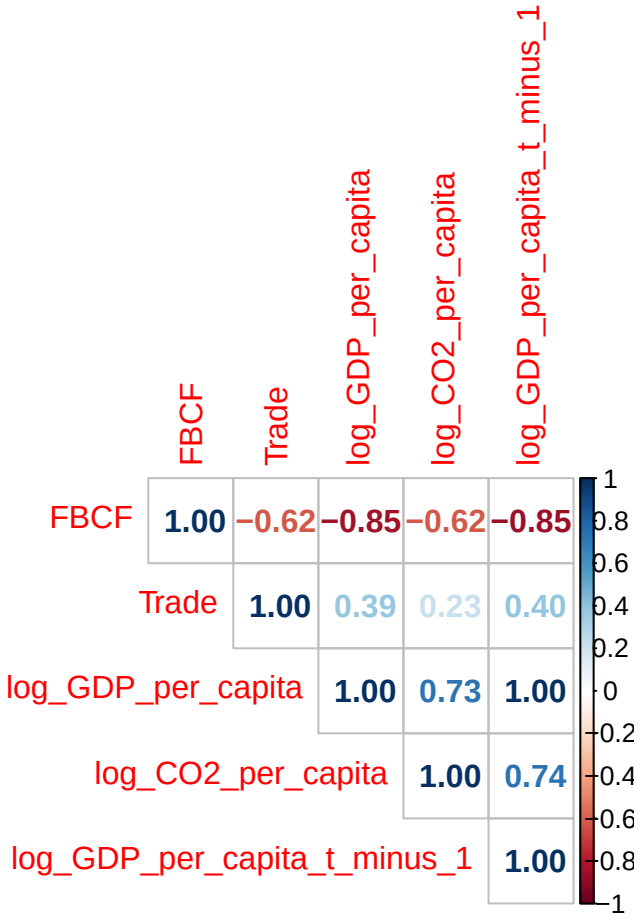
Pour en apprendre davantage sur le type de lien qui existe entre notre variable endogène et nos variables exogènes, nous utilisons une réprésentation graphique de nuage de points.



L'examen des graphiques suggère que le modèle linéaire est le plus adapté pour décrire la relation entre les variables explicatives et le PIB par habitant. Cette observation constitue une étape préliminaire essentielle, nous permettant de saisir la nature des liens entre les variables, étape nécessaire pour poursuivre au mieux la réalisation du modèle économétrique.

4.2.2 Matrice de corrélation

On représente ensuite les corrélations entre les différentes variables.



Notre analyse révèle une corrélation significative entre le PIB par habitant actuel et celui de l'année précédente, ainsi qu'entre le PIB par habitant et d'autres variables explicatives telles que la Formation Brute de Capital Fixe (FBCF) et les émissions de CO2 par habitant. Cette forte interdépendance souligne l'importance d'une étude approfondie du PIB par habitant décalé d'une année pour mieux comprendre ces dynamiques et affiner notre modèle économétrique dans la suite de notre projet.

4.3 Spécification du modèle

Pour expliquer le PIB par habitant, nous avons conçu un modèle économétrique initial basé sur les principes théoriques précédemment établis. Ce modèle intègre des variables clés qui sont susceptibles d'influencer le PIB par habitant, conformément à notre cadre théorique :

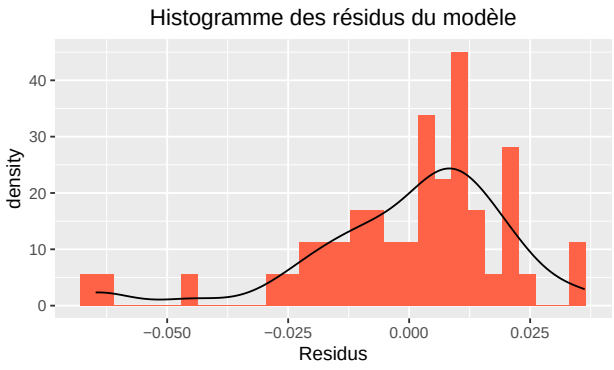
$$log.GDP.per.capita = b_0 + b_1 * FBCF + b_2 * log.CO2.per.capita + b_3 * log.GDP.per.capita_{t-1} + b_4 * Trade + e$$

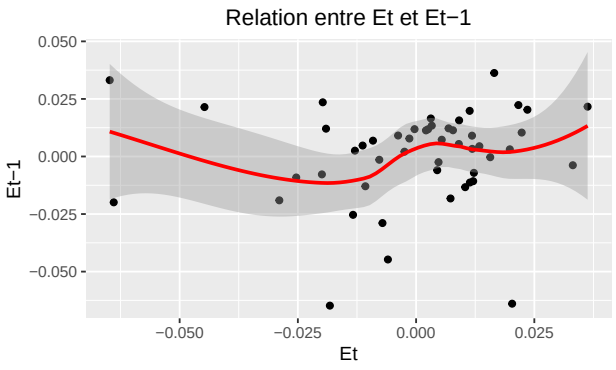
```
Call:
lm(formula = log_GDP_per_capita ~ Trade + FBCF + log_CO2_per_capita +
    log_GDP_per_capita_t_minus_1, data = don[, 2:6])

Residuals:
      Min        1Q    Median        3Q       Max
-0.064780 -0.009923  0.004464  0.011972  0.036268

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.3191346  0.2646505   1.206    0.234
Trade                       -0.0001756  0.0006223  -0.282    0.779
FBCF                         0.0006867  0.0015826   0.434    0.666
log_CO2_per_capita           0.0022845  0.0463288   0.049    0.961
log_GDP_per_capita_t_minus_1 0.9683669  0.0241781  40.051   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02155 on 46 degrees of freedom
  (1 observation effacée parce que manquante)
Multiple R-squared:  0.9944,    Adjusted R-squared:  0.994
F-statistic:  2055 on 4 and 46 DF,  p-value: < 2.2e-16
```

Une synthèse de ce modèle révèle que seul le PIB décalé d'une année est statistiquement significatif, accompagné d'un R^2 ajusté élevé. Cette situation inhabituelle peut effectivement suggérer un problème d'endogénéité. Pour investiguer davantage, il est judicieux de commencer par examiner les résidus du modèle afin de détecter d'éventuelles anomalies ou modèles qui pourraient indiquer des problèmes sous-jacents dans notre analyse économétrique.

4.4 Analyse des résidus

4.4.1 Histogramme des résidus et relation entre Et et Et-1





L'analyse des résidus de notre modèle économétrique indique une relation linéaire entre Et et Et-1, qui semble former une ligne horizontale. Cela suggère l'absence d'autocorrélation des résidus, un bon indicateur de la fiabilité du modèle. Cependant, l'histogramme des résidus laisse présager que ceux-ci ne suivent pas une distribution normale.

Pour confirmer ces observations, nous procéderons à des tests statistiques, tels que le test de Durbin-Watson pour l'autocorrélation et le test de Shapiro-Wilk pour la normalité, afin de vérifier les propriétés des résidus de notre modèle.

4.4.2 Test de normalité des résidus de Shapiro-Wilk

```
    Shapiro-Wilk normality test

data:  model$residuals
W = 0.91117, p-value = 0.001015
```

La p-value obtenue du test de normalité étant inférieure au seuil critique de 5%, nous rejetons l'hypothèse selon laquelle les résidus suivent une distribution normale. Cette non-normalité des résidus implique que les estimations des coefficients obtenues par la méthode des Moindres Carrés Ordinaires (MCO) pourraient être biaisées.

4.4.3 Test d'autocorrélation : Statistique H de Durbin Waston

Notre modèle étant dynamque, nous utilisons la statistique H de Durbin Waston pour évaluer l'autocorrorélaion. On effectue d'abord le test de durbin waston, et on a:

```
 lag Autocorrelation D-W Statistic p-value
   1      0.08824524      1.783704   0.188
 Alternative hypothesis: rho != 0
```

Ainsi, on déduit que:

$$Stat_{H.de.durbin} = \left(1 - \frac{1.78}{2}\right) \times \sqrt{\frac{51}{1 - (1.25 \times 10^{-6})^2}} = 0.7855 < 1.65$$

Avec une statistique inférieure à 1.65, on peut conclure qu'il n'existe pas d'autocorrélation, indiquant ainsi que les erreurs de notre modèle ne sont pas liées à une corrélation entre les observations successives, mais plutôt à une autre source d'erreur.

4.5 Etude de l'endogénéité

Selon la théorie économique, le PIB par habitant et le PIB par habitant décalé d'une année peuvent en effet avoir un impact l'un sur l'autre. Cela nous permet d'étudier l'endogénéité de notre modèle économique.

4.5.1 Recherche d'instruments alternatifs

Comme les données sont temporelles, il est possible de tester si les variables du PIB par habitant décalées dans le temps pourraient servir de bons instruments.

```
Call:
ivreg(formula = log_GDP_per_capita ~ Trade + log_GDP_per_capita_t_minus_1 +
    log_CO2_per_capita + FBCF | Trade + FBCF + log_CO2_per_capita +
    log_GDP_per_capita_t_minus_2 + log_GDP_per_capita_t_minus_3,
    data = don[-3, ])

Residuals:
      Min        1Q    Median        3Q       Max
```

```
-0.067376 -0.008270  0.002919  0.012416  0.036467

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  0.3063899  0.2785855   1.100    0.278
Trade                       -0.0002337  0.0006277  -0.372    0.711
log_GDP_per_capita_t_minus_1  0.9747212  0.0259810  37.517   <2e-16 ***
log_CO2_per_capita          -0.0173838  0.0487629  -0.356    0.723
FBCF                         0.0003658  0.0016379   0.223    0.824

Diagnostic tests:
                df1 df2 statistic p-value
Weak instruments   2  42   436.101  <2e-16 ***
Wu-Hausman         1  42     0.011   0.917
Sargan             1  NA     0.033   0.856
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02148 on 43 degrees of freedom
Multiple R-Squared: 0.993,  Adjusted R-squared: 0.9924
Wald test:  1518 on 4 and 43 DF,  p-value: < 2.2e-16
```

Sous l'hypothèse nulle, les instruments sont exogènes (non corrélés aux aléas). Les tests de Sargan et d'Hausman acceptent l'hypothèse nulle, montrant ainsi que les instruments choisis sont bien exogènes et non corrélés aux aléas du PIB par habitant.

4.5.2 Test des instruments faibles: weak instrument

```
Analysis of Variance Table

Model 1: log_GDP_per_capita_t_minus_1 ~ Trade + log_CO2_per_capita
Model 2: log_GDP_per_capita_t_minus_1 ~ log_GDP_per_capita_t_minus_2 +
    log_GDP_per_capita_t_minus_3 + Trade + log_CO2_per_capita
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     45 1.27154
2     43 0.04425  2    1.2273 596.31 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Le test de **weak instrument** indique que les variables **log_GDP_per_capita_t_minus_2** et **log_GDP_per_capita_t_minus_3** sont de bons instruments, car elles sont corrélées à la variable **log_GDP_per_capita_t_minus_1**.

4.5.3 Test d'endogénéité: Test de Hausman Wu

Nous examinons ainsi l'endogénéité de notre modèle en utilisant le **test de la régression augmentée**.

```
Call:
lm(formula = log_GDP_per_capita ~ log_GDP_per_capita_t_minus_1 +
    Trade + FBCF + log_CO2_per_capita + residu, data = our_don)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067320 -0.008361  0.002922  0.012498  0.036554

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                 0.3002456  0.3428797   0.876    0.386
log_GDP_per_capita_t_minus_1  0.9751158  0.0307261  31.736   <2e-16 ***
Trade                       -0.0002229  0.0006779  -0.329    0.744
FBCF                         0.0004105  0.0020286   0.202    0.841
log_CO2_per_capita          -0.0171325  0.0493195  -0.347    0.730
residu                      -0.0075878  0.1300626  -0.058    0.954
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02173 on 42 degrees of freedom
Multiple R-squared:  0.993, Adjusted R-squared:  0.9922
F-statistic:  1199 on 5 and 42 DF,  p-value: < 2.2e-16
```

Après avoir effectué le test d'Hausman, nous constatons que les résidus ne sont pas significatifs, indiquant l'absence d'endogénéité. Cela suggère que notre modèle initial présentait une endogénéité, mais celle-ci a été corrigée suite à l'introduction des variables instrumentales. Nous pouvons alors estimer les coefficients à l'aide de l'estimateur des variables instrumentales.

5 Partie 2: Analyse de de la multicolinéarité

Dans cette deuxième partie, nous abordons l'analyse de la multicollinéarité dans nos données. Pour répondre aux objectifs pédagogiques de ce projet, nous introduisons plusieurs nouvelles variables dans notre modèle, ce qui entraîne l'évolution de notre modèle comme suit:

$$log.GDP.per.capita = b_0 + b_1 * FBCF + b_2 * log.CO2.per.capita + b_3 * log.GDP.per.capita_{t-1} +$$
$$b_4 * Trade + b_5 * Trend + b_6 * POP_g + b_7 * Dth_rate +$$
$$b_8 * Pop_65 + b_9 * HTan + b_{10} * UNPrate + b_{11} * Hth.pc.gdp + b_{11} * Bth_rate + e$$

Avec en plus:

POP__g: Taux de croissance de la pop annuelle en % de la population

Dth__rate: Taux de mortalité par 1000 personnes

Pop__65: Population de 65 ans ou plus en % de population

Bth__rate: Taux de natalité par 1000 personnes

HT__an: Heure moy de travail annuelle par personne employée

UNP__rate: Taux de chômage en % de la population active

Hth__pc__gdp: Dépense dans la santé en % du PIB

5.1 Détection de la multicolinéarité

Lorsque nous estimons les coefficients du modèle à l'aide de la méthode des moindres carrés ordinaires (MCO), nous présentons le modèle de la manière suivante :

```
##
## Call:
## lm(formula = log_GDP_per_capita ~ ., data = don2)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0268914 -0.0058529 -0.0002735  0.0075627  0.0162585
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.236e+00  1.011e+00   4.189 0.000160 ***
## Trend                        2.484e-02  8.731e-03   2.845 0.007119 **
## Trade                       -7.200e-04  5.781e-04  -1.245 0.220644
## FBCF                         1.540e-03  1.977e-03   0.779 0.440931
## POP_g                       -1.404e-02  2.092e-02  -0.671 0.506262
## Dth_rate                    -3.662e-03  1.638e-02  -0.224 0.824235
## Pop_65                      -1.988e-02  1.044e-02  -1.904 0.064546 .
## Bth_rate                    -7.319e-03  6.573e-03  -1.114 0.272452
## HT_an                        4.207e-04  6.616e-05   6.359 1.84e-07 ***
## UNP_rate                    -2.842e-03  6.593e-03  -0.431 0.668830
## Hth_pc_gdp                  -2.094e-02  5.061e-03  -4.137 0.000188 ***
## log_CO2_per_capita           2.099e-01  4.002e-02   5.246 6.14e-06 ***
## log_GDP_per_capita_t_minus_1 4.540e-01  9.962e-02   4.557 5.24e-05 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01162 on 38 degrees of freedom
##   (1 observation effacée parce que manquante)
## Multiple R-squared:  0.9987, Adjusted R-squared:  0.9982
## F-statistic:  2363 on 12 and 38 DF,  p-value: < 2.2e-16
```

Nous observons un (R^2) élevé, mais les variables issues des théories économiques ne sont pas significatives. Nous en déduisons qu'il existe probablement un problème de multicolinéarité.

Afin de confirmer nos soupçons concernant la présence de multicolinéarité dans le modèle, nous examinons le facteur d'inflation de la variance (VIF). Le VIF mesure de combien la variance d'un coefficient est augmentée en raison d'une relation linéaire avec les autres régresseurs.

```
##                       Trend                       Trade
##                 6234.738870                    5.373750
##                        FBCF                       POP_g
##                   29.814924                   44.050482
##                   Dth_rate                      Pop_65
##                  294.878259                 2138.598813
##                   Bth_rate                       HT_an
##                  189.889560                   55.734206
##                   UNP_rate                  Hth_pc_gdp
##                   21.368109                   40.903585
##          log_CO2_per_capita log_GDP_per_capita_t_minus_1
##                    5.653755                  305.296570
```

La valeur au-dessus de laquelle nous considérons qu'il y a de la multicollinéarité n'est pas fixe. Nous prendrons donc 5 comme valeur de référence.

Nous observons alors un VIF très élevé dans notre modèle pour les variables ajoutées. Cela suggère une forte multicollinéarité entre les variables explicatives. En d'autres termes, ces variables sont fortement corrélées les unes avec les autres, ce qui peut poser des problèmes lors de l'estimation des coefficients et de l'interprétation des résultats. Pour résoudre ce problème, nous allons réduire la dimension de notre modèle afin d'éliminer la multicolinéarité.

5.2 Méthodes de réduction de dimension

5.2.1 Regression sur Composantes Principales (PCR)

Une synthèse des résultats de l'estimation par la méthode PCR nous donne :

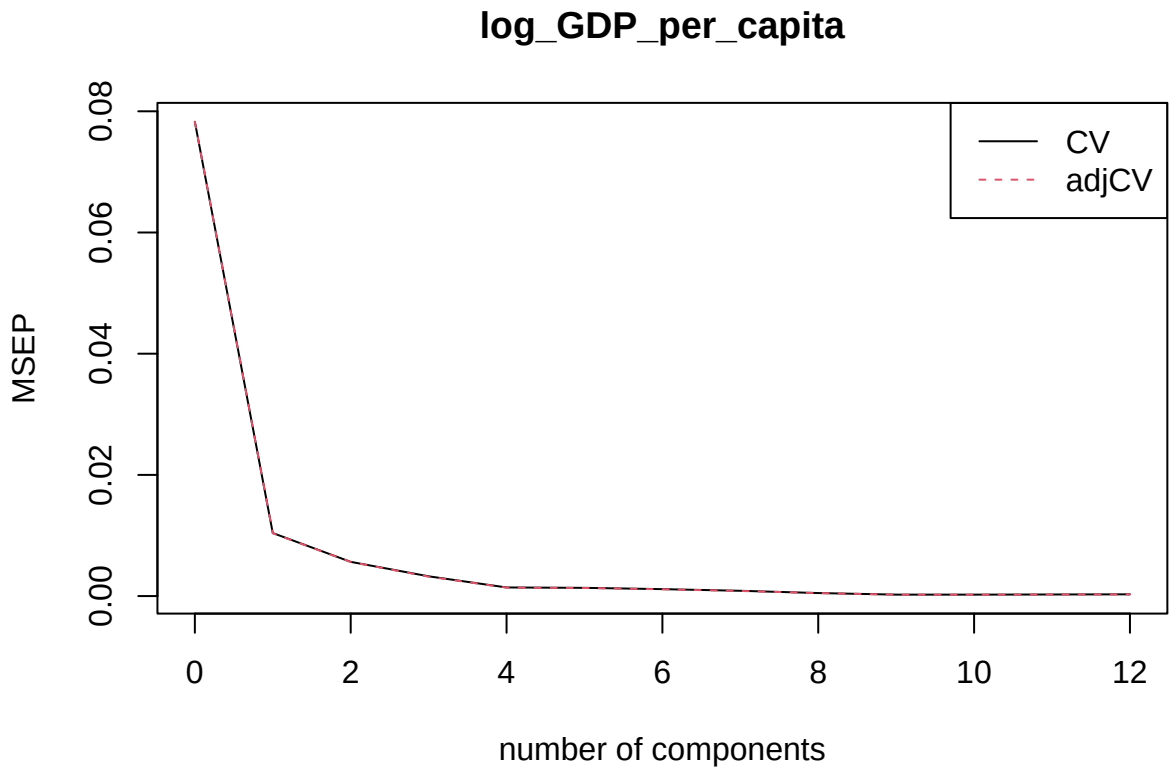
```
Data:   X dimension: 51 12
    Y dimension: 51 1
Fit method: svdpc
Number of components considered: 12

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.2798   0.1020  0.07515  0.05712  0.03772  0.03683  0.03383
adjCV       0.2798   0.1018  0.07499  0.05657  0.03737  0.03654  0.03348
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
```

```
CV     0.02954  0.02228  0.01543   0.01548   0.01613   0.01691
adjCV  0.02898  0.02222  0.01521   0.01521   0.01580   0.01650

TRAINING: % variance explained
                    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X                     78.71    89.86    94.00    97.49    98.86    99.35
log_GDP_per_capita    87.30    93.39    96.69    98.50    98.75    99.14
                    7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
X                     99.65    99.83    99.93     99.97    100.00    100.00
log_GDP_per_capita    99.38    99.61    99.82     99.84     99.85     99.87
```

On constate ainsi que 99,14 % de la variance du PIB par habitant est expliquée par les 6 premières composantes, avec une variation minime par la suite.

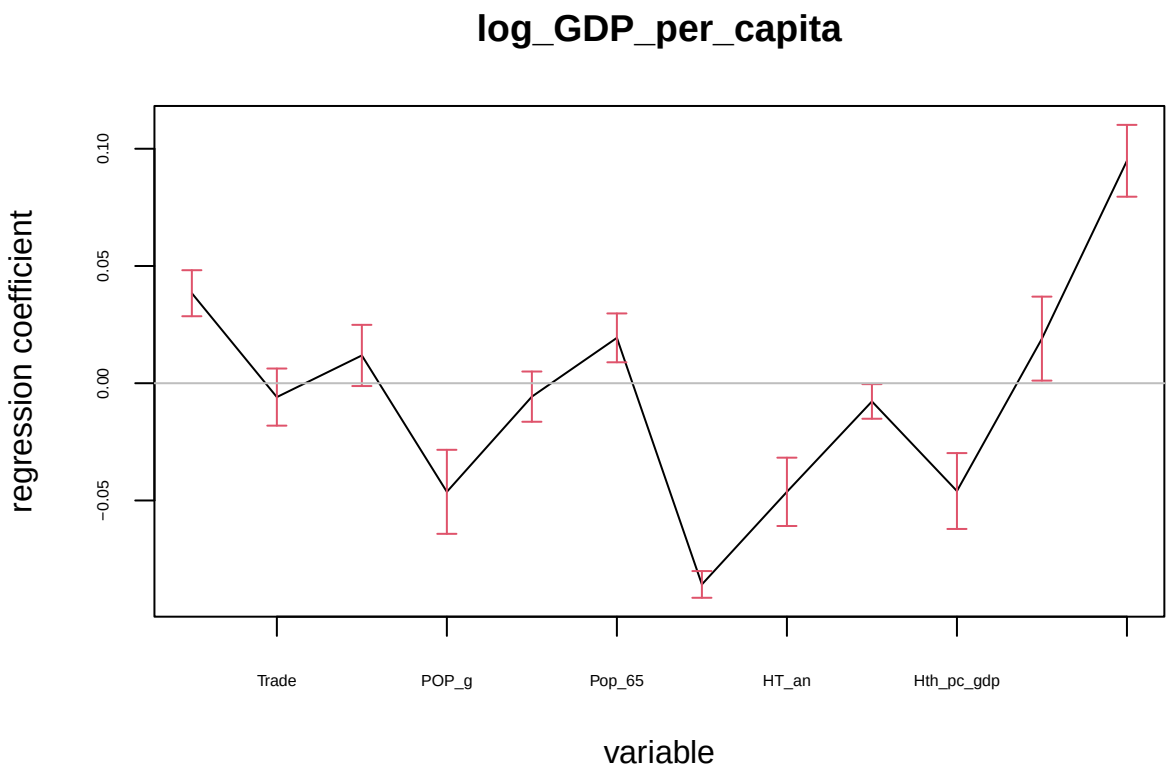
Les modèles comportant un nombre plus élevé de composantes risquent de souffrir de sur-ajustement, c'est-à-dire qu'ils peuvent devenir trop complexes pour les données disponibles, ce qui pourrait entraîner des prédictions moins précises sur de nouveaux ensembles de données. Il est donc préférable de choisir un modèle avec le moins de composantes possible tout en maximisant l'explication de notre variable d'intérêt. Cela nous conduit à rechercher le modèle présentant la plus petite erreur de validation croisée tout en restant moins complexe.

Pour aller plus loin, nous examinons l'erreur quadratique moyenne pour chaque nombre potentiel de composantes principales incluses dans le modèle.



On remarque ainsi l'erreur diminue très faiblement à partir de la quatrième composante, voir reste presque inchangé dès la sixième composante. On peut donc déduire qu'on choisit le modèle avec **6 composantes**, qui a la meilleure performance avec une valeur de 0.03383 pour CV et 0.03348 pour adjCV.

On estime alors les coefficients:



Les coefficients pour les variables *Trade*, *Pop_g*, *Dth_rate*, *Bth_rate*, *HT_an*, *Hth_pc_gdp* et *UNP_rate* sont tous négatifs, ce qui suggère que des valeurs plus élevées de ces variables sont associées à des valeurs plus faibles du PIB par habitant chaque année. De même, les coefficients positifs pour la variable *Trend*, *FBCF*, *Pop_65*, *log_CO2_per_capita*, et *log_GDP_per_capita_t_minus_1* suggèrent une association positive avec la variable dépendante.

Toutefois, ces coefficients ne reflètent pas l'effet direct de chaque variable sur le PIB par habitant.

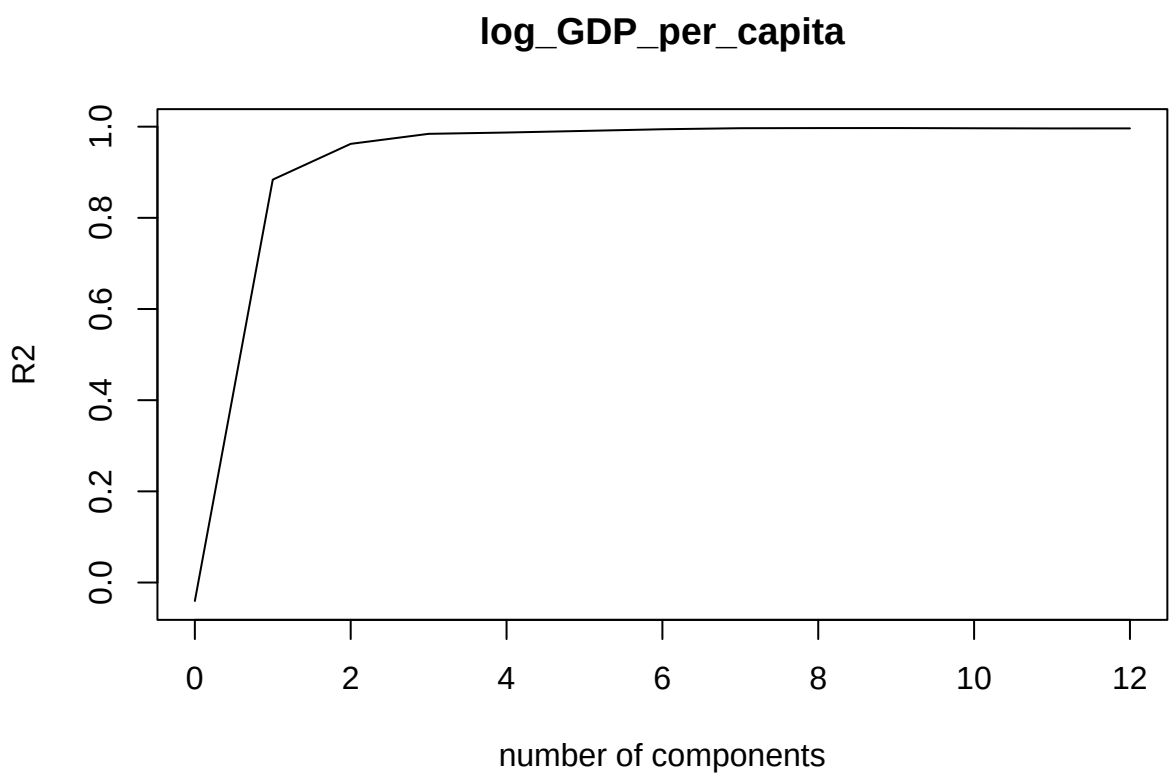
5.2.2 Régression des Moindres Carrés Partiel (PLS)

```
## Data:    X dimension: 51 12
##  Y dimension: 51 1
## Fit method: kernelpls
## Number of components considered: 12
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV          0.2798  0.09349  0.05328  0.03440  0.03111  0.02643  0.02058
## adjCV       0.2798  0.09326  0.05293  0.03402  0.03076  0.02606  0.02023
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## CV     0.01596  0.01520  0.01512   0.01615   0.01695   0.01691
## adjCV  0.01577  0.01498  0.01489   0.01582   0.01659   0.01650
##
## TRAINING: % variance explained
```

```
##                    1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## X                    78.61    89.04    93.75    95.95    98.22    99.22
## log_GDP_per_capita   89.43    96.92    98.88    99.27    99.53    99.71
##                    7 comps  8 comps  9 comps  10 comps  11 comps  12 comps
## X                    99.53    99.76    99.92     99.97    100.00    100.00
## log_GDP_per_capita   99.81    99.84    99.84     99.85     99.85     99.87
```

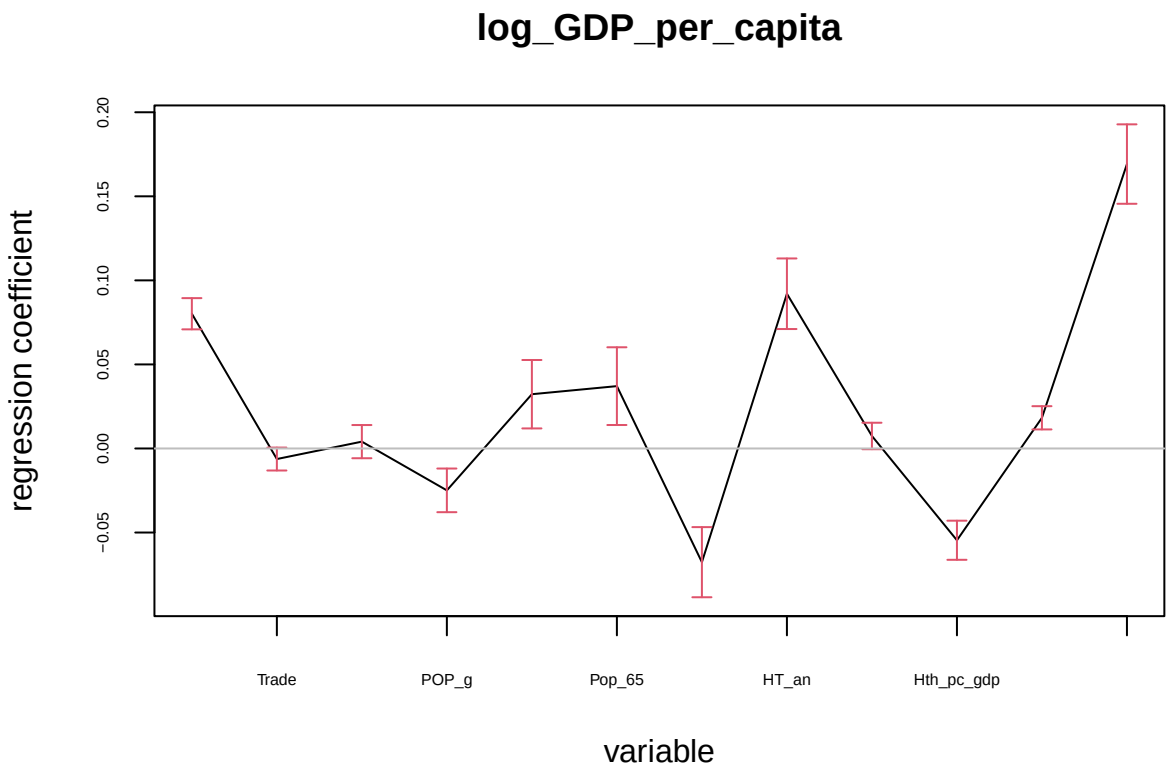
On peut observer que l'erreur diminue progressivement avec l'augmentation du nombre de composantes, mais le taux de décroissance ralentit à partir de 7 composantes. La meilleure performance est obtenue avec 9 composantes, où l'erreur de validation croisée est de 0.01512.

Pour aller plus loin, nous examinons le R^2 pour chaque nombre potentiel de composantes principales incluses dans le modèle.



On peut ainsi remarquer qu'il est maximum et commence à rester constant dès la huitième composantes.

On estime alors les coefficients:



En ce qui concerne l'interprétation des coefficients, chaque coefficient représente uniquement la relation entre la variable explicative et celle de réponse. Par exemple, un coefficient négatif pour la variable *Trade* indique une relation inverse entre cette variable et la variable de réponse, c'est-à-dire que dans le cadre du Japon,un taux d'ouverture commerciale élévé a tendance à réduire le PIB par habitant. De même, un coefficient positif pour la variable *FBCF* indique une relation directe entre cette variable et la variable de réponse, c'est-à-dire qu'un taux de Formation brute de capital fixe élévé a tendance à avoir une valeur plus élevée pour sur le PIB par habitant.

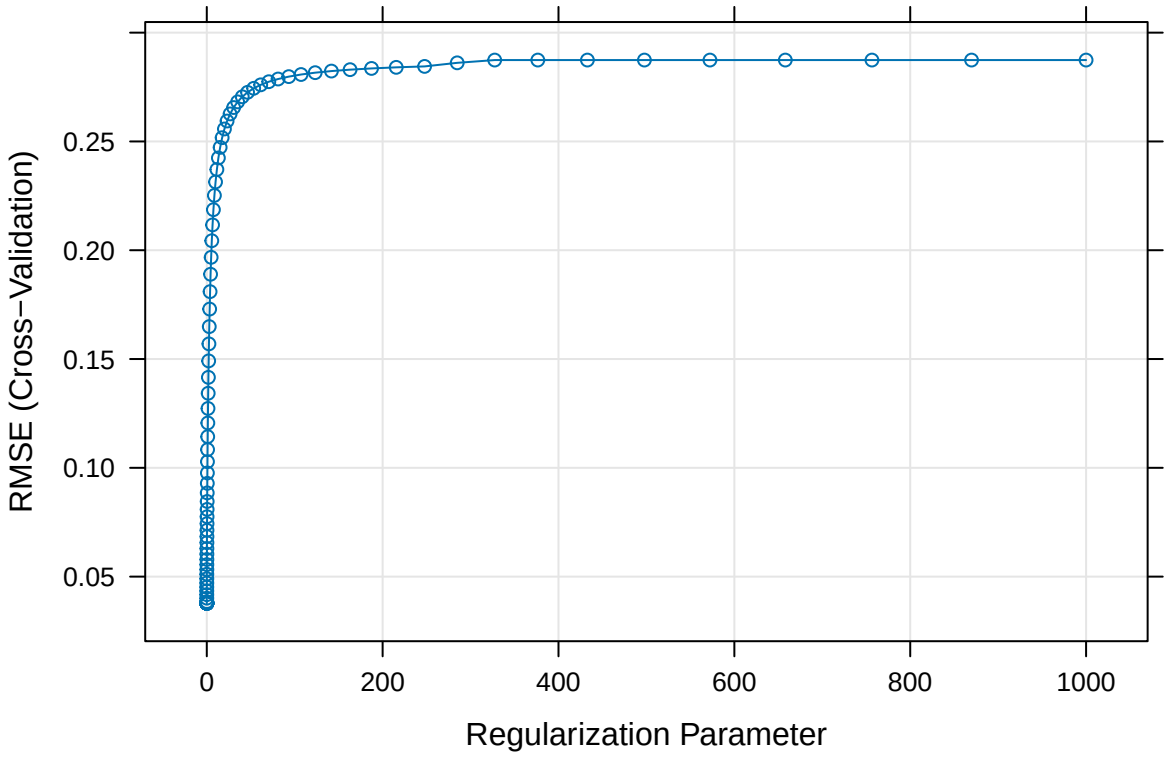
5.3 Méthodes pénalisées

Les méthodes de pénalisation sont des techniques visant à régulariser notre modèle linéaire et à réduire le risque de surajustement (overfitting). Nous utilisons le package glmnet pour mettre en œuvre ces méthodes.

5.3.1 Création des bases apprentisage et test

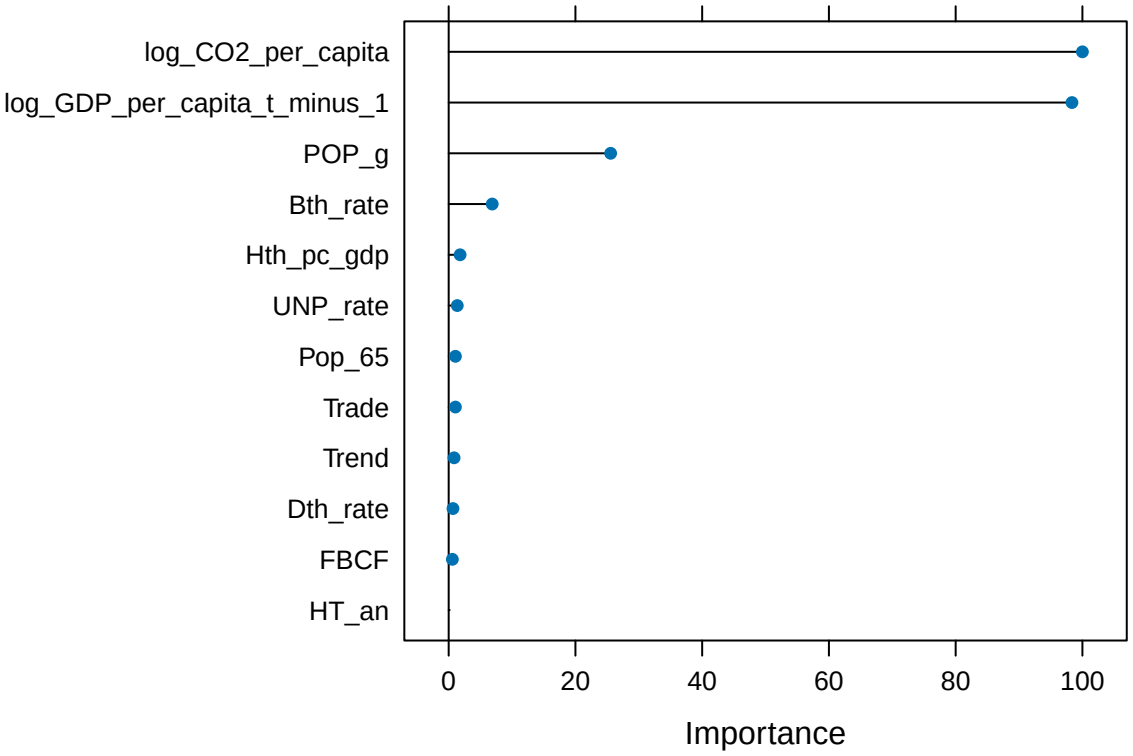
Nous construirons notre modèle sur les données d'entraînement et évaluerons ses performances sur les données de test. Notre échantillon d'apprentissage contient 50 % des données, tandis que l'échantillon de test contient les 50 % restants.

5.3.2 Régression Ridge



On constate ainsi que la valeur du meilleur paramètre lambda, qui minimise l'erreur quadratique moyenne estimée par validation croisée, est **0.02477076**.

On présente également l'importance des variables, et on a:



Ainsi, on peut constater que dans le contexte du Japon, le PIB par habitant est principalement expliqué par le PIB par habitant décalé d'un an, les émissions de CO2 par habitant et le taux de croissance de la population.

Cette observation est complétée par l'estimation du meilleur modèle, qui est le suivant :

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                 7.067180e+00
## Trend                       2.482040e-03
## Trade                      -3.129246e-03
## FBCF                       -1.603430e-03
## POP_g                      -7.299939e-02
## Dth_rate                    2.108488e-03
## Pop_65                      3.264395e-03
## Bth_rate                   -2.041859e-02
## HT_an                      -6.101912e-05
## UNP_rate                    3.777657e-03
## Hth_pc_gdp                 -5.815146e-03
## log_CO2_per_capita          2.888446e-01
## log_GDP_per_capita_t_minus_1  2.849834e-01
```

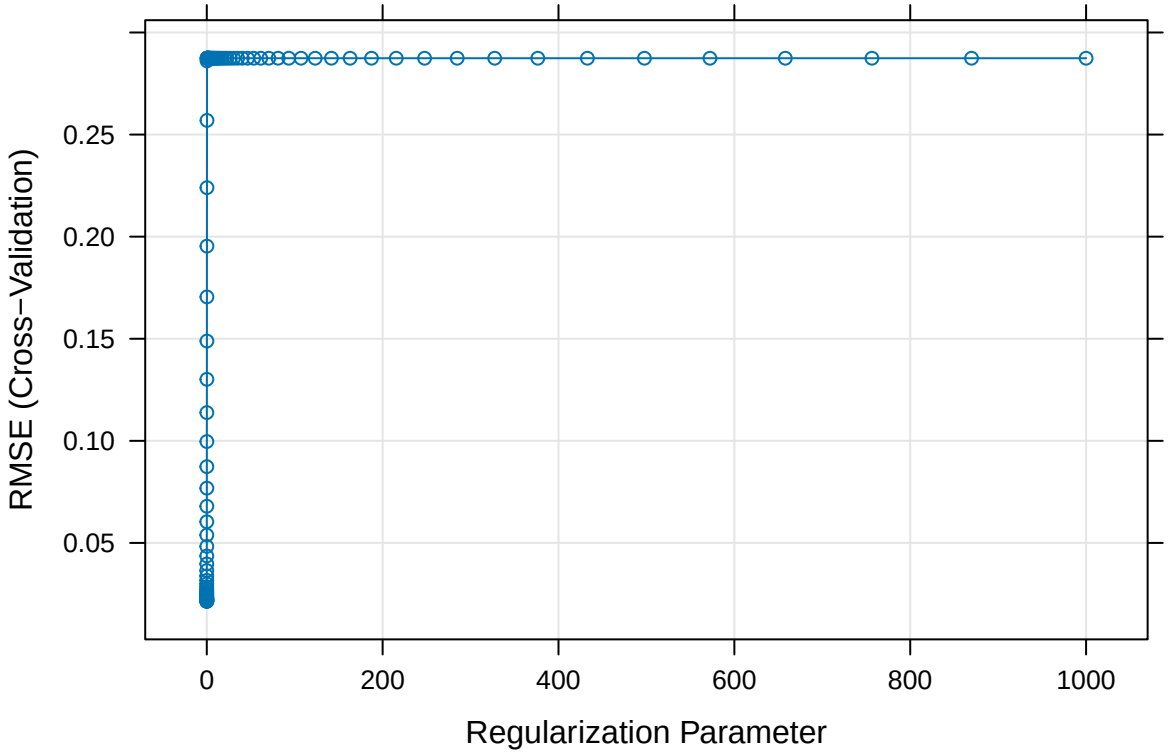
On conclut ici qu'un coefficient négatif indique une *relation inverse* avec la variable dépendante, tandis qu'un coefficient positif indique une *relation directe* avec la variable dépendante. Plus la valeur absolue d'un coefficient est élevée, plus grande est l'importance de la variable correspondante pour la prédiction de la variable dépendante, un constat conforme à l'analyse précédente.

On remarque que, en raison de l'utilisation de la régularisation dans le modèle de Ridge, certains de nos coefficients sont très petits (POP_g et HT_an) par rapport aux autres. Cela est dû à la pénalisation de la magnitude des coefficients, qui est utilisée pour éviter le **surajustement** et améliorer le modèle.

```
##         RMSE   Rsquare
## 1 0.03116389 0.9842132
```

Le coefficient de détermination (R-carré) s'élève à 0.9842132, avec une erreur de prévision de 0.03116389. Cela signifie que le meilleur modèle a pu expliquer 98,4 % de la variation des valeurs de réponse des données d'entraînement, avec seulement 3 % de chance d'erreur.

5.4 Régression Lasso



On constate que la valeur du paramètre lambda qui minimise l'erreur quadratique moyenne estimée par validation croisée est **0.00231013**.

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                       s0
## (Intercept)                  2.485666388
## Trend                        .
## Trade                       -0.000354545
## FBCF                         .
## POP_g                        .
## Dth_rate                     .
## Pop_65                       .
```

```
## Bth_rate                      -0.015700857
## HT_an                          .
## UNP_rate                       .
## Hth_pc_gdp                     .
## log_CO2_per_capita             0.128679631
## log_GDP_per_capita_t_minus_1   0.748675990
```

On remarque qu'avec l'utilisation d'une régression de Lasso Hitters, connu aussi pour sélectionner les variables pertinentes pour le modèle, certains de nos coefficients sont très petits (*Trade*) par rapport à d'autres. Les variables *FTrend*, *FBCF*,*POP_g* ,*Dth_rate*,*Pop_65*,*HT_an*,*UNP_rate* et *Hth_pc_gdp*, a quant à elles, ont été retirée du modèle. Cela est dû à la pénalisation de la magnitude des coefficients qui est utilisée pour éviter le **surajustement** et améliorer le modèle.

```
##         RMSE   Rsquare
## 1 0.01920281 0.9946675
```

Le coefficient de détermination (R-carré) s'élève à 0.9946675, avec une erreur de prévision de 0.01920281. Cela indique que le PIB par habitant au Japon, dans le cadre de notre étude, peut s'expliquer uniquement par l'ouverture commerciale (**Trade**), le taux de natalité (*Bth_rate*), le CO2 par habitant (*log_CO2_per_capita*) et le PIB par habitant décalé d'une année.

5.5 Régression Elastic Net

La régression Elastic Net combine deux formes de pénalisation précédente, Ridge et Lasso.

On remarque que les meilleurs alpha et lambda estimés sur les données d'entrainement sont respectivemets égaux à 0.2 et 0.000744449.

```
## 13 x 1 sparse Matrix of class "dgCMatrix"
##                                       s1
## (Intercept)                   3.1600655427
## Trend                         0.0035574535
## Trade                        -0.0012951520
## FBCF                          0.0008788450
## POP_g                        -0.0404414235
## Dth_rate                      0.0160744819
## Pop_65                        0.0033844474
## Bth_rate                     -0.0201570049
## HT_an                         0.0003407347
## UNP_rate                      0.0064907194
## Hth_pc_gdp                   -0.0176620830
## log_CO2_per_capita            0.1979239058
## log_GDP_per_capita_t_minus_1  0.5930769767
```

Avec l'application d'une régression Elastic Net, toutes les variables sont conservées dans le modèle final. Cependant, les variables *Trade*, *HT_an* et *FBCF* ont des coefficients plus petits que les autres variables.

```
##         RMSE   Rsquare
## 1 0.01338051 0.9973782
```

Le R-carré vaut 0.9973782. Donc le meilleur modèle a été en mesure d'expliquer 99,73782 % de la variation des valeurs de réponse des données d'entraînement.

5.6 Comparaison des resultats

Afin de sélectionner le meilleur modèle nous étudions le tableau récapitulatif suivant :

```
##
## 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   0.022371442 0.02884145 0.03902211 0.03779278 0.04376880 0.05909730    0
## lasso   0.006508557 0.01236792 0.01587623 0.02138269 0.03000851 0.04745894    0
## elastic 0.006052574 0.01172078 0.01448392 0.01875094 0.02566563 0.04222110    0
```

D'après les différents indicateurs, la régression Elastic Net semble être la meilleure option, car elle minimise la plage de valeurs de la RMSE, quel que soit le paramétrage. En conséquence, le modèle final retenu est celui généré par la régression Elastic Net

6 Partie 3: Causalité et double machine learning

6.1 Mise à jour de la base et estimation par double machine learning simple

Afin de mettre en oeuvre cette partie, on introduit dans nos données une variables relatif à l'année 2008, afin d'évaluer les effets de la crise économique de 2008 sur le PIB par habitant au JAPON.Cet effet, mesurer par une variable qu'on appelle ici **dummy_2008**, prendra pour valeur 1 après 2008 et 0 avant.

Ensuite, nous estimons les paramètres à l'aide de la méthode Lasso avec glmnet. Le paramètre de pénalité lambda est sélectionné via une validation croisée.

```
## INFO  [06:41:33.354] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 1/5)
## INFO  [06:41:33.454] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 2/5)
## INFO  [06:41:33.511] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 3/5)
## INFO  [06:41:33.565] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 4/5)
## INFO  [06:41:33.618] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_l' (iter 5/5)
## INFO  [06:41:33.866] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 1/5)
## INFO  [06:41:33.929] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 2/5)
## INFO  [06:41:34.002] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 3/5)
## INFO  [06:41:34.068] [mlr3] Applying learner 'regr.cv_glmnet' on task 'nuis_m' (iter 4/5)
## INFO  [06:41:34.157] [mlr3] Applying learner 'regr.cv_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|)
## dummy_2008  -0.01063    0.02377  -0.447    0.655
```

Ainsi, le coefficient associé à la variable **dummy_2008** ne semble pas significatif. Pour renforcer notre conclusion, nous allons estimer l'intervalle de confiance et les estimations par bootstrap.

6.2 Estimation par double machine learning avec bootstrapping

```
##                   2.5 %     97.5 %
## dummy_2008 -0.05709425 0.03583778
```

On conclut alors l'intervalle de confiance pour *dummy_2008* inclut zéro, cela indique que le coefficient pour cette variable n'est pas significatif.

On applique ensuite une *correction des p-values* selon deux méthodes, utiles car permet de réduire le risque de fausse découverte. En effet, lorsque plusieurs tests d'hypothèses sont effectués simultanément, le risque d'observer au moins une fausse découverte augmente. La correction des valeurs p vise à ajuster ces valeurs en fonction du nombre de tests effectués, afin de maintenir un niveau global de significativité approprié.

```
##              Estimate.  pval
## dummy_2008 -0.01062824 0.636
```

```
##              Estimate.      pval
## dummy_2008 -0.01062824 0.6547601
```

Ainsi, quelle que soit la méthode utilisée, la variable *dummy_2008* n'est pas significative, et on en déduit que la crise économique de 2008 n'a pas eu d'effet sur le PIB par habitant au Japon à long terme.

7 Conclusion

Pour conclure, dans cette étude sur le PIB par habitant au JAPON, nous avons d'abord analysé notre base de données de manière univariée et bivariée pour orienter au mieux la suite de notre réflexion. Ensuite, nous avons sélectionné notre modèle économétrique fondé sur des théories économiques. Après avoir testé l'absence d'autocorrélation dans nos données, ainsi qu'un test d'endogénéité, la présence d'endogénéité nous a amenés à décider d'appliquer l'estimateur des variables instrumentales.

Dans le cadre pédagogique de ce projet, nous avons introduit plusieurs autres variables dans notre modèle, puis étudié la multicolinéarité. Pour cela, nous avons appliqué des méthodes de réduction de dimension et de pénalisation. Parmi celles-ci, le modèle optimal est celui proposé par *elastic net*. Ce modèle est défini par :

$$log.GDP.per.capita = 3.16 + 0.00088 * FBCF + 0.197 * log.CO2.per.capita + 0.593 * log.GDP.per.capita_{t-1} +$$

$$-0.00129 * Trade + 0.0035 * Trend + -0.0404 * POP_g + 0.016 * Dth_rate +$$

$$0.00338 * Pop_65 + 0.00034 * HTan + 0.00649 * UNPrate + -0.017 * Hth.pc.gdp + -0.020 * Bth_rate + e$$

Enfin, nous avons examiné l'impact de la crise économique de 2008 sur le PIB par habitant au JAPON en utilisant une estimation par le Double Machine Learning. Nos résultats indiquent que cette crise n'a eu aucun effet significatif sur le PIB par habitant au JAPON, ou en d'autres termes, nous ne pouvons pas conclure que cette crise a eu des effets à long terme sur le PIB par habitant au JAPON.

8 Bibliographie

Données:

- OCDE
- World bank data

Références:

- Isabelle Cadoret, **Econométrie avancée : causalité, Lasso, Ridge**, Mastère Mathématiques Appliquées Statistique, Université de Rennes.
- Gareth James, & Daniela Witten, &Trevor Hastie, & Robert Tibshirani, **An Introduction to Statistical Learning with Applications in R**

9 Contribution au projet

Marc TANO : Tout le projet, à l'exception des méthodes de réduction de dimension PLS.

Gyldano Dadjedji : Tout le projet, sauf pour les méthodes pénalisées ridge.

Kwami Nouchet : Tout le projet, excluant les méthodes pénalisées elastic net.