Effectuez une prédiction de revenus

Pinos Bruno

SOMMAIRE

Exucative Summary

L’entreprise souhaite un modèle permettant de déterminer le revenu potentiel d’une personne à partir seulement de son pays d’origine et du revenu de ses parents .

Trouver un modèle efficace pour faire une telle prédiction ?

Insight et recommandation:

Un modèle de régression linéaire appliqué sur le logarithme du revenu permet une assez bonne prédiction .

Mission 1

Les Données

« Traitement, nettoyage, analyse »

Importation des données

Source : World Income Distribution :
https://openclassrooms.com/fr/projects/148/assignment

	country	year_survey	quantile	nb_quantiles	income	gdpppp
1	ALB	2008	1	100	728,89795	7297
2	ALB	2008	2	100	916,66235	7297
3	ALB	2008	3	100	1010,916	7297
4	ALB	2008	4	100	1086,9078	7297
5	ALB	2008	5	100	1132,6997	7297

```
###J'augmente de 3.6% les revenu de 2006 car la croissance du gdpppp
###mondial a été de 3.6% entre 2006 et 2008
df[df$year_survey == 2006,"gdpppp"]<-
  df[df$year_survey == 2006,"gdpppp"]*1.036

df[df$year_survey == 2006,"income"]<-
  df[df$year_survey == 2006,"income"]*1.036

###Je divise par 0.97 les pays étudié en 2009 car le gdpppp mondiale
###a chuté de 3% entre 2008 et 2009
df[df$year_survey == 2009,"gdpppp"]<-
  df[df$year_survey == 2009,"gdpppp"]/0.97
```

Importation des données

Source : World Income Distribution : http://www.fao.org/faostat/fr/#data/OA

Problématique :
Ici : le nom est donné en français
WID : le nom est donné en ISO 3

	nom_pays	pop
1	Afghanistan	27722.276
2	Afrique du Sud	49779.471
3	Albanie	3002.678
4	Algérie	34730.608
5	Allemagne	81065.752

Source : https://sql.sh/514-liste-pays-csv-xml

	country	nom_pays
1	AFG	Afghanistan
2	ALB	Albanie
3	ATA	Antarctique
4	DZA	Algérie
5	ASM	Samoa Américaines

Importation des données

Ajout de deux pays :

```
nom_pays[242,]<-c("SRB","Serbie")
nom_pays[243,]<-c("MNE","Monténégro")
```

Renommage des pays problématiques :

```
###Je modifie les nom_pays qui ne correspondent pas exactement
fao$nom_pays<-fao$nom_pays%>%as.vector()

fao$nom_pays[fao$nom_pays == "États-Unis d'Amérique"]<-"États-Unis"

fao$nom_pays[fao$nom_pays ==
               "Bolivie (État plurinational de)"]<-"Bolivie"

fao$nom_pays[fao$nom_pays == "République centrafricaine"]<-
  "République Centrafricaine"

fao$nom_pays[fao$nom_pays == "Chine, continentale"]<-"Chine"

fao$nom_pays[fao$nom_pays == "Chine, Taiwan Province de"]<-
  "Taïwan"

fao$nom_pays[fao$nom_pays == "République démocratique du Congo"]<-
  "République Démocratique du Congo"
```

Calcul de l'indice de Gini

```
###Je crée une colonne gini dans ma table principale et je la
###remplis en calculant les indices de chaque pays
df$gini<-0

for (i in 0:pays_analyse-1){
  df$gini[(i*100+1):((i+1)*100)]<-
    round(gini(df[row.names(df) ==
                    (i*100+1):((i+1)*100),"income"])*100,1)
}
```

	country	year_survey	quantile	nb_quantiles	income	gdpppp	gini
1	ALB	2008	1	100	728.8980	7297	30.5
2	ALB	2008	2	100	916.6623	7297	30.5
3	ALB	2008	3	100	1010.9160	7297	30.5
4	ALB	2008	4	100	1086.9078	7297	30.5

Création de la table principale

country	country_full	year_survey	quantile	nb_quantiles	income	gdpppp	pop	gini
ALB	Albanie	2008	1	100	728.8980	7297	3002.678	30.5
ALB	Albanie	2008	2	100	916.6623	7297	3002.678	30.5
ALB	Albanie	2008	3	100	1010.9160	7297	3002.678	30.5
ALB	Albanie	2008	4	100	1086.9078	7297	3002.678	30.5
ALB	Albanie	2008	5	100	1132.6997	7297	3002.678	30.5

Conclusion

Les années utilisées par la WID vont de 2004 à 2011, avec une forte proportion de 2007 et 2008. Mon étude portera seulement sur l'année **2008**

116 pays sont présents dans l'étude de la WID
Mon étude portera sur **111** d'entre eux (Environ 50% des pays recensés par la FAO)

Environ **6,2 milliards** de personnes sont couvertes par l'étude, soit environ **88%** de la population mondiale recensée par la FAO.

Les quantiles utilisés par la WID sont des **percentiles**, pour l'ensemble des pays

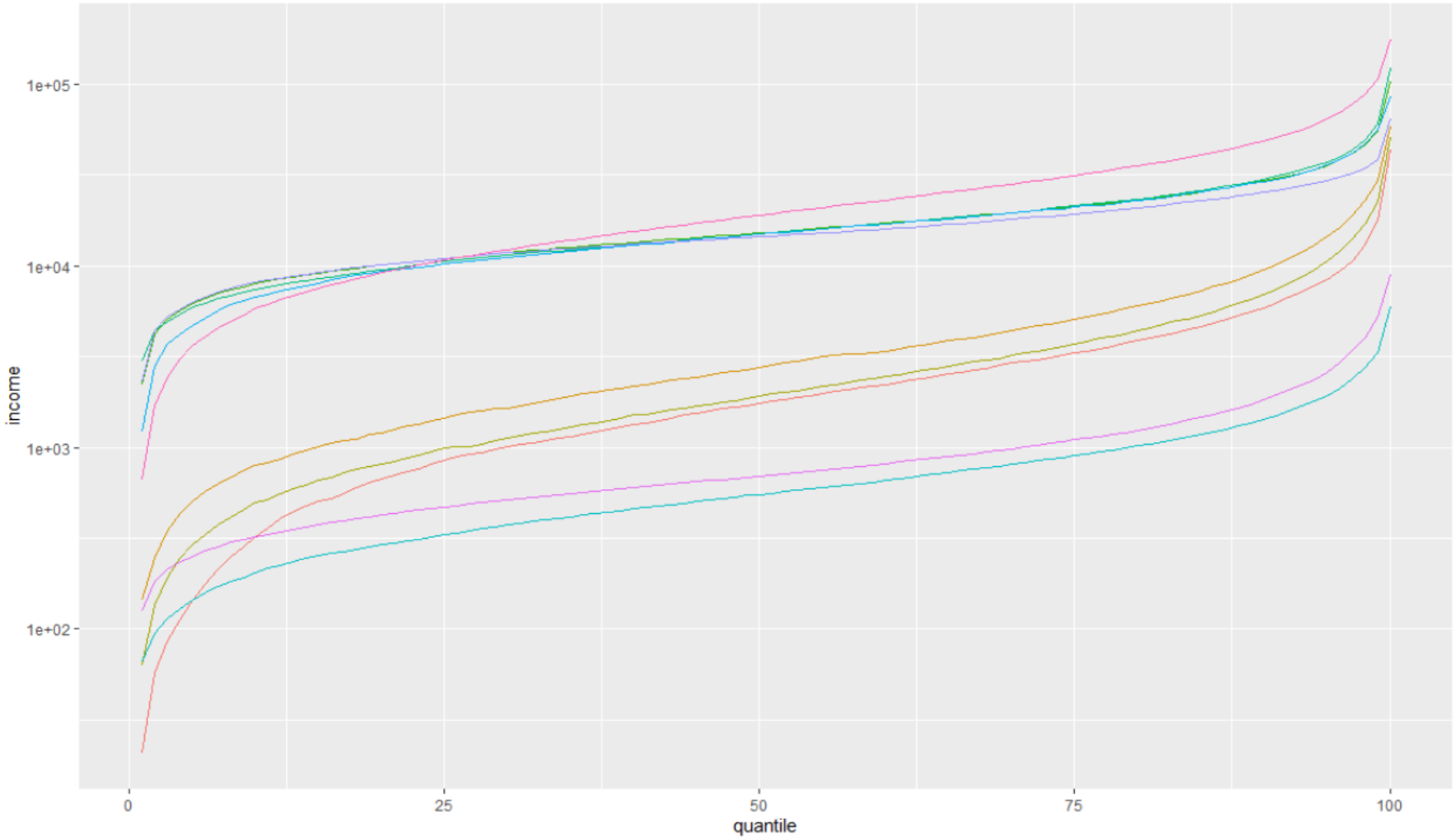
L'utilisation des percentiles est une **bonne méthode**, car cela permet de considérablement réduire la taille de l'échantillon (versus un individu par personne), tout en préservant suffisamment d'information.

Le dollar PPP est une unité qui permet de comparer le pouvoir d'achat entre deux pays sans distorsion due aux taux de change.

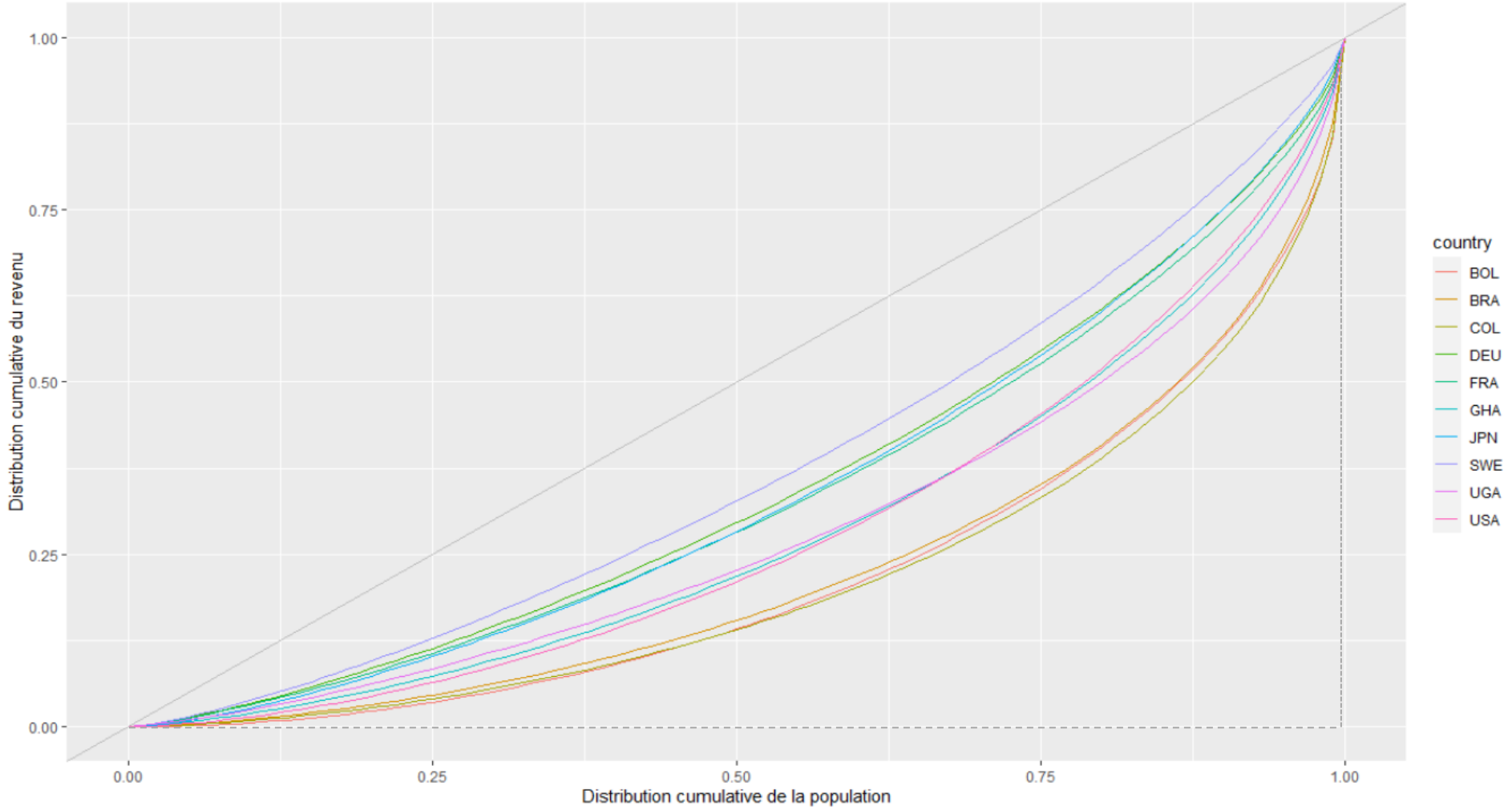
Mission 2

Répartition des revenus

La distribution des revenus varie selon les pays (échelle logaritmique)

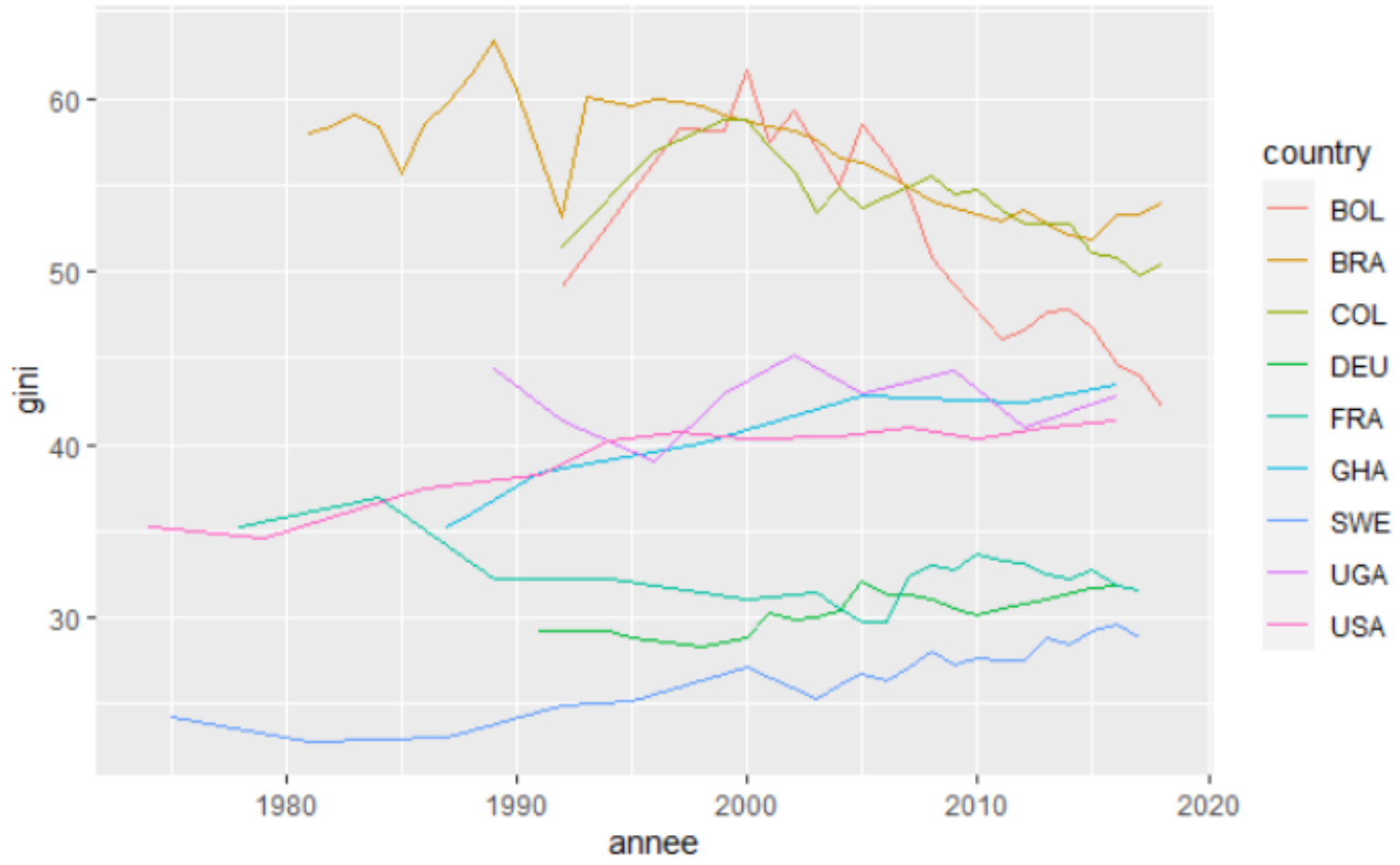


Courbe de Lorenz des USA similaire à celles des pays africains



Indice de Gini

Source : The World Bank:
https://data.worldbank.org/indicator/SI.POV.GINI



Gini rank

	country_full	gini
1	Slovénie	23.1
2	Slovaquie	24.7
3	République Tchèque	25.3
4	Suède	25.5
5	Ukraine	25.5

	country_full	gini
1	Afrique du Sud	67.0
2	Honduras	60.2
3	Colombie	56.9
4	République Centrafricaine	56.2
5	Bolivie	56.1

	country_full	rank
1	France	39

Conclusion

Au premier abord on pourrait croire que l'indice de Gini est directement corrélé avec le gdpppp du pays mais quand on regarde plus précisément on voit que ce n'est pas aussi simple.

Les Etats unis au niveau des pays africains, des pays d'Amérique du sud dans le worst5 alors que les pays africains sont bien plus Pauvre, trois pays d'europe de l'est leader dans le top5 alors que leurs gdpppp sont plutôt moyen etc…

Mission 3

Table des probabilités conditionnelles

Coefficient d’élasticité

Source : elasticity.txt:
https://openclassrooms.com/fr/projects/148/assignment

Coefficients of intergenerational elasticity between parents’ and children’s income

	Base case	Optimistic (high mobility)	Pessimistic (low mobility)
Nordic European countries and Canada	0.2	0.15	0.3
Europe (except nordic countries)	0.4	0.3	0.5
Australia/New Zealand/USA	0.4	0.3	0.5
Asia	0.5	0.4	0.6
Latin America/Africa	0.66	0.5	0.9

Classement par région

```
###Je classe mes pays par région
amerique_latine<-read.csv("amerique_latine.csv",
                encoding = "UTF-8")%>%
  select("country_full" = "Zone")

amerique_latine$region<-"amerique_latine"

europe<-read.csv("europe.csv",
                encoding = "UTF-8")%>%
  select("country_full" = "Zone")

europe$region<-"europe"

asie<-read.csv("asie.csv",
                encoding = "UTF-8")%>%
  select("country_full" = "Zone")

asie$region<-"asie"

afrique<-read.csv("afrique.csv",
                encoding = "UTF-8")%>%
  select("country_full" = "Zone")

afrique$region<-"afrique"

fao_region<-rbind(afrique,asie,europe,amerique_latine)
```

country	country_full	year_survey	quantile	nb_quantiles	income	gdpppp	pop	gini	region
ALB	Albanie	2008	1	100	728.8980	7297	3002.678	30.5	europe
ALB	Albanie	2008	2	100	916.6623	7297	3002.678	30.5	europe
ALB	Albanie	2008	3	100	1010.9160	7297	3002.678	30.5	europe
ALB	Albanie	2008	4	100	1086.9078	7297	3002.678	30.5	europe
ALB	Albanie	2008	5	100	1132.6997	7297	3002.678	30.5	europe

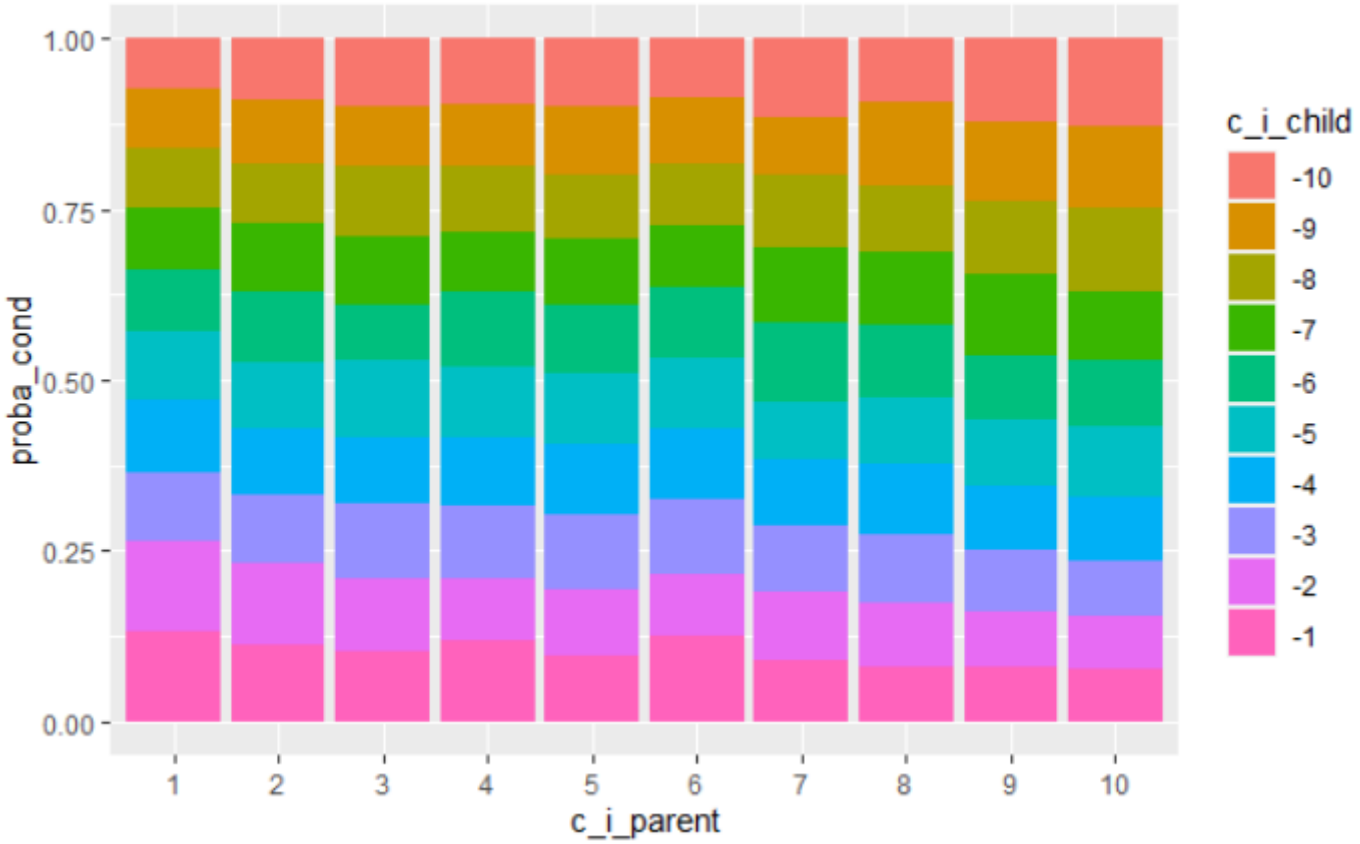
Création de la table des probabilité conditionnelles

$$ln(Y_{child}) = \alpha + p_j \, ln(Y_{parent}) + \epsilon$$

	c_i_child	c_i_parent	n	proba_cond
1	1	1	253	0.253
2	1	2	139	0.139
3	1	3	92	0.092
4	1	4	63	0.063
5	1	5	54	0.054
6	1	6	46	0.046
7	1	7	31	0.031
8	1	8	23	0.023
9	1	9	31	0.031
10	1	10	24	0.024
11	1	11	22	0.022
12	1	12	14	0.014
13	1	13	20	0.020
14	1	14	20	0.020
15	1	15	15	0.015

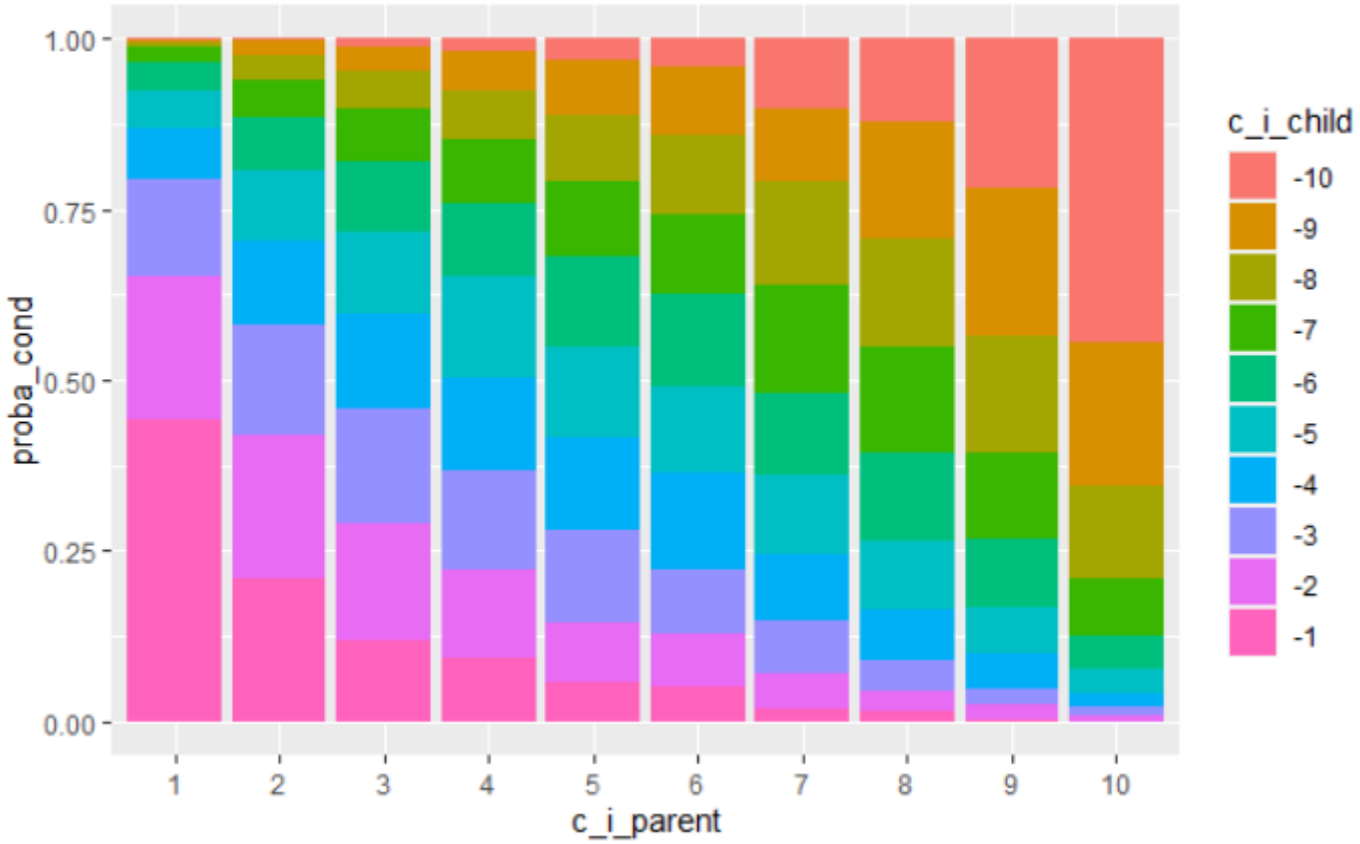
Graphique distribution conditionnelle

Nombre de quantile = 10 / p = 0.1



Graphique distribution conditionnelle

Nombre de quantile = 10 / p = 0.9



Conclusion

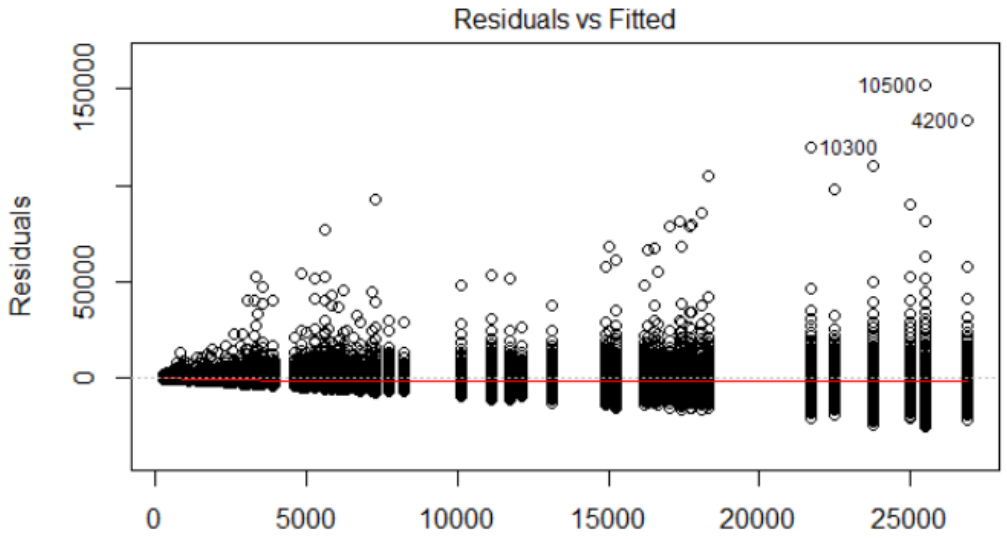
Un coefficient d’élasticité élevé nous permet donc de mieux prévoir le revenu de l’enfant, quand on connait le revenu des parents.
Il sera donc sans doute plus facile de faire nos prédiction sur des pays ayant un coefficient d’élasticité élevé.

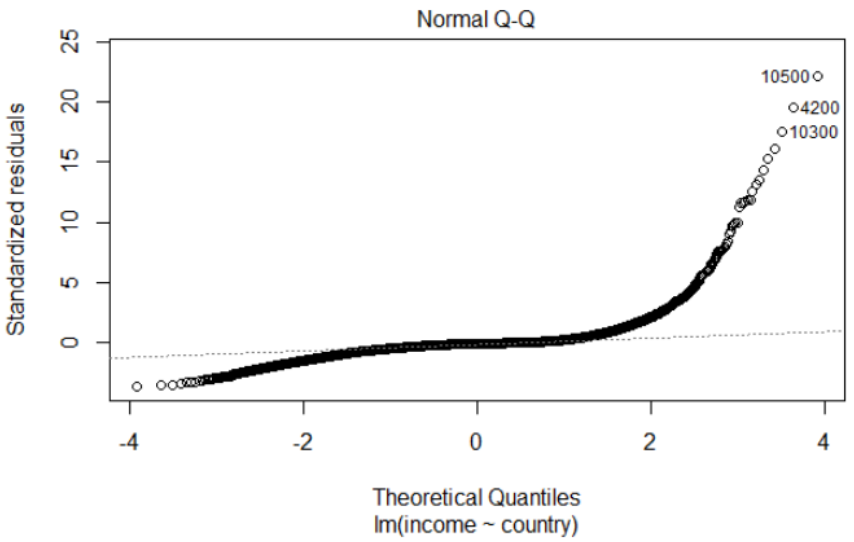
Mission 4

Création du modèle prédictif

ANOVA income ~ country

```
Call:
lm(formula = income ~ country, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-24840  -1907   -393    493 151425

Residual standard error: 6859 on 10989 degrees of freedom
Multiple R-squared:  0.491,     Adjusted R-squared:  0.4859
F-statistic: 96.38 on 110 and 10989 DF,  p-value: < 2.2e-16
```

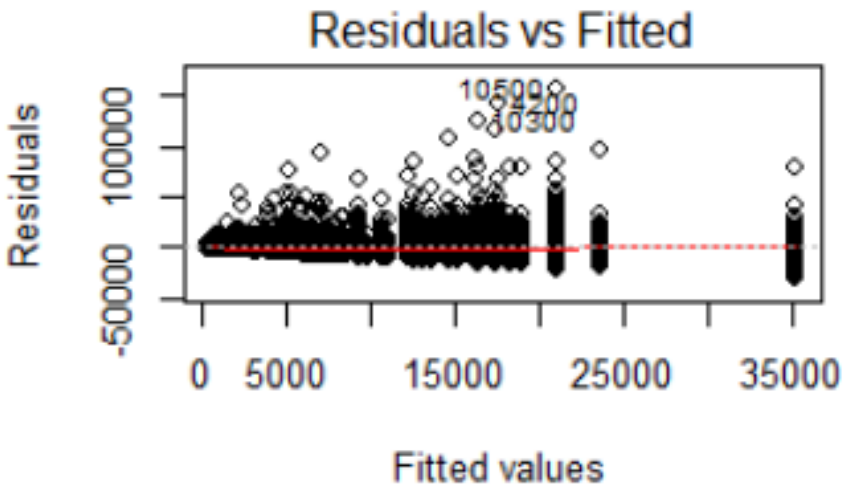


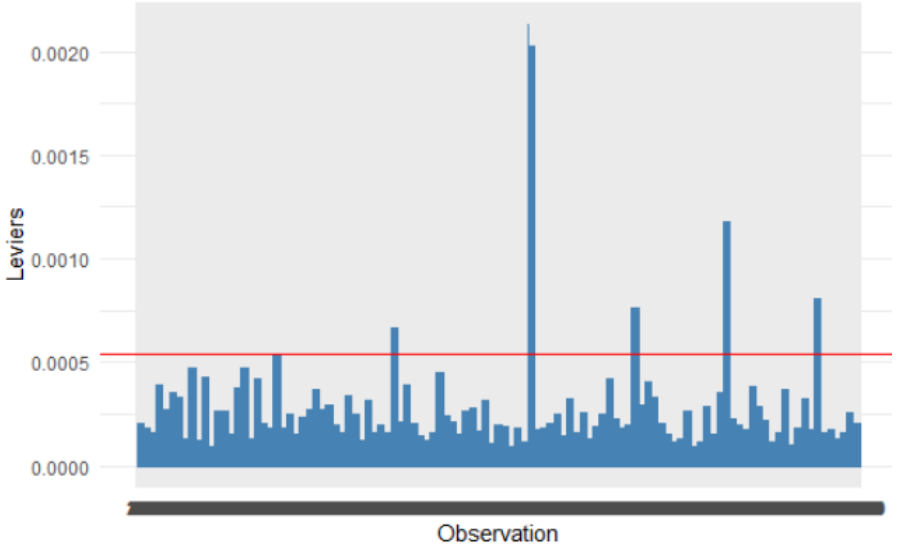


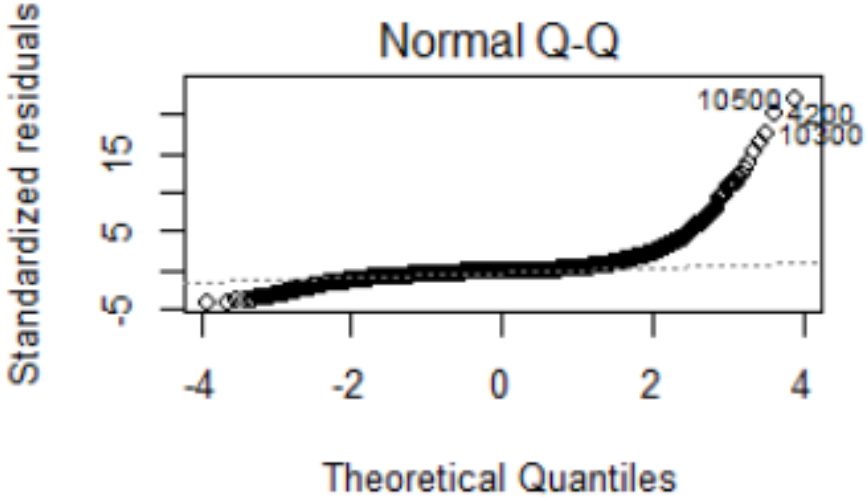
Régression linéaire : income ~ gini + gdpppp

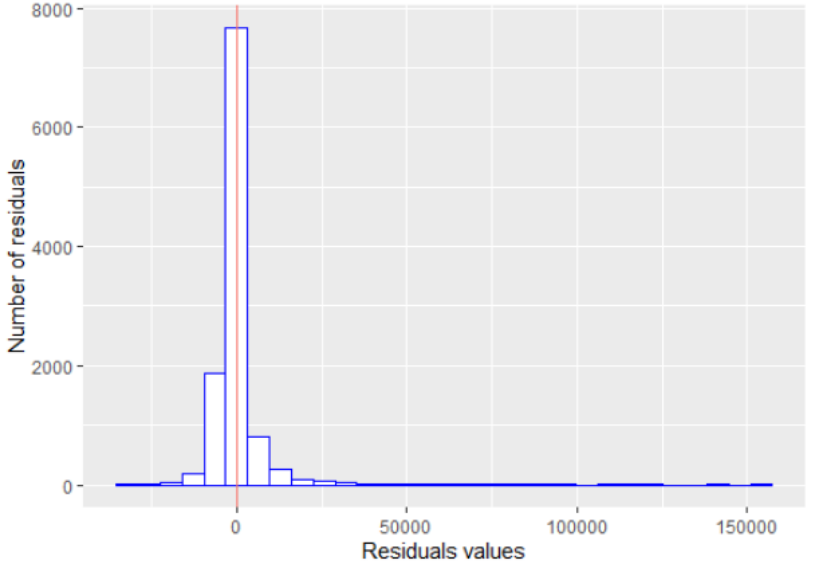
```
Call:
lm(formula = income ~ gdpppp + gini, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-30365  -2408   -479    568 155861

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.992e+02  3.541e+02  -1.410   0.1587
gdpppp       4.823e-01  5.559e-03  86.756   <2e-16 ***
gini         1.622e+01  8.287e+00   1.957   0.0503 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7147 on 11097 degrees of freedom
Multiple R-squared:  0.4419,    Adjusted R-squared:  0.4418
F-statistic:  4394 on 2 and 11097 DF,  p-value: < 2.2e-16
```

Régression linéaire : income ~ gini + gdpppp





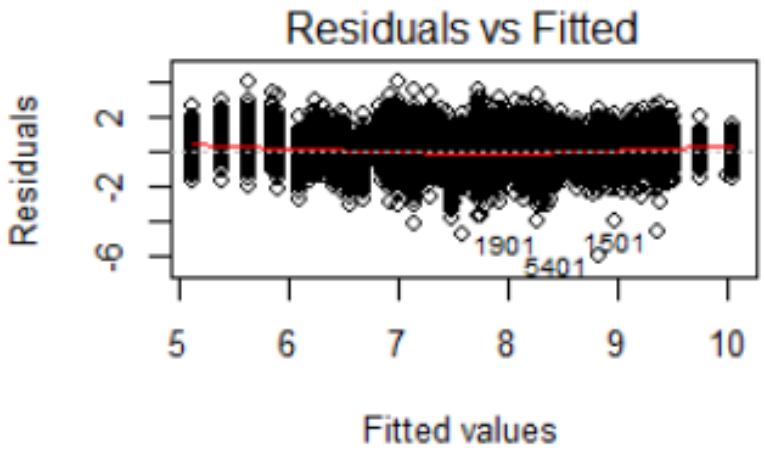


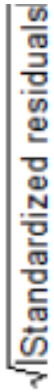


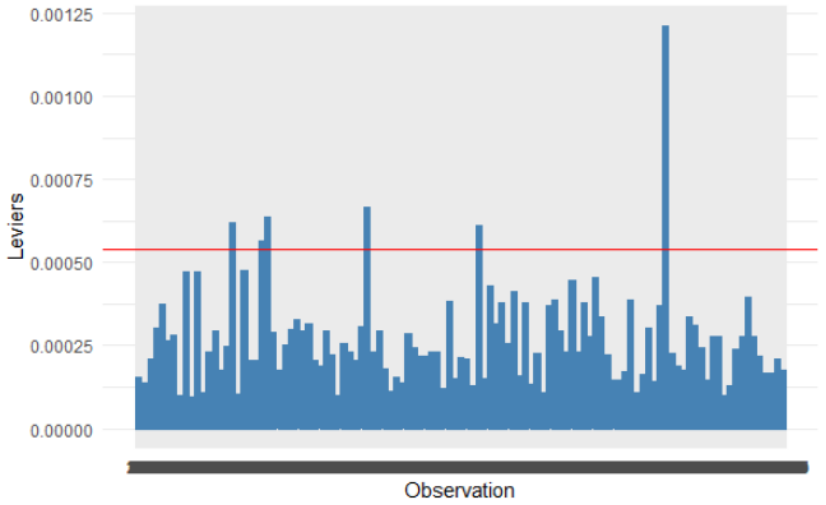
Régression linéaire : log(income) ~ gini + log(gdpppp)

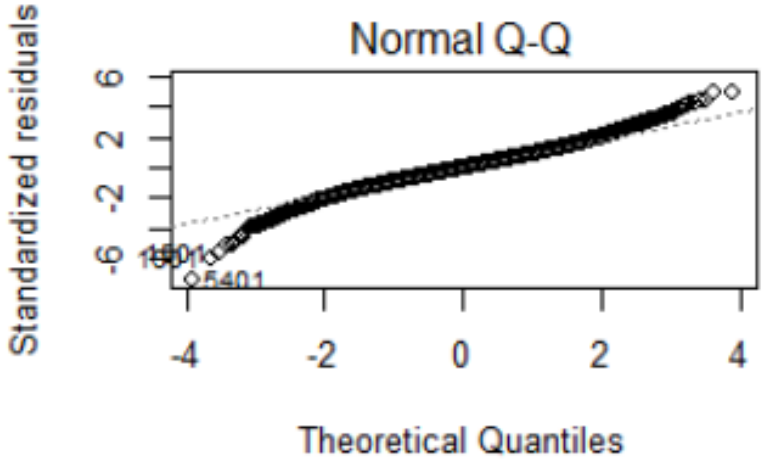
```
Call:
lm(formula = log(income) ~ log(gdpppp) + gini, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9678 -0.4844  0.0043  0.4911  3.9460

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.8801190  0.0740364   11.89   <2e-16 ***
log(gdpppp)  0.8597772  0.0063634  135.11   <2e-16 ***
gini        -0.0153011  0.0008967  -17.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8051 on 11097 degrees of freedom
Multiple R-squared:  0.6631,    Adjusted R-squared:  0.663
F-statistic: 1.092e+04 on 2 and 11097 DF,  p-value: < 2.2e-16
```

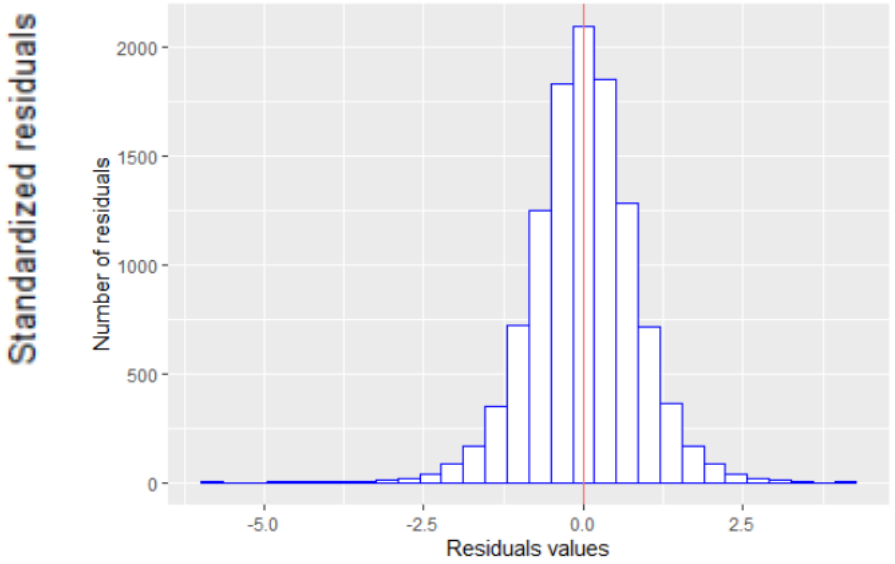
Régression linéaire : income ~ gini + log(gdpppp)







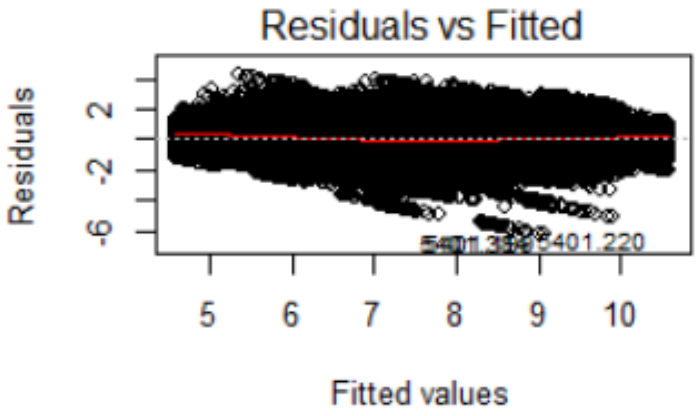


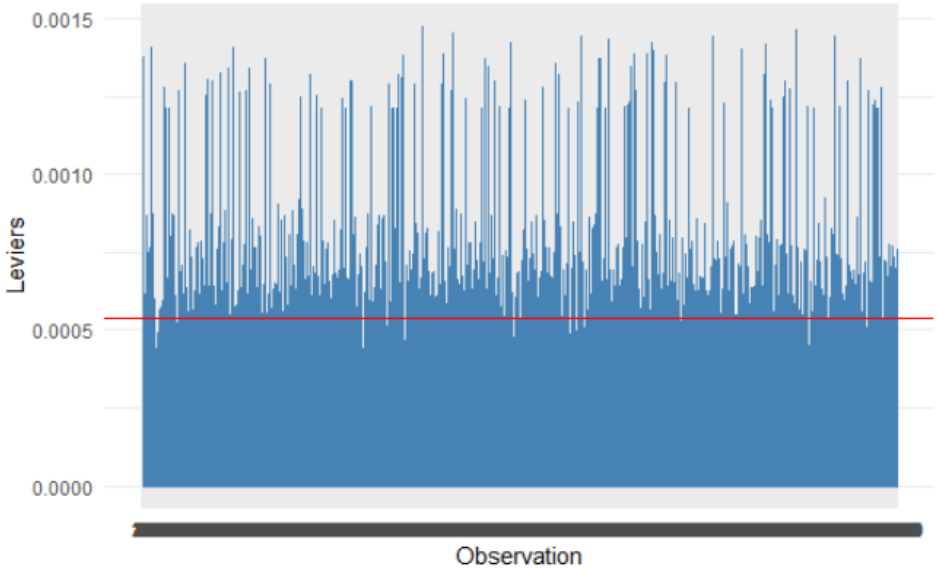


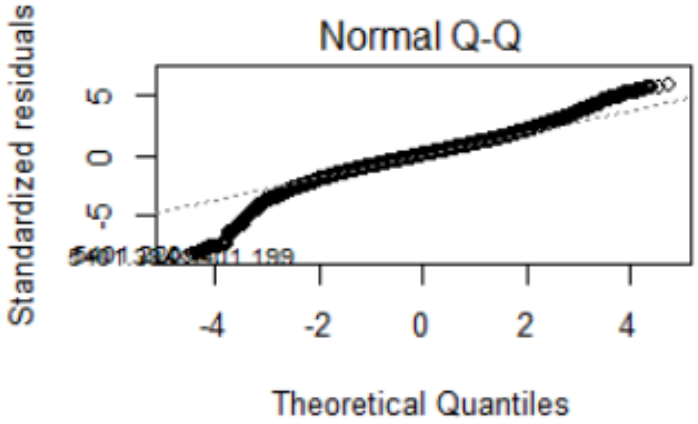
Régression linéaire :
Log(income) ~ log(gdpppp) + gini + c_i_parent

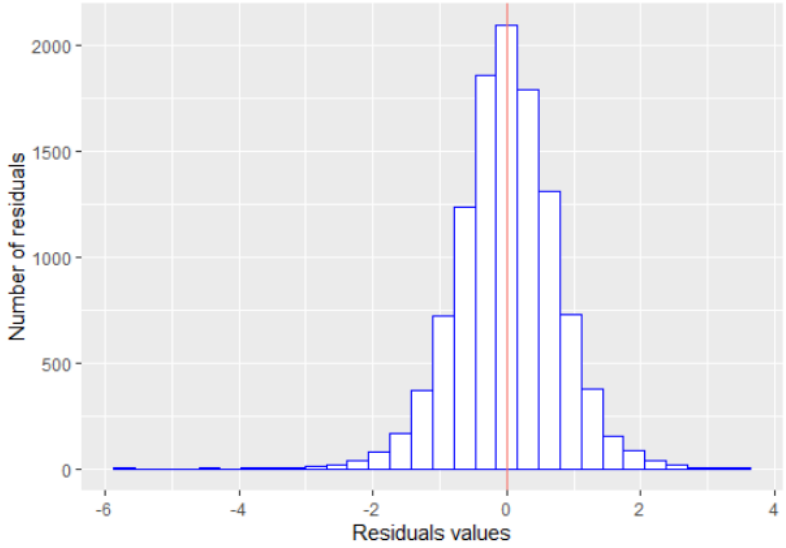
```
Call:
lm(formula = log(income) ~ log(gdpppp) + gini + c_i_parent, data = df500)

Residuals:
    Min      1Q  Median      3Q     Max
-6.4452 -0.4582  0.0055  0.4656  4.3483

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.385e-01  3.105e-03   109.0   <2e-16 ***
log(gdpppp)  8.598e-01  2.626e-04  3273.6   <2e-16 ***
gini        -1.530e-02  3.701e-05  -413.5   <2e-16 ***
c_i_parent   1.073e-02  1.093e-05   982.1   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.743 on 5549996 degrees of freedom
Multiple R-squared:  0.7129,    Adjusted R-squared:  0.7129
F-statistic: 4.595e+06 on 3 and 5549996 DF,  p-value: < 2.2e-16
```

Régression linéaire :
Log(income) ~ log(gdpppp) + gini + c_i_parent









Conclusion

En incluant la **classe de revenu des parents, l'analyse des résidus est** sensiblement la même que sur le modèle précédent (ils suivent une loi normale et sont globalement de même variance) mais on gagne **5 points** sur le coefficient de détermination pour atteindre **0.7129.**

La variance totale est expliquée à **71%** par le pays de naissance et le revenu des parents et à **19%** par les autres facteurs non considérés dans le modèle.

Un indice de **gini plus élevé, défavorise plus de personne qu'il n'en favorise.** Ceci est mis en évidence par le coefficient négatif devant l'indice au sein du modèle.

Bilan

Ce projet m’a permis de m’améliorer au niveau de la récupération et du traitement des données. J’en ai aussi appris beaucoup sur les hypothèses de validités des modèles de régression linéaires.

J’ai trouvé la mission 3 un peu confuse.
Il manque un petit texte pour expliquer quelle est le but des 12 opérations.