L’impact des institutions sur la performance économique, un héritage de la colonisation européenne

Gregoire FUCHS - Mehdi BENAILLY - Hugo WEBER

1. Introduction

En nous appuyant sur un article d’Acemoglu, Johnson et Robison de 2001, nous allons étudier l’une des explications données à cette différence de richesse produite. En effet dans l’article, mentionné précédemment, les auteurs cherchent à démontrer que le cadre institutionnel est un facteur jouant un rôle clé dans le développement économique d’une région. Ils cherchent donc à établir un lien entre PIB/hab et niveau de protection de la propriété privée.

En nous basant sur cet article nous avons essayé de refaire le travail des auteurs pour chercher à établir nous-même le lien entre pib/hab et institution.

Dans le cadre de notre présentation et pour montrer le fruit de notre travail, nous allons commencer par présenter l’article qui nous a servi de référence, ses bases de données que nous avons repris, puis nous présenteront les résultats et le code, et enfin ses limites qui furent soulevés par les auteur eux-mêmes mais également par d’autres auteurs entrant parfois en opposition avec cet article.

2. Présentation de l’article

Comme mentionné précédemment les auteurs cherchent à démontrer le lien entre niveau de protection de la propriété privée et niveau de développement économique. En raison d’endogénéité, ce qui sera détaillé plus tard, il a fallu également trouver une variable instrumentale pertinente. Pour cela nous allons devoir nous plonger dans le passé colonial européen qui fut très important dans le cadre de mise en place des institutions toujours en place aujourd’hui. « Si chaque territoire se différencie par sa population son climat, son histoire et sa géographie qui lui est propre, et que l’ont constate que chaque colonie n’a pas été administré de la même manière que ses « sœurs » même au sein d’un même empire colonial il convient de se demander ce qui a justifié le choix d’une institution sur une autre par les européens.

Les auteurs ont décidé de se baser sur la mortalité des colons européens dans le territoire colonisé comme facteur déterminant le cadre institutionnel tel que plus un territoire est considéré comme inadapté à la survie des colons moins ses institutions protègeront la propriété privée vu comme essentielle au développement économique. »

Raison pour laquelle on étudie le taux de mortalité des colons:

En raison de l’absence de données fiable, ils sont rabattus sur les données des troupes d’occupations des territoires ainsi que les chiffres des missionnaires chrétiens envoyé depuis l’Europe comme indicateur de la mortalité des colons.

L’article semble conclure sur les résultats suivants : Plus les colons meurent moins les institutions retenues sont orientées vers une protection de la propriété. Et moins la protection de la propriété privée sera garantie par les institutions moins le territoire se développera.

3. Résultats et code

La première étape va être d’importer les fichiers issus du logiciel propriétaire de stat “.dta” du MIT d’Acemoglu (Nobel 2024) Le lien vers la base de données est https://economics.mit.edu/people/faculty/daron-acemoglu/data-archive (https://economics.mit.edu/people/faculty/daron-acemoglu/data-archive).

Les données sont donc en lien avec l’article d’Acemoglu de 2001 AER paper (AJR) sur le lien des institutions vers le développement économique. AJR fournit plusieurs ensembles de données sur sa page de réplication dans laquelle nous avons placé le répertoire « données ». Chaque ensemble de données nous permet de construire des estimations à partir de l’un des tableaux de l’article.

```
# Chargement des packages
library(tidyverse)
library(haven)
library(stargazer)
library(ggplot2)
library(car)
library(sandwich)
library(lmtest)
library(dplyr)
library(broom)
library(plm)
library(AER)
library(mgcv)
library(pscl)
library(pROC)
library(MLmetrics)
library(ResourceSelection)
```

```
# Importation des bases de données
maketable1 <- read_dta("~/Documents/data copie/maketable1.dta")
maketable2 <- read_dta("~/Documents/data copie/maketable2.dta")
maketable3 <- read_dta("~/Documents/data copie/maketable3.dta")
maketable4 <- read_dta("~/Documents/data copie/maketable4.dta")
maketable5 <- read_dta("~/Documents/data copie/maketable5.dta")
maketable6 <- read_dta("~/Documents/data copie/maketable6.dta")
maketable7 <- read_dta("~/Documents/data copie/maketable7.dta")
maketable8 <- read_dta("~/Documents/data copie/maketable8.dta")
```

Nous affichons les noms des colonnes pour connaître les variables présentes dans chaques tableaux “maketable”, afin de les sélectionner facilement par la suite.

```
names(maketable1)
```

```
##  [1] "shortnam" "euro1900" "avexpr"   "logpgp95" "cons1"    "democ00a"
##  [7] "cons00a"  "extmort4" "logem4"   "loghjypl" "baseco"
```

```
names(maketable2)
```

```
## [1] "shortnam" "africa"   "lat_abst" "avexpr"   "logpgp95" "other"    "asia"
## [8] "loghjypl" "baseco"
```

```
names(maketable3)
```

```
##  [1] "lat_abst" "euro1900" "excolony" "avexpr"   "logpgp95" "cons1"
##  [7] "indtime"  "democ00a" "cons00a"  "extmort4" "logem4"
```

```
names(maketable4)
```

```
##  [1] "shortnam" "africa"   "lat_abst" "rich4"    "avexpr"   "logpgp95"
##  [7] "logem4"   "asia"     "loghjypl" "baseco"
```

```
names(maketable5)
```

```
##  [1] "shortnam" "catho80"  "muslim80" "lat_abst" "no_cpm80" "f_brit"
##  [7] "f_french" "avexpr"   "sjlofr"   "logpgp95" "logem4"   "baseco"
```

```
names(maketable6)
```

```
##  [1] "shortnam" "avelf"    "lat_abst" "temp1"    "temp2"    "temp3"
##  [7] "temp4"    "temp5"    "humid1"   "humid2"   "humid3"   "humid4"
## [13] "steplow"  "deslow"   "stepmid"  "desmid"   "drystep"  "drywint"
## [19] "edes1975" "avexpr"   "logpgp95" "landlock" "goldm"    "iron"
## [25] "silv"     "zinc"     "oilres"   "logem4"   "baseco"
```

```
names(maketable7)
```

```
##  [1] "shortnam" "africa"   "lat_abst" "malfal94" "avexpr"   "logpgp95"
##  [7] "logem4"   "asia"     "yellow"   "baseco"   "leb95"    "imr95"
## [13] "meantemp" "lt100km"  "latabs"
```

```
names(maketable8)
```

```
##  [1] "shortnam" "lat_abst" "euro1900" "avexpr"   "logpgp95" "democ1"
##  [7] "cons1"    "indtime"  "democ00a" "cons00a"  "logem4"   "baseco"
```

3.1 Analyse des régressions linéaires

Dans cette première partie, nous analyserons des régressions linéaires d'abord simple puis multiple avec différentes variables de contrôle, pour étudier la relation entre institutions et développement économique.

3.1.1 Régression simple : institutions et PIB

La régression simple examine l'effet direct de l'indicateur des différences institutionnelles représentées par un indice de protection contre l'expropriation en moyenne sur la période 1985-1995 (avexpr) sur le logarithme du PIB par habitant en 1995 (log(PGDP95)).

Commençons d'abord par tracer un graphique pour avoir un premier aperçu de la relation entre les deux variable, en utilisant ggplot2.

```
ggplot(maketable1, aes(x=avexpr, y = logpgp95)) +
  geom_point() +
  labs(x = "Risque d'expropriation (Institutions)", y = "logarithme du PIB par habitant en 19
95") +
  ggtitle("Figure n°1: Relation entre le risque d'expropriation et le PIB") +
  theme_minimal()
```

Il semble donc qu'il y ait une relation positive entre une meilleure protection de la propriété privée (=institutions) et d'avoir un bon revenu.

Nous créons alors la régression linéaire et nous affichons les resultats grâce au package stargazer.

```
regsimple <- lm(logpgp95 ~ avexpr, data = maketable1)

stargazer(regsimple, type = "text",
          title = "Tableau de régression n°1",
          dep.var.labels = "log du PIB par habitant (logpgp95)",
          covariate.labels = "Institutions (avexpr)")
```

```
##
## Tableau de régression n°1
## ============================================================
##                               Dependent variable:
##                      ---------------------------------
##                      log du PIB par habitant (logpgp95)
## ------------------------------------------------------------
## Institutions (avexpr)              0.532***
##                                    (0.041)
##
## Constant                           4.626***
##                                    (0.301)
##
## ------------------------------------------------------------
## Observations                         111
## R2                                  0.611
## Adjusted R2                         0.608
## Residual Std. Error            0.718 (df = 109)
## F Statistic                171.438*** (df = 1; 109)
## ============================================================
## Note:                         *p<0.1; **p<0.05; ***p<0.01
```

Les résultats montrent une relation positive et significative entre les deux variables, avec une statistique de Fisher élevée de 171.4 et une p-value quasi nulle. Le coefficient estimé (avexpr) indique qu'une augmentation d'une unité de l'indice de protection contre l'expropriation est associée à une augmentation de 53,187% du PIB par habitant, toutes choses égales par ailleurs. On retrouve un R^2 élevé, 61,13% de la variabilité du logarithme du PIB par habitant en 1995 est expliquée par les institutions, ce qui confirme leur rôle central dans le développement économique. Des institutions solides favorisent l'investissement et la croissance économique. Des protections institutionnelles plus fortes réduisent l'incertitude liée aux risques d'expropriation.

Nous essayons ensuite de détecter d'éventuelles violation des hypothèses de Gauss Markov grâce au test de Breusch-Pagan.

```
# Installations de package lmtest
bp_test<-bptest(regsimple)
print(bp_test)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  regsimple
## BP = 2.6029, df = 1, p-value = 0.1067
```

Le test de Breusch-Pagan avec un résultat de 2,6029 et une p-value de 0.10 semble verifier l'hypothèse de Gauss Markov d'homoscédasticité.

Testons ensuite l'autocorrélation des résidus avec le test de Durbin-Watson.

```
dw_test <- dwtest(regsimple)
print(dw_test)
```

```
##
##  Durbin-Watson test
##
## data:  regsimple
## DW = 1.6891, p-value = 0.0483
## alternative hypothesis: true autocorrelation is greater than 0
```

En revanche, le test de Durbin-Watson égale 1.689 indique qu'il y a une autocorrélation positive des résidus car différent de 2, ce qui peut biaiser les écarts-types estimés car il y a une violation des hypothèses de Gauss-Markov.

Passons ensuite aux régressions multiples

3.1.2 Régressions multiples

3.1.2.1 Ajout de variables de contrôle : climat

Cette régression intègre la latitude absolue (lat_abst) comme proxy pour le climat, en plus des institutions (avexpr).

```
reg2 <- lm(logpgp95 ~ avexpr + lat_abst, data = maketable2)

stargazer(reg2, type = "text",
          title = "Tableau de régression n°2",
          dep.var.labels = "log du PIB par habitant",
          covariate.labels = c("institutions", "climat (latitude absolue)"))
```

```
##
## Tableau de régression n°2
## ================================================
##                              Dependent variable:
##                          -----------------------
##                            log du PIB par habitant
## ------------------------------------------------
## institutions                       0.463***
##                                     (0.055)
##
## climat (latitude absolue)           0.872*
##                                     (0.488)
##
## Constant                           4.873***
##                                     (0.328)
##
## ------------------------------------------------
## Observations                          111
## R2                                   0.623
## Adjusted R2                          0.616
## Residual Std. Error            0.711 (df = 108)
## F Statistic                 89.047*** (df = 2; 108)
## ================================================
## Note:                     *p<0.1; **p<0.05; ***p<0.01
```

Les résultats montrent que l'indicateur des différences institutionnelles reste significatif, avec un effet positif sur le PIB. L'ajout de la latitude dans le modèle vise à capturer l'effet potentiel des conditions climatiques sur le développement économique. Cependant, l'absence de significativité de cette variable pourrait indiquer que son effet est largement transmit par les institutions elles-mêmes. Ce sont donc les institutions, et non directement le climat, qui influencent positivement le PIB. On peut supposer que l'hypothèse de multicolinéarité est potentiellement violée.

On teste alors cela.

```
vif_values <- vif(reg2)
print(vif_values)
```

```
##   avexpr lat_abst
##  1.90409  1.90409
```

L'analyse des facteurs d'inflation de variance (VIF = 1.9) montre qu'il n'y a pas de problème de multicolinéarité parfaite car le VIF est inférieur à 5. Cela valide l'ajout de cette variable au modèle sans compromettre l'interprétation des coefficients.

Comparons nos deux premières regressions:

```
stargazer(regsimple, reg2, type = "text",
          title = "Tableau de régression n°3",
          dep.var.labels = "log du PIB par habitant en 1995",
          covariate.labels = c("institutions", "climat (latitude absolue)"))
```

```
##
## Tableau de régression n°3
## ================================================================================
##                                          Dependent variable:
##                              ---------------------------------------------------
##                                      log du PIB par habitant en 1995
##                                        (1)                         (2)
## --------------------------------------------------------------------------------
## institutions                          0.532***                  0.463***
##                                       (0.041)                    (0.055)
##
## climat (latitude absolue)                                        0.872*
##                                                                  (0.488)
##
## Constant                              4.626***                  4.873***
##                                       (0.301)                    (0.328)
##
## --------------------------------------------------------------------------------
## Observations                            111                        111
## R2                                     0.611                      0.623
## Adjusted R2                            0.608                      0.616
## Residual Std. Error               0.718 (df = 109)           0.711 (df = 108)
## F Statistic                  171.438*** (df = 1; 109) 89.047*** (df = 2; 108)
## ================================================================================
## Note:                                            *p<0.1; **p<0.05; ***p<0.01
```

En comparant nos deux premiers modèles, on observe que l'ajout de la latitude n'altère pas l'effet des institutions sur le PIB. Cette stabilité des coefficients souligne que l'impact des institutions est indépendant des effets climatiques et que l'ajout de la latitude ne contribue pas significativement à l'amélioration de l'ajustement global du modèle.

3.1.2.2 Ajout de variables de contrôle : situation géographique

Pour approfondir, nous ajoutons une autre variable de contrôle, la situation géographique de l'individu concerné.

```
reg3 = lm(logpgp95 ~ avexpr + lat_abst + asia + africa + other, data=maketable2)

stargazer(reg3, type = "text",
          title = "Tableau de régression n°4",
          dep.var.labels = "log du PIB par habitant en 1995",
          covariate.labels = c("institutions", "climat (latitude absolue)", "Asie","Afrique",
"Autres"))
```

```
##
## Tableau de régression n°4
## ===============================================================
##                                 Dependent variable:
##                             -----------------------------------
##                             log du PIB par habitant en 1995
## ---------------------------------------------------------------
## institutions                             0.390***
##                                           (0.051)
##
## climat (latitude absolue)                 0.333
##                                           (0.445)
##
## Asie                                      -0.153
##                                           (0.155)
##
## Afrique                                  -0.916***
##                                           (0.166)
##
## Autres                                    0.304
##                                           (0.375)
##
## Constant                                 5.851***
##                                           (0.340)
##
## ---------------------------------------------------------------
## Observations                                111
## R2                                         0.715
## Adjusted R2                                0.702
## Residual Std. Error                  0.626 (df = 105)
## F Statistic                       52.738*** (df = 5; 105)
## ===============================================================
## Note:                              *p<0.1; **p<0.05; ***p<0.01
```

Nous avons pu observer que l'indicateur des différences institutionnelles reste significatif et les variables géographiques (Afrique et Asie) impactent négativement le PIB.

Retestons si l'hypothèse d'homoscédasticité est valide.

```
bptest(reg3)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg3
## BP = 24.435, df = 5, p-value = 0.0001791
```

Le test de Breusch-Pagan (BP = 24.435) avec une P-value très faible 0.0001 révèle une hétéroscédasticité significative, et donc l'hypothèse d'homoscédasticité est alors clairement violée. Afin de corriger l'hétéroscédasticité, nous calculons alors des coefficients robustes à la White.

```
#On calcule les erreurs standards robustes grâce à lmtest et notamment en obtenant des coeffi
cients plus robustes
reg4 <- coeftest(reg3, vcov = vcovHC(reg3, type = "HC1"))

# Afficher les résultats avec erreurs standards robustes
print(reg4)
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  5.851084   0.293497 19.9358 < 2.2e-16 ***
## avexpr       0.389561   0.051005  7.6377 1.078e-11 ***
## lat_abst     0.332564   0.442440  0.7517   0.45394
## asia        -0.153063   0.180505 -0.8480   0.39838
## africa      -0.916386   0.154111 -5.9463 3.638e-08 ***
## other        0.303549   0.174329  1.7412   0.08457 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

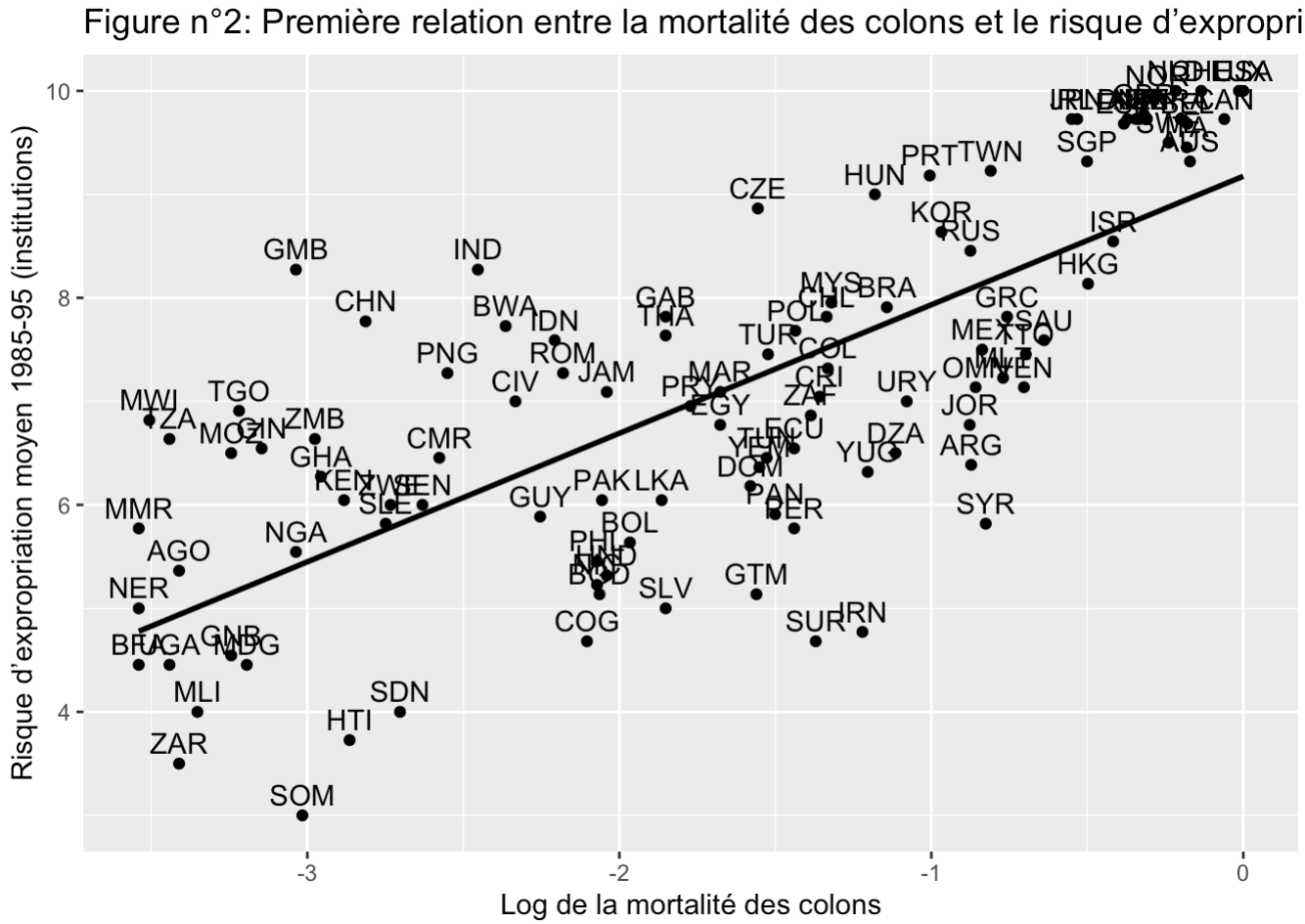
Ces coefficients robustes restent significatifs, ce qui renforce la robustesse des conclusions. On observe une modification des écarts types entre reg3 et reg4, donc il y a hétéroscédasticité.

3.1.3 Présence de causalités inverses

Une limitation de l'approche par régression ordinaire est la présence de causalités inverses. Les pays riches peuvent se permettre de meilleures institutions, ce qui biaise l'estimation de l'effet des institutions sur le PIB. De plus, certaines variables, comme la latitude, pourraient être corrélées avec à la fois les institutions et le PIB, et créer des problèmes d'endogénéité. Pour traiter cela nous utiliserons la méthode des variables instrumentales (présentées dans la partie suivante). Pour choisir la variable instrumentale, nous émettons l'hypothèse que les taux de mortalité plus élevés des colonisateurs ont conduit à la mise en place d'institutions de nature plus extractive (moins de protection contre l'expropriation), et ces institutions persistent encore aujourd'hui.

```
#Enlevons les valeurs manquantes avant
maketable2 <- maketable2 %>% drop_na(loghjypl,avexpr)

# Création du graphique avec ggplot2
ggplot(maketable2, aes(x = loghjypl, y = avexpr, label = shortnam)) +
  geom_point() +
  geom_text(vjust = -0.5) +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  labs(
    x = "Log de la mortalité des colons",
    y = "Risque d’expropriation moyen 1985-95 (institutions)",
    title = "Figure n°2: Première relation entre la mortalité des colons et le risque d’expro
priation"
  )
```



Ainsi grâce à ce premier graphique, nous pouvons voir que l’instrument de la mortalité des colons est pertinent : il est fortement corrélé au taux des colons.

3.2 Analyse de la régression instrumentale

Dans cette partie, nous analysons alors la régression instrumentale utilisée afin surmonter les problèmes d’endogénéité identifiés dans les modèles précédents. Cette méthode exploite la mortalité des colons (logem4) comme variable instrumentale pour les institutions (avexpr), permettant d’estimer l’effet causal des institutions sur le PIB.

3.2.1 Régression instrumentale

La première étape de la régression instrumentale consiste à régresser l'indicateur des institutions comme variable endogène (avexpr) sur la mortalité des colons comme variable instrumentale (logem4) afin de vérifier si l'instrument est pertinent et s'il explique significativement la variation des institutions.

```
reg1S <- lm(avexpr ~ logem4, data = maketable4)

stargazer(reg1S, type = "text",
          title = "Tableau de régression n°5",
          dep.var.labels = "indicateur des institutions",
          covariate.labels = "mortalité des colons")
```

```
##
## Tableau de régression n°5
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                      indicateur des institutions
## ------------------------------------------------
## mortalité des colons          -0.647***
##                                (0.115)
##
## Constant                      9.528***
##                                (0.548)
##
## ------------------------------------------------
## Observations                     74
## R2                              0.304
## Adjusted R2                     0.295
## Residual Std. Error        1.321 (df = 72)
## F Statistic             31.513*** (df = 1; 72)
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

Le coefficient de mortalité des colons (logem4) est significatif (p-value < 0.01) et la statistique de Fisher egal à 31.51 dépasse largement le seuil critique de 10, cela confirme bien que la mortalité des colons est pertinent pour prédire les institutions. Ces résultats sont cohérents avec l'hypothèse selon laquelle des taux de mortalité élevés ont conduit à des institutions plus extractives (moins protectrices contre l'expropriation). En effet on retrouve une forte relation négative entre mortalité des colons et qualité des institutions, une augmentation d'une unité de la mortalité des colons entraine une diminution de 0.6468 unité de l'indicateur des différences institutionnelles, toutes choses égales par ailleurs. Soit une détérioration des institutions.

Ensuite nous récupérerons la variable prédite de la première régression pour réaliser notre régression comparable au modèle OLS simple.

```
# Prédire 'avexpr' en utilisant le modèle de la première étape.
maketable4$predicted_avexpr <- predict(reg1S, newdata = maketable4)
```

3.2.2 Variable prédite

Ensuite, nous régressons le logarithme du PIB par habitant (log(PGDP95) sur la variable prédite (predicted_avexpr) issues de la première étape de la régression instrumentale.

```
reg2s <- lm(logpgp95 ~ predicted_avexpr, data = maketable4)
summary(reg2s)
```

```
##
## Call:
## lm(formula = logpgp95 ~ predicted_avexpr, data = maketable4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.71304 -0.53326  0.01954  0.47188  1.44673
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.35024    0.65309   3.599 0.000556 ***
## predicted_avexpr  0.87221    0.09878   8.830 2.09e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7563 on 79 degrees of freedom
##   (82 observations deleted due to missingness)
## Multiple R-squared:  0.4967, Adjusted R-squared:  0.4903
## F-statistic: 77.96 on 1 and 79 DF,  p-value: 2.094e-13
```

```
stargazer(reg2s, type = "text",
          title = "Tableau de régression n°6",
          dep.var.labels = "log du PIB par habitant en 1995",
          covariate.labels = "Valeurs predites (Institutions)")
```

```
##
## Tableau de régression n°6
## ================================================================
##                                       Dependent variable:
##                                  -------------------------------
##                                  log du PIB par habitant en 1995
## ----------------------------------------------------------------
## Valeurs predites (Institutions)              0.872***
##                                               (0.099)
##
## Constant                                     2.350***
##                                               (0.653)
##
## ----------------------------------------------------------------
## Observations                                   81
## R2                                            0.497
## Adjusted R2                                   0.490
## Residual Std. Error                      0.756 (df = 79)
## F Statistic                           77.961*** (df = 1; 79)
## ================================================================
## Note:                                  *p<0.1; **p<0.05; ***p<0.01
```

Le coefficient estimé pour (predicted_avexpr) est significatif (p < 0.01) et plus élevé que dans les régressions OLS précédentes. Une augmentation d'une unité de la variable prédite est associée à une augmentation moyenne de 87,221% du PIB, toutes choses égales par ailleurs. Un coefficient aussi élevé indique que les institutions ont un effet causal élevé sur le PIB et que cette effet des institutions sur le PIB était sous-estimé lorsque l'endogénéité n'était pas corrigée.

3.2.3 Autres tests (test de Sargan et test de Haussman)

```
# Résumé du modèle avec la statistique F de la première regression instrumentale
summary(reg1S, diagnostics = TRUE)
```

```
##
## Call:
## lm(formula = avexpr ~ logem4, data = maketable4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6521 -1.0076  0.1535  0.9478  3.4621
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.5276     0.5481  17.382  < 2e-16 ***
## logem4       -0.6468     0.1152  -5.614 3.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.321 on 72 degrees of freedom
##   (89 observations deleted due to missingness)
## Multiple R-squared:  0.3044, Adjusted R-squared:  0.2948
## F-statistic: 31.51 on 1 and 72 DF,  p-value: 3.48e-07
```

```
# Test de suridentification de Sargan
sargan_test <- summary(reg1S)$diagnostics["Sargan test", ]

# Afficher le résultat du test de suridentification
print(sargan_test)
```

```
## NULL
```

Le test de Sargan nous permet d'évaluer la validité de l'instrument. Dans ce modèle, avec un seul instrument pour une seule variable endogène, le test n'est pas applicable. Mais l'absence de suridentification garantit que la mortalité des colons (logem4) est valide et ne corrèle pas directement avec les résidus. Donc on est sûr qu'il n'y a pas de problème d'exogénéité.

En comparant nos régressions OLS et la régression 2SLS, on observe que les estimations issues de la régression instrumentale sont plus élevées et donc que les estimations OLS étaient biaisées en raison de l'endogénéité. La régression instrumentale nous a permit de corriger ce biais et d'obtenir une mesure plus précise de l'effet des institutions. Pour confirmer cela, nous avons essayer de réaliser un test de Hausman (mais nous n'avons pas réussie). Ce test aurait permit de comparer les coefficients OLS et 2SLS pour vérifier si les différences sont significatives. Une différence significative validerait l'utilisation de la régression instrumentale.

```
hausmantest <- phtest(reg3, reg2s)
```

```
## Error in UseMethod("phtest"): no applicable method for 'phtest' applied to an object of cl
ass "lm"
```

Nous avons essayés de trouver une autre méthode mais à nouveau cela n'a pas fonctionnée.

```
# Extraire les coefficients
beta_ols <- coef(reg3)
beta_2sls <- coef(reg2s)

# Différence des coefficients
diff_beta <- beta_2sls - beta_ols

# Extraire les matrices de variance-covariance
vcov_ols <- vcov(reg3)
vcov_2sls <- vcov(reg2s)

# Calculer la différence des matrices de variance-covariance
vcov_diff <- vcov_2sls - vcov_ols
```

```
## Error in vcov_2sls - vcov_ols: non-conformable arrays
```

```
# Calculer la statistique de Hausman
hausman_stat <- t(diff_beta) %% solve(vcov_diff) %% diff_beta
```

```
## Error: object 'vcov_diff' not found
```

```
# Degrés de liberté (nombre de coefficients comparés)
df <- length(beta_2sls)

# Calculer la P-value
p_value <- pchisq(hausman_stat, df = df, lower.tail = FALSE)
```

```
## Error: object 'hausman_stat' not found
```

```
# Résultats
cat("Statistique de Hausman :", hausman_stat, "\n")
```

```
## Error: object 'hausman_stat' not found
```

```
cat("P-value :", p_value, "\n")
```

```
## Error: object 'p_value' not found
```

3.3 Analyse de la régression logistique

Dans cette section, nous analysons une régression logistique pour étudier la relation entre la probabilité qu'un pays se trouve en Afrique (africa) et différentes variables explicatives telles que l'indicateur institutionnel (avexpr), le climat (lat_abst), et la mortalité des colons (logem4).

On fait auparavant l'hypothèse alors que l'Afrique possède des institutions faibles, et que l'indicateur institutionnel pourrait être alors un déterminant essentiel qui permettrait alors de déterminer automatiquement le fait d'être Africain.

3.3.1 Premier modèle

Ce premier modèle inclut uniquement une constante pour prédire la variable Africa

```
# Régression logistique qui permet d'estimer la probabilité moyenne d'obtenir l'Afrique
model0 <- glm(africa ~ 1,
              data = maketable7,
              family = binomial)

stargazer(model0, type = "text",
          title = "Tableau de régression n°7")
```

```
##
## Tableau de régression n°7
## ================================================
##                         Dependent variable:
##                     ---------------------------
##                                africa
## ------------------------------------------------
## Constant                      -0.815***
##                                (0.170)
##
## ------------------------------------------------
## Observations                    163
## Log Likelihood                -100.485
## Akaike Inf. Crit.              202.971
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

Le coefficient à l'origine est de -0,8 et à l'exponentielle cela donne une odds de 0,4, equivalent à 0,4/1 +0,4=28,57% Ce modèle prédit une probabilité moyenne de 28,57 % qu'un pays soit situé en Afrique, en l'absence d'autres informations, donc au niveau mondial . Cela fournit un point de référence pour évaluer l'ajout de variables explicatives.

3.3.2 Matrice de corrélation et sélection des variables

Nous calculons une matrice de corrélation afin d'avoir une première idée des corrélations que l'on peut trouver et sélectionner les variables pertinentes pour le modèle.

```
# Il est importance que toute la database soit numérique afin de pouvoir utiliser la matrice
de corrélation .

# Filtrer les colonnes numériques: étape imporante
numeric_cols <- select_if(maketable7, is.numeric)

# Calculer la matrice de corrélation sans valeurs manquantes
cor_matrix <- cor(numeric_cols, use = "complete.obs")
```

```
## Warning in cor(numeric_cols, use = "complete.obs"): the standard deviation is
## zero
```

```
# Afficher la matrice
print(cor_matrix)
```

```
##               africa    lat_abst   malfal94     avexpr    logpgp95     logem4
## africa    1.00000000 -0.22438423  0.7428484 -0.3591982 -0.57097524  0.6191578
## lat_abst -0.22438423  1.00000000 -0.4406416  0.3601901  0.41726133 -0.4752727
## malfal94  0.74284843 -0.44064156  1.0000000 -0.4729627 -0.73006004  0.7404646
## avexpr   -0.35919822  0.36019010 -0.4729627  1.0000000  0.73739734 -0.5346873
## logpgp95 -0.57097524  0.41726133 -0.7300600  0.7373973  1.00000000 -0.7461527
## logem4    0.61915777 -0.47527274  0.7404646 -0.5346873 -0.74615269  1.0000000
## asia     -0.35155203 -0.09618615 -0.1168577  0.1631583 -0.02436164 -0.2558623
## yellow    0.09385019 -0.59716714  0.4649478 -0.2976088 -0.36815984  0.4081984
## baseco            NA          NA         NA         NA          NA         NA
## leb95    -0.78445385  0.43913515 -0.8775234  0.5445320  0.79585329 -0.7172603
## imr95     0.72136114 -0.35667189  0.8088717 -0.5245457 -0.79540132  0.7385618
## meantemp  0.22337595 -0.72823834  0.4344476 -0.4943250 -0.60334310  0.5420333
## lt100km  -0.50291628 -0.05194163 -0.3868092  0.1433514  0.27090085 -0.2341743
## latabs   -0.22438037  1.00000000 -0.4406371  0.3601872  0.41725642 -0.4752735
##                asia      yellow baseco      leb95      imr95   meantemp
## africa   -0.35155203  0.09385019     NA -0.7844539  0.7213611  0.2233760
## lat_abst -0.09618615 -0.59716714     NA  0.4391351 -0.3566719 -0.7282383
## malfal94 -0.11685769  0.46494779     NA -0.8775234  0.8088717  0.4344476
## avexpr    0.16315834 -0.29760879     NA  0.5445320 -0.5245457 -0.4943250
## logpgp95 -0.02436164 -0.36815984     NA  0.7958533 -0.7954013 -0.6033431
## logem4   -0.25586227  0.40819843     NA -0.7172603  0.7385618  0.5420333
## asia      1.00000000  0.10453832     NA  0.1782972 -0.1380143  0.2373180
## yellow    0.10453832  1.00000000     NA -0.3573366  0.3057486  0.6123886
## baseco            NA          NA      1         NA         NA         NA
## leb95     0.17829720 -0.35733663     NA  1.0000000 -0.9403784 -0.4204286
## imr95    -0.13801429  0.30574861     NA -0.9403784  1.0000000  0.4070533
## meantemp  0.23731802  0.61238861     NA -0.4204286  0.4070533  1.0000000
## lt100km   0.25800482  0.12270150     NA  0.4746615 -0.4688134  0.1642413
## latabs   -0.09619240 -0.59717159     NA  0.4391313 -0.3566675 -0.7282361
##              lt100km      latabs
## africa   -0.50291628 -0.22438037
## lat_abst -0.05194163  1.00000000
## malfal94 -0.38680924 -0.44063711
## avexpr    0.14335136  0.36018723
## logpgp95  0.27090085  0.41725642
## logem4   -0.23417430 -0.47527353
## asia      0.25800482 -0.09619240
## yellow    0.12270150 -0.59717159
## baseco            NA          NA
## leb95     0.47466153  0.43913126
## imr95    -0.46881338 -0.35666745
## meantemp  0.16424131 -0.72823614
## lt100km   1.00000000 -0.05194894
## latabs   -0.05194894  1.00000000
```

Pourquoi fait-on alors cette matrice de corrélation ? En regardant la colonne Afrique on remarque que l'on trouve de fortes corrélations qui pourraient indiquer un problème de multicolinéarité parfaite : on observe notamment de fortes multicolinéarités notamment avec le loggdp 95.

Concernant toutefois les liens entre varaibles explicatives, On peut observer un lien fort entre la température moyenne annuelle (meantemp) et le climat (lat_abst).

3.3.3 Modèles logit , probit ou logit ?

Nous allons chercher à offrir le modèle le plus performant, avec un équilibre entre modèles Logit et Probit, mais aussi sélection des variables explicatives.

Etape 1 : le terme d'erreur va suivre une distribution logit.

```
library(stargazer)
logit <- glm(africa ~ lat_abst + avexpr + logpgp95 + logem4 + malfal94 + meantemp + lt100km,
             data = maketable7,
             family = binomial)#Family permet d'indiquer ce que l'on souhait : ici LOGIT par
défaut.
stargazer(logit, type = "text",
          title = "Tableau de régression n°8")
```

```
##
## Tableau de régression n°8
## ============================================
##                       Dependent variable:
##                   -------------------------
##                              africa
## --------------------------------------------
## lat_abst                     7.018
##                             (6.219)
##
## avexpr                      -0.156
##                             (0.508)
##
## logpgp95                    -0.254
##                             (1.000)
##
## logem4                      1.943*
##                             (1.136)
##
## malfal94                     2.415
##                             (1.690)
##
## meantemp                    -0.035
##                             (0.177)
##
## lt100km                    -4.657**
##                             (2.026)
##
## Constant                    -6.385
##                            (12.443)
##
## --------------------------------------------
## Observations                  60
## Log Likelihood              -15.409
## Akaike Inf. Crit.           46.818
## ============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

```
#Test de spécification : enlever des variables explicatives.
coeftest(logit, vcov = vcovHC, type = "HC1")
```

```
##
## z test of coefficients:
##
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.384806  10.015096 -0.6375 0.523787
## lat_abst     7.017883   6.040031  1.1619 0.245278
## avexpr      -0.155597   0.387906 -0.4011 0.688332
## logpgp95    -0.253621   0.865264 -0.2931 0.769435
## logem4       1.943178   1.976562  0.9831 0.325553
## malfal94     2.415143   1.780103  1.3567 0.174863
## meantemp    -0.034798   0.131891 -0.2638 0.791901
## lt100km     -4.656520   1.593195 -2.9228 0.003469 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On observe un test de Wald de 41,377 et une Log vraisemblance de -33,240.

On peut observer notamment que l'AIC du modèle Probit est 46,716, tandis que celui du modèle logit est de 46,8.

Etape 2 : choix de loi sur le terme d'erreur : probit

```
#Test de transformation: quel est le meilleur entre tobit ou logit ?
probitmodel <- glm(africa ~ lat_abst + avexpr + logpgp95 + logem4 + malfal94 + meantemp + lt1
00km,
                   data = maketable7,
                   family = binomial(link="probit"))
summary(probitmodel)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + avexpr + logpgp95 + logem4 +
##     malfal94 + meantemp + lt100km, family = binomial(link = "probit"),
##     data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.15672    6.74056  -0.468   0.6396
## lat_abst     3.86459    3.26824   1.182   0.2370
## avexpr      -0.12253    0.27368  -0.448   0.6544
## logpgp95    -0.07812    0.54182  -0.144   0.8854
## logem4       0.89934    0.55273   1.627   0.1037
## malfal94     1.65931    0.95012   1.746   0.0807 .
## meantemp    -0.01437    0.09740  -0.147   0.8827
## lt100km     -2.69208    1.14060  -2.360   0.0183 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 82.108  on 59  degrees of freedom
## Residual deviance: 30.731  on 52  degrees of freedom
##   (103 observations deleted due to missingness)
## AIC: 46.731
##
## Number of Fisher Scoring iterations: 8
```

```r
# Installer les packages nécessaires
library(AER)

# Spécification du modèle Tobit
model1 <- tobit(africa ~ lat_abst + avexpr + logpgp95 + logem4 + malfal94 + meantemp + lt100k
m,data = maketable7, left=0)  # Supposant une censure à gauche à 0
summary(model1)
```

```
##
## Call:
## tobit(formula = africa ~ lat_abst + avexpr + logpgp95 + logem4 +
##     malfal94 + meantemp + lt100km, left = 0, data = maketable7)
##
## Observations: (103 observations deleted due to missingness)
##          Total  Left-censored     Uncensored Right-censored
##             60             34             26              0
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.587042   2.023130   0.290  0.77169
## lat_abst     1.095683   0.992162   1.104  0.26945
## avexpr      -0.006789   0.083404  -0.081  0.93513
## logpgp95    -0.128405   0.189583  -0.677  0.49821
## logem4       0.149940   0.118975   1.260  0.20758
## malfal94     0.956606   0.366039   2.613  0.00896 **
## meantemp    -0.018276   0.031597  -0.578  0.56298
## lt100km     -0.973990   0.386985  -2.517  0.01184 *
## Log(scale)  -0.629110   0.155444  -4.047 5.18e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Scale: 0.5331
##
## Gaussian distribution
## Number of Newton-Raphson Iterations: 6
## Log-likelihood: -33.24 on 9 Df
## Wald-statistic: 41.38 on 7 Df, p-value: 6.8536e-07
```

```
#Vérifions la normalité des résidus
shapiro.test(residuals(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(model1)
## W = 0.94631, p-value = 0.01049
```

```
shapiro.test(residuals(probitmodel))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(probitmodel)
## W = 0.91221, p-value = 0.0003778
```

```
shapiro.test(residuals(logit))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(logit)
## W = 0.90174, p-value = 0.0001523
```

```
#Comparer les Logit, Probit, et Tobit
AIC(model1)
```

```
## [1] 84.48078
```

```
AIC (probitmodel)
```

```
## [1] 46.73082
```

```
AIC(logit)
```

```
## [1] 46.81771
```

```
# Obtenir les pseudo R²
pR2(probitmodel)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -15.3654088 -41.0539059  51.3769942   0.6257260   0.5752622   0.7716437
```

```
pR2(logit)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -15.4088529 -41.0539059  51.2901061   0.6246678   0.5746467   0.7708181
```

D’emblée, on peut déjà observer un Wald très élevé du Tobit avec 41,38, ce qui montre déjà une significativité importante. De plus, son AIC est plus haut (84>46) pour le modèle Tobit, ce qui signifie qu’il est trop complexe toutefois pour être retenu.

Du point de vue du respect des hypothèses, notamment de normalité des résidus, le test de Shapiro montre bien qu’on a une normalité des résidus, cette hypothèse essentielle des modèles Tobit/Probit/Logit n’est pas vérifiée Le test de Shapiro est particulièrement adapté car il est assez puissant pour des petits échantillons, comme ici.

Ainsi, nous avons choisi de continuer avec le modèle le plus simple et informatif du point de vue des données.

3.3.4 Diminution du problème de multicolinéraité et Backward selection

Nous tentons ensuite de chercher la meilleure spécification en allant du général vers le spécifique. A chaque itération, on supprime la variable considérée comme la moins significative, donc avec la plus forte P-VALUE !

```
summary(logit)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + avexpr + logpgp95 + logem4 +
##     malfal94 + meantemp + lt100km, family = binomial, data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.3848    12.4426  -0.513   0.6079
## lat_abst      7.0179     6.2190   1.128   0.2591
## avexpr       -0.1556     0.5084  -0.306   0.7596
## logpgp95     -0.2536     0.9998  -0.254   0.7998
## logem4        1.9432     1.1358   1.711   0.0871 .
## malfal94      2.4151     1.6900   1.429   0.1530
## meantemp     -0.0348     0.1775  -0.196   0.8446
## lt100km      -4.6565     2.0264  -2.298   0.0216 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 82.108  on 59  degrees of freedom
## Residual deviance: 30.818  on 52  degrees of freedom
##   (103 observations deleted due to missingness)
## AIC: 46.818
##
## Number of Fisher Scoring iterations: 7
```

Grâce à la bacaward selection, le but est de supprimer les varaibles les moins significatives, afin de réussir à obtenir un modèle qui va éviter le surapprentissage. Au final, on peut déjà enlever meantemp.

```
model1 <- glm(africa ~ lat_abst + avexpr + logpgp95 + logem4 + malfal94  + lt100km,
              data = maketable7,
              family = binomial)
summary(model1)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + avexpr + logpgp95 + logem4 +
##     malfal94 + lt100km, family = binomial, data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.75219    5.53272  -0.317   0.7515
## lat_abst     5.76366    4.28029   1.347   0.1781
## avexpr      -0.09933    0.47962  -0.207   0.8359
## logpgp95    -0.51035    0.72327  -0.706   0.4804
## logem4       1.19404    0.80636   1.481   0.1387
## malfal94     2.80458    1.64976   1.700   0.0891 .
## lt100km     -4.96992    2.01679  -2.464   0.0137 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 32.698  on 54  degrees of freedom
##   (102 observations deleted due to missingness)
## AIC: 46.698
##
## Number of Fisher Scoring iterations: 7
```

```
pR2(model1)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.3489325 -41.8794525  51.0610399   0.6096192   0.5670211   0.7593909
```

Avexpr a la p-value la plus élevée donc on enleve cette variable.

```
model2 <- glm(africa ~ lat_abst + logpgp95 + logem4 + malfal94  + lt100km,
              data = maketable7,
              family = binomial)
summary(model2)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + logpgp95 + logem4 + malfal94 +
##     lt100km, family = binomial, data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.8752     5.5338  -0.339   0.7347
## lat_abst      5.7175     4.2734   1.338   0.1809
## logpgp95     -0.5925     0.6076  -0.975   0.3294
## logem4        1.2228     0.8059   1.517   0.1292
## malfal94      2.7666     1.6328   1.694   0.0902 .
## lt100km      -4.8606     1.9381  -2.508   0.0121 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 32.741  on 55  degrees of freedom
##   (102 observations deleted due to missingness)
## AIC: 44.741
##
## Number of Fisher Scoring iterations: 7
```

```
library("pscl")
pR2(model2)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.3704875 -41.8794525  51.0179301   0.6091045   0.5667150   0.7589810
```

```
model3 <- glm(africa ~ lat_abst  + logem4 + malfal94  + lt100km,
              data = maketable7,
              family = binomial)
summary(model3)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + logem4 + malfal94 + lt100km,
##     family = binomial, data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -6.1208     3.4495  -1.774   0.0760 .
## lat_abst      4.2469     3.6804   1.154   0.2485
## logem4        1.1266     0.7349   1.533   0.1253
## malfal94      3.5104     1.4706   2.387   0.0170 *
## lt100km      -4.7303     1.8904  -2.502   0.0123 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 33.685  on 56  degrees of freedom
##   (102 observations deleted due to missingness)
## AIC: 43.685
##
## Number of Fisher Scoring iterations: 6
```

```
pR2(model3)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.8424133 -41.8794525  50.0740784   0.5978359   0.5599586   0.7499325
```

On voit que le 2eme modèle a un pseudo Rcarré de 0,75, de 75%. On peut clairement observer que le Logpgp95 a une pvalue de 0,32, ce qui pousse alors à enlever cette variable en Backward selection. On fera de même, jusq'à ce qu'il n'y a plus de varaible non significative.

```
model4 <- glm(africa ~   logem4 + malfal94  + lt100km  ,
              data = maketable7,
              family = binomial)
summary(model4)
```

```
##
## Call:
## glm(formula = africa ~ logem4 + malfal94 + lt100km, family = binomial,
##     data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.7685     2.5473  -1.479   0.1390
## logem4        0.8099     0.6383   1.269   0.2046
## malfal94      3.2239     1.4267   2.260   0.0238 *
## lt100km      -4.5407     1.8005  -2.522   0.0117 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 35.055  on 57  degrees of freedom
##   (102 observations deleted due to missingness)
## AIC: 43.055
##
## Number of Fisher Scoring iterations: 6
```

```
pR2(model4)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -17.5275172 -41.8794525  48.7038705   0.5814769   0.5499624   0.7365449
```

```
model5 <- glm(africa ~ malfal94  + lt100km  ,
              data = maketable7,
              family = binomial)
summary(model5)
```

```
##
## Call:
## glm(formula = africa ~ malfal94 + lt100km, family = binomial,
##     data = maketable7)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.6824     0.6451  -1.058   0.2902
## malfal94      4.5040     1.1125   4.049 5.15e-05 ***
## lt100km      -4.2250     1.7164  -2.462   0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 36.930  on 58  degrees of freedom
##   (102 observations deleted due to missingness)
## AIC: 42.93
##
## Number of Fisher Scoring iterations: 6
```

```
pR2(model5)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -18.4648659 -41.8794525  46.8291732   0.5590949   0.5359168   0.7177341
```

In fine, on remarque que le R2 final a un peu diminué , il est désormais à 0,73, vs 0,77. Toutefois, l'intérêt du Backward selection est de sélectionner uniquement les varaibles les plus pertinentes afin de permettre par la suite d'avoir une spécification suffisamment robuste, pour prédire de nouvelles données. Ainsi, l'AIC qui est passé de 46 à 42, montre bien une simplicité forte, donc une possibilité de s'adapter à de nouvelles données plus forte.

Maintenant , on va tenter de comparer les BIC

```
#2EME valeur de performance du modèle !
BIC (probitmodel)
```

```
## [1] 63.48557
```

```
BIC (model1)
```

```
## [1] 61.47398
```

```
BIC (model2)
```

```
## [1] 57.40622
```

```
BIC (model3)
```

```
## [1] 54.2392
```

```
BIC(model4)
```

```
## [1] 51.49853
```

```
BIC(model5)
```

```
## [1] 49.26235
```

```
#Amélioration de la multicolinéarité
library(car)
vif(probitmodel)
```

```
## lat_abst   avexpr logpgp95   logem4 malfal94 meantemp  lt100km
## 3.575225 2.091644 3.648947 2.951449 2.235127 4.723751 1.281257
```

```
vif(model2)
```

```
## lat_abst logpgp95   logem4 malfal94  lt100km
## 1.967394 1.678492 2.061546 1.968175 1.184540
```

```
vif(model3)
```

```
## lat_abst   logem4 malfal94  lt100km
## 1.549327 1.919277 1.554387 1.182913
```

```
vif(model5)
```

```
## malfal94  lt100km
## 1.118893 1.118893
```

En utilisant le critère de BIC, on observe une valeur d'environ 20 unités plus faibles :cela est considéré comme énorme, une très forte différence. Or le BIC est une mesure très intéressante , non pour la prédiction, mais pour l'efficacité du modèle en lui-même.

Le modèle BIC se concentre davantage sur l'évaluation de l'adéquation réelle du modèle, sacrifiant souvent une partie du pouvoir prédictif au profit d'un modèle plus proche de la vérité.

Dans quelle mesure ces modifications ont été utiles ?

Ensuite, en faisant les VIF, on part d'un problème de multicolinéarité modérée, et on observe que la simplification du modèle permet une diminution extrêmement forte de la multicolinéarité, devenue très légère.

Etant donné que le Backward selection, conduit à l'élimination du loggdp dans la régression africaine, cela montre bien qu'il n'y a pas forcément un impact si élevé de la croissance économique sur l'Afrique, qu'on suppose avoir des institutions faibles . Il pourrait y avoir une exception africaine ; on pourrait l'expliquer par une explication géographique, qui serait alors le paludisme.

3.3.4.2 Test de transformation

Supposons un modèle Probit de base sélectionné auparavant (model2) avec un AIC de 44.741. Ajoutons un terme en Log:

```
stargazer(model2, type = "text",
          title = "Tableau de régression n°8")
```

```
##
## Tableau de régression n°8
## ================================================
##                        Dependent variable:
##                    ----------------------------
##                               africa
## ------------------------------------------------
## lat_abst                       5.718
##                               (4.273)
##
## logpgp95                      -0.593
##                               (0.608)
##
## logem4                         1.223
##                               (0.806)
##
## malfal94                      2.767*
##                               (1.633)
##
## lt100km                      -4.861**
##                               (1.938)
##
## Constant                      -1.875
##                               (5.534)
##
## ------------------------------------------------
## Observations                    61
## Log Likelihood                -16.370
## Akaike Inf. Crit.             44.741
## ================================================
## Note:              *p<0.1; **p<0.05; ***p<0.01
```

```
model2_log <- glm(africa ~ lat_abst + log(logpgp95) + log(logem4) + malfal94 + lt100km,
                  data = maketable7,
                  family = binomial)
library("stargazer")
pR2(model2_log)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.7838678 -41.8794525  50.1911694   0.5992338   0.5608025   0.7510626
```

```
stargazer(model2_log, type = "text",
          title = "Tableau de régression n°9")
```

```
##
## Tableau de régression n°9
## ===============================================
##                       Dependent variable:
##                   ---------------------------
##                              africa
## -----------------------------------------------
## lat_abst                      5.419
##                              (4.227)
##
## log(logpgp95)                -4.058
##                              (4.765)
##
## log(logem4)                   4.370
##                              (3.211)
##
## malfal94                     3.187**
##                              (1.594)
##
## lt100km                     -4.790**
##                              (1.914)
##
## Constant                      0.700
##                             (10.341)
##
## -----------------------------------------------
## Observations                    61
## Log Likelihood               -16.784
## Akaike Inf. Crit.            45.568
## ===============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Nous obtenons un AIC de 45.568, qui est un peu-audessus.De plus, on remarque qu'en log, le paludisme semble être également la seule variable explicative significative, avec notamment la latitude.

Ensuite avec la spécification du PIB au carré:

```
model2_quad <- glm(africa ~ lat_abst + logpgp95 + I(logpgp95^2) + logem4 + I(logem4^2) + malf
al94 + I(malfal94^2) + lt100km,
                   data = maketable7,
                   family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
stargazer(model2_quad, type = "text",
          title = "Tableau de régression n°10")
```

```
##
## Tableau de régression n°10
## ============================================
##                       Dependent variable:
##                   -------------------------
##                              africa
## --------------------------------------------
## lat_abst                     7.840
##                             (6.852)
##
## logpgp95                    13.743
##                            (10.747)
##
## I(logpgp952)                -1.013
##                             (0.702)
##
## logem4                      -16.115
##                            (10.163)
##
## I(logem42)                   1.981
##                             (1.293)
##
## malfal94                    -13.062
##                             (9.623)
##
## I(malfal942)                15.615
##                            (10.352)
##
## lt100km                    -6.428**
##                             (2.662)
##
## Constant                    -13.172
##                            (36.827)
##
## --------------------------------------------
## Observations                  61
## Log Likelihood              -12.607
## Akaike Inf. Crit.           43.214
## ============================================
## Note:             *p<0.1; **p<0.05; ***p<0.01
```

Cette fois ci l'AIC est de 43,214 : le modèle avec la spécification du PIB au carré semble être le meilleur en terme de simplicité ; tout en étant également celui qui propose le pseudo Rcarré le plus élevé : 0,82.

```
library("pscl")
pR2(model2_quad )
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -12.6071838 -41.8794525  58.5445374   0.6989649   0.6170100   0.8263394
```

```
pR2(model2_log)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.7838678 -41.8794525  50.1911694   0.5992338   0.5608025   0.7510626
```

```
pR2(model2)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -16.3704875 -41.8794525  51.0179301   0.6091045   0.5667150   0.7589810
```

```
# R2CU : 0,82
# 0,7589.
```

On remarque que la spécification au carré permet une élévation de la précision notable. Ainsi, on choisit ce modèle qui semble être désormais beaucoup plus précis tout en étant le plus simple selon l'AIC.

Test de différence:

```
anova(model2_quad,model2_log,model2)
```

```
## Analysis of Deviance Table
##
## Model 1: africa ~ lat_abst + logpgp95 + I(logpgp95^2) + logem4 + I(logem4^2) +
##     malfal94 + I(malfal94^2) + lt100km
## Model 2: africa ~ lat_abst + log(logpgp95) + log(logem4) + malfal94 +
##     lt100km
## Model 3: africa ~ lat_abst + logpgp95 + logem4 + malfal94 + lt100km
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        52     25.214
## 2        55     33.568 -3  -8.3534  0.03925 *
## 3        55     32.741  0   0.8268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Une différence de déviance de -8 indique que le modèle plus complexe a une log-vraisemblance supérieure de 4 unités. Autrement dit, le modèle plus complexe ajuste mieux les données que le modèle simple. On remarque alors une supériorité nette du modèle quadratique sur le modèle en log.

3.3.5 Vérification des hypothèses

Nous allons ensuite vérifier si toutes les hypothèses de la régression logistique sont validées.

3.3.5.1 Hypothèse de linéarité

vérifions si désormais nous avons une relation suffisamment linéaire, ou bien nous devons absolument réaliser

```
#Modèle Gam afin de tester la non linéarité
  library("mgcv")
  model_gam <- gam(africa ~ s(lat_abst) + s(logpgp95) +
                     s(logem4) + s(malfal94) +
                     s(lt100km), data = maketable7)
  summary(model_gam)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## africa ~ s(lat_abst) + s(logpgp95) + s(logem4) + s(malfal94) +
##     s(lt100km)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.44262    0.02929   15.11   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##               edf Ref.df     F  p-value
## s(lat_abst) 4.967  5.863 2.944 0.016258 *
## s(logpgp95) 3.914  4.716 5.437 0.001031 **
## s(logem4)   4.842  5.754 2.245 0.061936 .
## s(malfal94) 5.626  6.550 4.881 0.000585 ***
## s(lt100km)  1.000  1.000 1.039 0.314089
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.791   Deviance explained = 86.2%
## GCV = 0.080487  Scale est. = 0.052319  n = 61
```

```
AIC(model_gam)
```

```
## [1] 11.55754
```

On trouve un AIC très faible d'environ 11 avec le modèle GAM. Le modèle permet de souligner de forte non-linéarité : par exemple, la latitude a un effet non linéaire significatif car F>4, et mieux encore, le LOGGDP réagirait bien, ce qui fait qu'au final la prise en compte de la non linéarité de diverses manières pourrait modifier de façon conséquente notre résultat.

3.3.5.2 Hypothèse d'indépendance

Une violation résulter en des valeurs aberrantes si les données ne sont pas identiques et indépendantes. Cette absence de valeurs aberrantes est nécessaire pour que le modèle de régression logistique fonctionne mieux.

1er critère: Identifier les indices des valeurs aberrantes

```
# Résidus studentisés
residuals_studentized <- rstudent(model2_quad)

# Identifier les indices des valeurs aberrantes (seuil typique : |résidu| > 2)
outliers <- which(abs(residuals_studentized) > 2)

# Afficher les indices et les valeurs
outliers
```

```
##  48  65 147
##  19  27  53
```

2e critère: Identifier les observations influentes

```
#Calculons la distance de Cook
cook_distance <- cooks.distance(model2_quad)

# Identifier les observations influentes (seuil typique : Cook > 4/n)
influential <- which(cook_distance > (4 / nrow(data)))

# Afficher les indices et les valeurs
influential
```

```
## integer(0)
```

On remarque qu'il n'y a aucun valeur aberrante selon ce critère.

3e critère: Identifier les points avec un leverage élevé, des seuils encore plus faible.

```
#Calculons le leverage
leverage <- hatvalues(model2_quad)
# Identifier les points avec un leverage élevé (seuil typique : 2 * p / n)
high_leverage <- which(leverage > (2 * length(coef(model2_quad)) / nrow(data)))

# Afficher les indices et les valeurs
high_leverage
```

```
## integer(0)
```

On peut remarquer que le nombre de valeurs aberrantes semble être assez faible. On retrouve avec le 1er critère tout de même 6 valeurs aberrantes.

3.3.5.3 Hypothèse d'absence de multicolinéarité parfaite

On n'utilise pas le VIF du modèle quadratique car il est forcément biaisé par les modélisations. Le fait d'ajouter des termes au carré rend forcément plus complexe le tout.

```
vif(model2)
```

```
## lat_abst logpgp95   logem4 malfal94  lt100km
## 1.967394 1.678492 2.061546 1.968175 1.184540
```

```
vif(model5)
```

```
## malfal94  lt100km
## 1.118893 1.118893
```

On va corriger les valeurs aberrantes en combinant les indices des observations influentes et aberrantes.

```
to_remove <- unique(c(outliers, influential, high_leverage))

# Créer un nouveau jeu de données sans ces observations.
data_clean <- maketable7[-to_remove, ]


model2_quad <- glm(africa ~ lat_abst + logpgp95 + I(logpgp95^2) + logem4 + I(logem4^2) + malf
al94 + I(malfal94^2) + lt100km,
                   data = data_clean,
                   family = binomial)

model5 <- glm(africa ~ lat_abst + logpgp95  + logem4  + malfal94 + lt100km,
                   data = data_clean,
                   family = binomial)

model2_quad1 <- glm(africa ~ lat_abst + logpgp95 + I(logpgp95^2) + logem4 + I(logem4^2) + mal
fal94 + I(malfal94^2) + lt100km,
                    data = maketable7,
                    family = binomial)
library("stargazer")
stargazer(model2_quad1,model2_quad, type = "text",
          title = "Tableau de régression n°11")
```

```
##
## Tableau de régression n°11
## ============================================
##                       Dependent variable:
##                   -------------------------
##                              africa
##                         (1)           (2)
## --------------------------------------------
## lat_abst                7.840         7.840
##                        (6.852)       (6.852)
##
## logpgp95                13.743        13.743
##                        (10.747)      (10.747)
##
## I(logpgp952)            -1.013        -1.013
##                        (0.702)       (0.702)
##
## logem4                 -16.115       -16.115
##                        (10.163)      (10.163)
##
## I(logem42)              1.981         1.981
##                        (1.293)       (1.293)
##
## malfal94               -13.062       -13.062
##                        (9.623)       (9.623)
##
## I(malfal942)            15.615        15.615
##                        (10.352)      (10.352)
##
## lt100km                -6.428**      -6.428**
##                        (2.662)       (2.662)
##
## Constant               -13.172       -13.172
##                        (36.827)      (36.827)
##
## --------------------------------------------
## Observations              61            61
## Log Likelihood         -12.607       -12.607
## Akaike Inf. Crit.       43.214        43.214
## ============================================
## Note:               *p<0.1; **p<0.05; ***p<0.01
```

On observe un AIC identique amélioré avec le dataclean ; tandis que la significativité des varialbes ne s'en retrouve pas amélioré.

```
pR2(model2_quad)
```

```
## fitting null model for pseudo-r2
```

```
##         llh     llhNull          G2    McFadden        r2ML        r2CU
## -12.6071838 -41.8794525  58.5445374   0.6989649   0.6170100   0.8263394
```

```
summary(model2_quad)
```

```
##
## Call:
## glm(formula = africa ~ lat_abst + logpgp95 + I(logpgp95^2) +
##     logem4 + I(logem4^2) + malfal94 + I(malfal94^2) + lt100km,
##     family = binomial, data = data_clean)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -13.1724    36.8270  -0.358   0.7206
## lat_abst        7.8401     6.8522   1.144   0.2526
## logpgp95       13.7431    10.7474   1.279   0.2010
## I(logpgp95^2)  -1.0132     0.7024  -1.442   0.1492
## logem4        -16.1147    10.1631  -1.586   0.1128
## I(logem4^2)     1.9812     1.2932   1.532   0.1255
## malfal94      -13.0624     9.6233  -1.357   0.1747
## I(malfal94^2)  15.6154    10.3516   1.508   0.1314
## lt100km        -6.4282     2.6625  -2.414   0.0158 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 83.759  on 60  degrees of freedom
## Residual deviance: 25.214  on 52  degrees of freedom
##   (99 observations deleted due to missingness)
## AIC: 43.214
##
## Number of Fisher Scoring iterations: 9
```

On peut observer un effet positif de la latitude, du PIB sur la probabilité d'êrte Africain. On peut observer un effet négatif de la Malaria et du fait d'avoir un logement sur la probabilité d'être Africain.

3.3.6 Vérification de la performance du modèle

```
# Vérifier la structure de la variable africa dans maketable7
library(pROC)

# Cette variable doit être binaire entre 0 et 1 .
str(maketable7$africa)
```

```
##  num [1:163] 0 1 0 0 0 0 0 0 1 0 ...
##  - attr(*, "label")= chr "dummy=1 for Africa"
##  - attr(*, "format.stata")= chr "%9.0g"
```

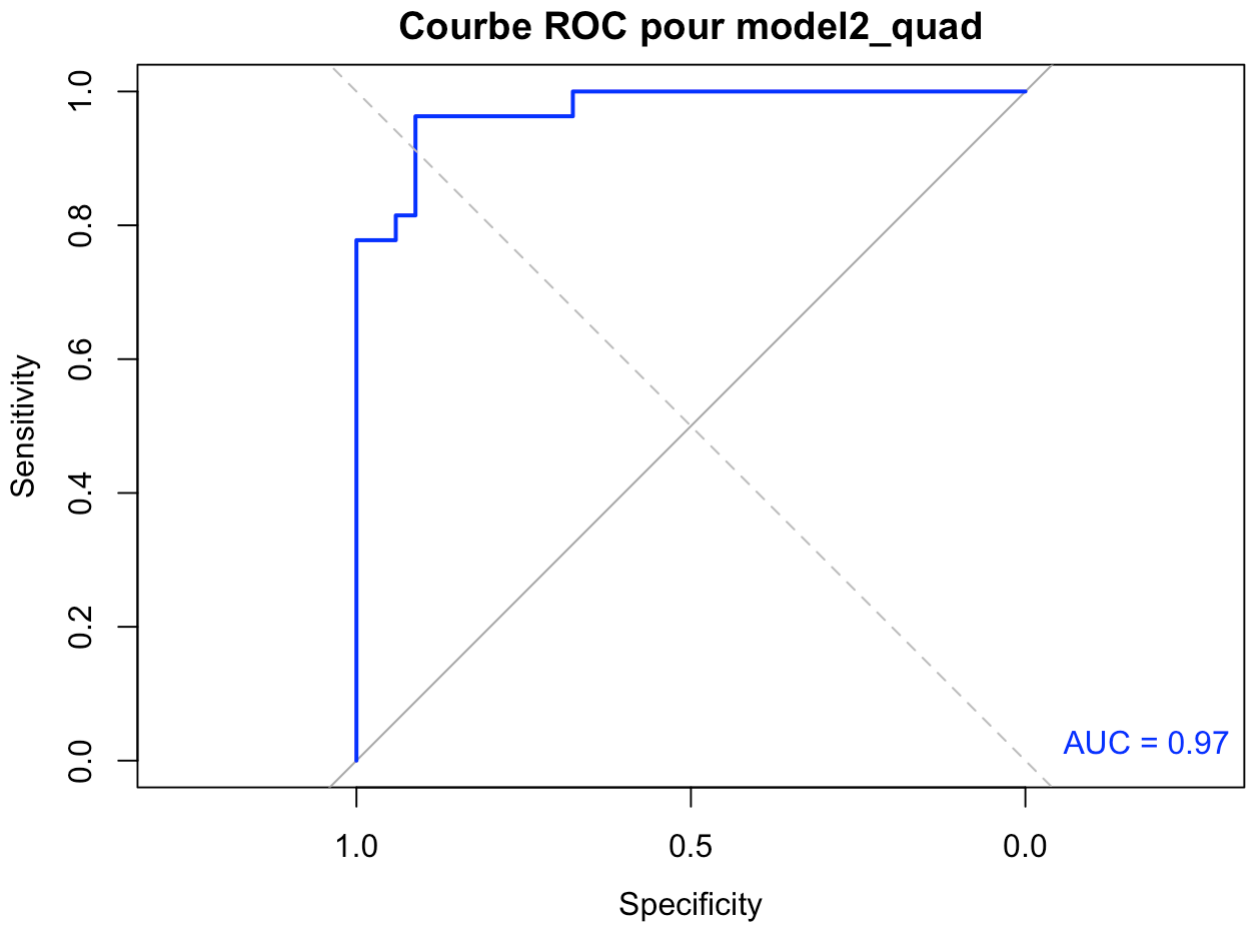
```
# Prédire les probabilités avec le modèle model2_quad
predicted_prob <- predict(model2_quad, newdata = maketable7, type = "response")

# Calculer la courbe ROC
roc_curve <- roc(maketable7$africa, predicted_prob)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```
# Tracons la courbe ROC
plot(roc_curve, col = "blue", lwd = 2, main = "Courbe ROC pour model2_quad")
abline(a = 0, b = 1, lty = 2, col = "gray")
legend("bottomright", legend = paste("AUC =", round(auc(roc_curve), 2)),
       bty = "n", text.col = "blue")
```



Nous observons ainsi que la valeur de notre AUC est extrêmement importante, on retrouve une valeur d'AUC proche de 1.Cela signifie que l'aire sous la courbe de ROC est proche de 1, ainsi notre modèle a une capacité de classification proche de la perfection. Toutefois, on se doit de vérifier le tout car les données peuvent être fortement déséquilibrées. Cette valeur pourrait également indiqué un overfitting, un surapprentissage dans les données d'où la préférence .

```
library(MLmetrics)
# Prédire les probabilités avec le modèle
predicted_prob <- predict(model2_quad, newdata = maketable7, type = "response")

predicted_prob1 <- predict(model5, newdata = maketable7, type = "response")

# Convertir les probabilités en classes avec un seuil (par défaut 0.5)
predicted_class <- ifelse(predicted_prob >= 0.5, 1, 0)
predicted_class1 <- ifelse(predicted_prob1 >= 0.5, 1, 0)

# Obtenir les vraies classes
actual_class <- maketable7$africa

# Calcul des métriques pour le modele quadratique
accuracy <- Accuracy(predicted_class, actual_class)
recall <- Recall(predicted_class, actual_class, positive = 1)
precision <- Precision(predicted_class, actual_class, positive = 1)
f1_score <- F1_Score(predicted_class, actual_class, positive = 1)


accuracy1 <- Accuracy(predicted_class1, actual_class)
recall1 <- Recall(predicted_class1, actual_class, positive = 1)
precision1 <- Precision(predicted_class1, actual_class, positive = 1)
f1_score1 <- F1_Score(predicted_class1, actual_class, positive = 1)
```

Affichons les résultats:

```
cat("Accuracy:", accuracy, "\n")
```

```
## Accuracy: NA
```

Il affiche une valeur manquante, cela arrive notamment quand il y en a trop dans les données.

```
cat("Recall:", recall, "\n")
```

```
## Recall: 0.8846154
```

88% des exemples réellement positifs (réellement africain) ont été correctement identifiés par le modèle.

```
cat("Precision:", precision, "\n")
```

```
## Precision: 0.8518519
```

85,19 % des prédictions positives du modèle sont correctes ! .

```
cat("F1-Score:", f1_score, "\n")
```

```
## F1-Score: 0.8679245
```

Le F1 score est le score le plus important, il permet d'équilibrer l'importance des faux positifs tout comme des faux négatifs ; sans influence des proportions initiales de Y dans la base de données. Ici on trouve une performance de 86,79%, ce qui fait que le modèle a une très bonne performance, mais pas excellente. En conclusion cela indique bien que le modèle est capable de détecter la majorité des cas. Pour avoir l'Accuracy , cela nécessiterait de détecter des données.

4) Conclusion

Avec plus de 18 000 citations, on peut dire sans crainte que l'article d'Acemoglu Jonson et robinson de 2001 est un article très important dans les recherches sur l'importance des institutions sur le développement. Les régressions économétriques semblent confirmer des liens entre mortalité des colons et institutions ainsi qu'entre institutions et développement économique. Il existe cependant plusieurs critiques d'auteurs que ce soit sur une probabilité de causalité inverse, ou même sur des variables omises qui serait importantes.