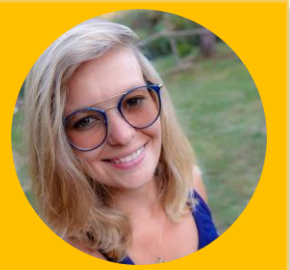
OPENCLASSROOMS



Lena Verboom

Prédiction de revenus

Projet 7 : Effectuer une prédiction de revenus

Contexte

- Mission pour une banque internationale

- **Cibler des nouveaux jeunes clients** susceptibles d'avoir de hauts revenus

- **Stratégie** : Créer un **modèle** pour déterminer le revenu potentiel d'une personne

- Analyse des caractéristiques **des revenus** (Pays, années, quantiles, gdpppp…)

Construction et interprétation du modèle

Données à ma disposition :

World Income Distribution

Revenu, quantiles et gdpppp par pays

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728,89795	7297
1	ALB	2008	2	100	916,66235	7297

116 pays

Data_projet7.csv

Population mondiale par pays en 2008

	country_name	Population_2008
0	Afghanistan	27722276
1	Albania	2947314

264 pays

Population.csv

THE WORLD BANK

Indice de Gini

	country_name	country	gini_2004	gini_2006	gini_2007	gini_2008	gini_2009	gini_2010	gini_2011
0	Afghanistan	AFG	..	..	..	..	..	..	..
1	Albania	ALB	..	..	..	30	..	..	..

264 pays

Gini.csv

Jointure des dataframes :

	country	year_survey	quantile	nb_quantiles	income	gdpppp	country_name	gini_2004	gini_2006	gini_2007	gini_2008	gini_2009	gini_2010	gini_2011	Population_2008
0	ALB	2008	1	100	728,89795	7297	Albania	..	..	..	30	..	..	..	2947314.0
1	ALB	2008	2	100	916,66235	7297	Albania	..	..	..	30	..	..	..	2947314.0
2	ALB	2008	3	100	1010,916	7297	Albania	..	..	..	30	..	..	..	2947314.0
3	ALB	2008	4	100	1086,9078	7297	Albania	..	..	..	30	..	..	..	2947314.0
4	ALB	2008	5	100	1132,6997	7297	Albania	..	..	..	30	..	..	..	2947314.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11594	COD	2008	96	100	810,6233	303,19305	Congo, Dem. Rep.	42.2	..	..	..	..	..	..	60411195.0
11595	COD	2008	97	100	911,7834	303,19305	Congo, Dem. Rep.	42.2	..	..	..	..	..	..	60411195.0
11596	COD	2008	98	100	1057,8074	303,19305	Congo, Dem. Rep.	42.2	..	..	..	..	..	..	60411195.0
11597	COD	2008	99	100	1286,6029	303,19305	Congo, Dem. Rep.	42.2	..	..	..	..	..	..	60411195.0
11598	COD	2008	100	100	2243,1226	303,19305	Congo, Dem. Rep.	42.2	..	..	..	..	..	..	60411195.0

11599 rows × 15 columns

116 pays

Jointure de type left sur data_projet7 sur country

Visualisation des valeurs NA :

```
country              0
year_survey          0
quantile             0
nb_quantiles         0
income               0
gdpppp             200
country_name       100
gini_2004          100
gini_2006          100
gini_2007          100
gini_2008          100
gini_2009          100
gini_2010          100
gini_2011          100
Population_2008    100
dtype: int64
```

Remplacement des NA gdpppp de Kosovo et West Bank and Gaza

country	year_survey	quantile	nb_quantiles	income	gdpppp	country_name
XKX	2008	1	100	437,8937	NaN	Kosovo
PSE	2009	100	100	6343,8755	NaN	West Bank and Gaza

country	country_name	gini_2004	gini_2006	gini_2007	gini_2008	gini_2009	gini_2010	gini_2011	Population_2008
TWN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Remplacement des NA de Taïwan

Transformation des variables en numérique :

```
country              object
year_survey           int64
quantile              int64
nb_quantiles          int64
income               object
gdpppp               object
country_name         object
gini_2004            object
gini_2006            object
gini_2007            object
gini_2008            object
gini_2009            object
gini_2010            object
gini_2011            object
Population_2008     float64
dtype: object
```

Remplacement des , par des .

Remplacement des .. par NA

Transformation en numérique

```
country              object
year_survey           int64
quantile              int64
nb_quantiles          int64
income              float64
gdpppp              float64
country_name         object
gini_2004           float64
gini_2006           float64
gini_2007           float64
gini_2008           float64
gini_2009           float64
gini_2010           float64
gini_2011           float64
Population_2008     float64
dtype: object
```

Recherche de valeurs anormales :

	year_survey	quantile	nb_quantiles	income	gdpppp
count	11599.000000	11599.000000	11599.0	11599.000000	1.159900e+04
mean	2007.982757	50.500819	100.0	6069.224260	4.944979e+04
std	0.909633	28.868424	0.0	9414.185972	3.966471e+05
min	2004.000000	1.000000	100.0	16.719418	3.031931e+02
25%	2008.000000	25.500000	100.0	900.685515	2.577000e+03
50%	2008.000000	51.000000	100.0	2403.244900	7.505000e+03
75%	2008.000000	75.500000	100.0	7515.420900	1.838850e+04
max	2011.000000	100.000000	100.0	176928.550000	4.300332e+06

Fiji : Valeur anormale

	country	year_survey	quantile	nb_quantiles	income	gdpppp
3200	FJI	2008	1	100	308.17334	4300332.0

Remplacement de la valeur

Recherche de données manquantes :

Groupby country : count nb_quantiles

country	year_survey	country_name	Population_2008	quantile	income	gdpppp	nb_quantiles
ALB	2008	Albania	2947314.0	5050	2994.829902	7297.000	100
ARG	2008	Argentina	40080160.0	5050	5847.884654	13220.000	100
ARM	2008	Armenia	2907618.0	5050	1628.382785	5611.000	100

Sélection du nombre de quantile inférieure à 100

country	year_survey	country_name	Population_2008	quantile	income	gdpppp	nb_quantiles
LTU	2008	Lithuania	3198231.0	5009	6641.247634	17571.0	99

Quantile numéro 41 absent

Remplacement de la valeur par la moyenne entre le quartile 40 et 42

Ajout du coefficient d'élasticité :

	countryname	IGEincome
0	Afghanistan	NaN
1	Albania	0.815874
2	Angola	NaN
3	Argentina	NaN
4	Armenia	NaN
...	...	...
145	Venezuela, RB	NaN
146	Vietnam	0.480000
147	West Bank and Gaza	NaN
148	Yemen, Rep.	NaN
149	Zambia	NaN

Remplacement des valeurs NA

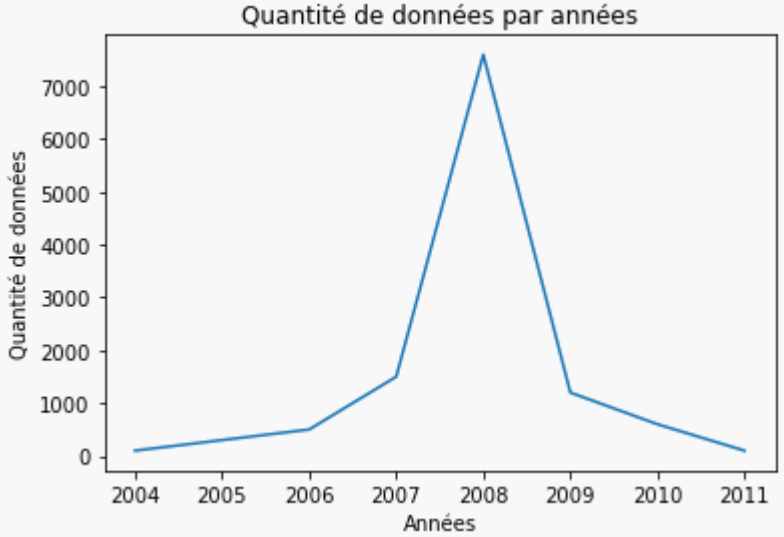
```
Nordic_european_countries_and_canada=0.2
Europe=0.4
Australia_New_Zealand_USA=0.4
Asia=0.5
Latin_America_Africa=0.66
```

	country_name	IGEincome
0	Afghanistan	0.500000
1	Albania	0.815874
2	Angola	0.660000
3	Argentina	0.660000
4	Armenia	0.660000
...	...	...
145	Venezuela, RB	0.660000
146	Vietnam	0.480000
147	West Bank and Gaza	0.500000
148	Yemen, Rep.	0.500000
149	Zambia	0.660000

Données finales :

	country	year_survey	quantile	nb_quantiles	income	gdpppp	country_name	gini_2004	gini_2006	gini_2007	gini_2008	gini_2009	gini_2010	gini_2011	Population_2008	IGEincome
0	ALB	2008	1	100	728.89795	7297.00000	Albania	NaN	NaN	NaN	30.0	NaN	NaN	NaN	2947314.0	0.815874
1	ALB	2008	2	100	916.66235	7297.00000	Albania	NaN	NaN	NaN	30.0	NaN	NaN	NaN	2947314.0	0.815874
2	ALB	2008	3	100	1010.91600	7297.00000	Albania	NaN	NaN	NaN	30.0	NaN	NaN	NaN	2947314.0	0.815874
3	ALB	2008	4	100	1086.90780	7297.00000	Albania	NaN	NaN	NaN	30.0	NaN	NaN	NaN	2947314.0	0.815874
4	ALB	2008	5	100	1132.69970	7297.00000	Albania	NaN	NaN	NaN	30.0	NaN	NaN	NaN	2947314.0	0.815874
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
11595	COD	2008	96	100	810.62330	303.19305	Congo, Dem. Rep.	42.2	NaN	NaN	NaN	NaN	NaN	NaN	60411195.0	0.707703
11596	COD	2008	97	100	911.78340	303.19305	Congo, Dem. Rep.	42.2	NaN	NaN	NaN	NaN	NaN	NaN	60411195.0	0.707703
11597	COD	2008	98	100	1057.80740	303.19305	Congo, Dem. Rep.	42.2	NaN	NaN	NaN	NaN	NaN	NaN	60411195.0	0.707703
11598	COD	2008	99	100	1286.60290	303.19305	Congo, Dem. Rep.	42.2	NaN	NaN	NaN	NaN	NaN	NaN	60411195.0	0.707703
11599	COD	2008	100	100	2243.12260	303.19305	Congo, Dem. Rep.	42.2	NaN	NaN	NaN	NaN	NaN	NaN	60411195.0	0.707703

Années des données utilisées :



Nombre de pays : **116**

year_survey	
2004	100
2006	500
2007	1500
2008	7599
2009	1200
2010	600
2011	100

Population couverte par l'analyse :

$$\frac{Pop_{étude}}{Pop\_mondiale} = \frac{6.2 \times 10^9}{6.7 \times 10^9} = \mathbf{91.7}\ \%$$

Type de quantiles :

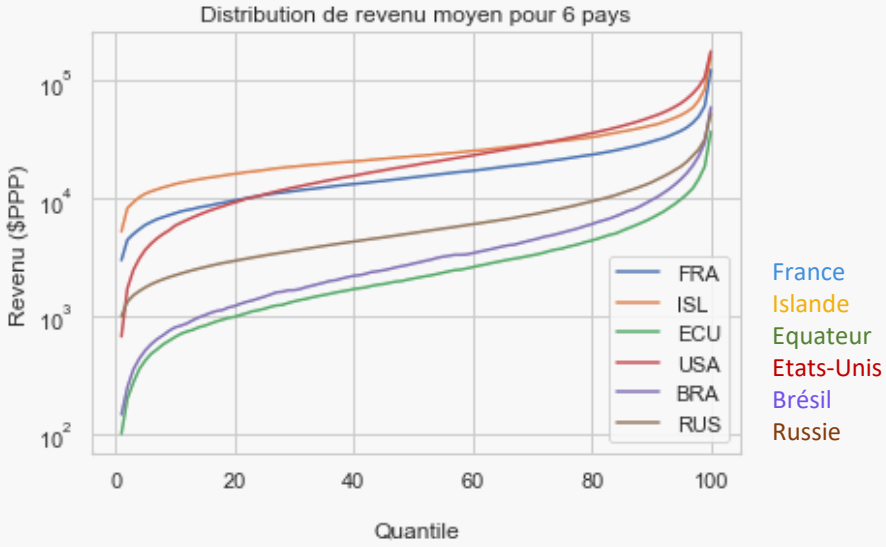
	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728,89795	7297
1	ALB	2008	2	100	916,66235	7297

→ Centiles

Echantillonnage par quantiles →

- Bonne méthode
- Echantillons représentatifs de la population
- Facilite la représentation
- Centiles

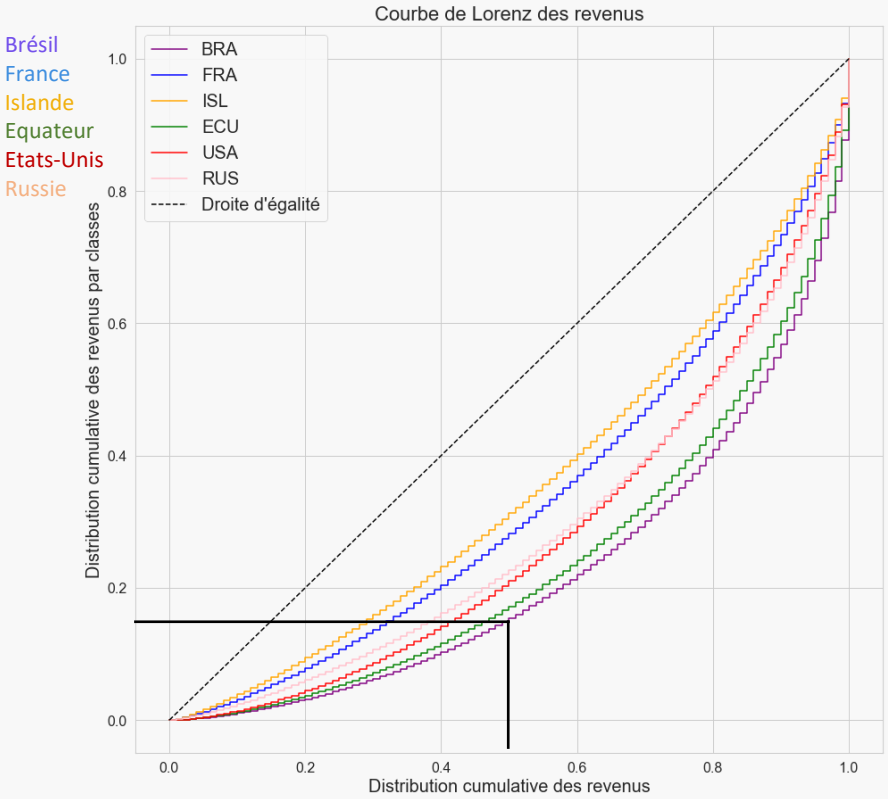
Diversité des pays en terme de revenus :



France
Islande
Equateur
Etats-Unis
Brésil
Russie

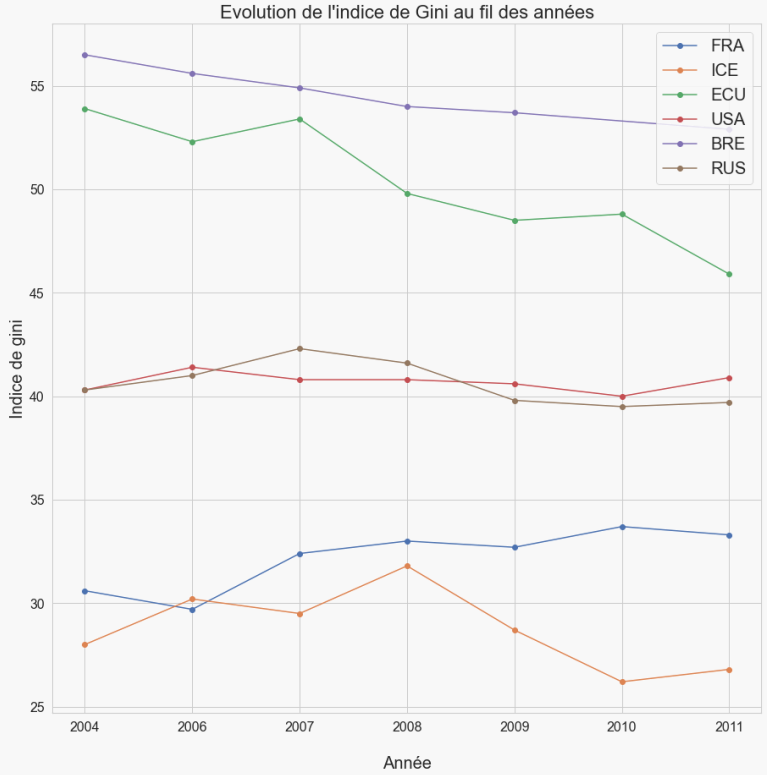
Distribution de revenu différente entre chaque pays : diversité

Courbe de Lorenz pour chaque pays choisi :



Plus d'inégalité de revenus pour certains pays : Brésil et équateur

Evolution de l'indice de Gini au fil des années :



France
Islande
Equateur
Etats-Unis
Brésil
Russie

Evolution de l'indice de Gini différente pour chaque pays

Classement des pays par indice de Gini :

Pays avec l'indice de Gini le plus faible

Azerbaijan	26.600000
Czech Republic	26.528571
Slovak Republic	26.371429
Denmark	26.200000
Slovenia	24.557143

Pays avec l'indice de Gini le plus élevé

South Africa	63.200000
Central African Republic	56.200000
Honduras	55.357143
Guatemala	54.600000
Brazil	54.600000

France →

	country_name	gini
79	France	32.2

→ **Position 80/116**

Préparation des données pour la mission 3 :

Renommer les colonnes

quantile → c_i_child
income → y_child
élasticité → pj
gini → Gj

	country	country_name	c_i_child	y_child	pj	Gj
0	ALB	Albania	1	728.89795	0.815874	0.321410
1	ALB	Albania	2	916.66235	0.815874	0.321410
2	ALB	Albania	3	1010.91600	0.815874	0.321410
3	ALB	Albania	4	1086.90780	0.815874	0.321410
4	ALB	Albania	5	1132.69970	0.815874	0.321410
...	...	...	...	...	...	...
11595	COD	Congo, Dem. Rep.	96	810.62330	0.707703	0.459403
11596	COD	Congo, Dem. Rep.	97	911.78340	0.707703	0.459403
11597	COD	Congo, Dem. Rep.	98	1057.80740	0.707703	0.459403
11598	COD	Congo, Dem. Rep.	99	1286.60290	0.707703	0.459403
11599	COD	Congo, Dem. Rep.	100	2243.12260	0.707703	0.459403

Calcul de l'indice de Gini à l'aide de la courbe de lorenz

Présentation des fonctions nécessaires :

```
def generate_incomes(n, pj):
```

Génération : d'un grand nombre de réalisation, de la variable ln_y_parent et du terme d'erreur selon une loi normale

```
def quantiles(l, nb_quantiles):
def compute_quantiles(y_child, y_parents, nb_quantiles):
```

Génération de c_i_child et c_i_parent à partir de y_child et y_parents

```
def distribution(counts, nb_quantiles):
def conditional_distributions(sample, nb_quantiles):
```

Estimation de la distribution conditionnelle de c_i_child (comptage de la combinaison c_i_child et c_i_parent, division nb individu compté et nb total pour chaque quantile parent)

```
def proba_cond(c_i_parent, c_i_child, mat):
```

Détermination de la probabilité de c_i_parent avec c_i_child et la distribution conditionnelle

Création d'un nouvel échantillon à partir de nos données finales :

Création de 499 clones pour chaque individu

	country	country_name	c_i_child	y_child	pj	Gj
0	ALB	Albania	1	728.89795	0.815874	0.321410
1	ALB	Albania	2	916.66235	0.815874	0.321410
2	ALB	Albania	3	1010.91600	0.815874	0.321410
3	ALB	Albania	4	1086.90780	0.815874	0.321410
4	ALB	Albania	5	1132.69970	0.815874	0.321410
...	...	...	...	...	...	...
5788395	COD	Congo, Dem. Rep.	96	810.62330	0.707703	0.459403
5788396	COD	Congo, Dem. Rep.	97	911.78340	0.707703	0.459403
5788397	COD	Congo, Dem. Rep.	98	1057.80740	0.707703	0.459403
5788398	COD	Congo, Dem. Rep.	99	1286.60290	0.707703	0.459403
5788399	COD	Congo, Dem. Rep.	100	2243.12260	0.707703	0.459403

5800000 rows × 6 columns

Création de 499 clones pour chaque individu

Attribution des classes :

```
def loop(n,nb_quantiles,pj_pays):
    y_child,y_parents=generate_incomes(n,pj_pays)
    sample=compute_quantiles(y_child,y_parents,nb_quantiles)
    cd=conditional_distributions(sample,nb_quantiles)
    return cd
```

Génération d'une boucle pour la détermination de y_parents et la distribution conditionnelle

```
pays_list=data_final['country'].unique()
```

Stockage des noms des pays

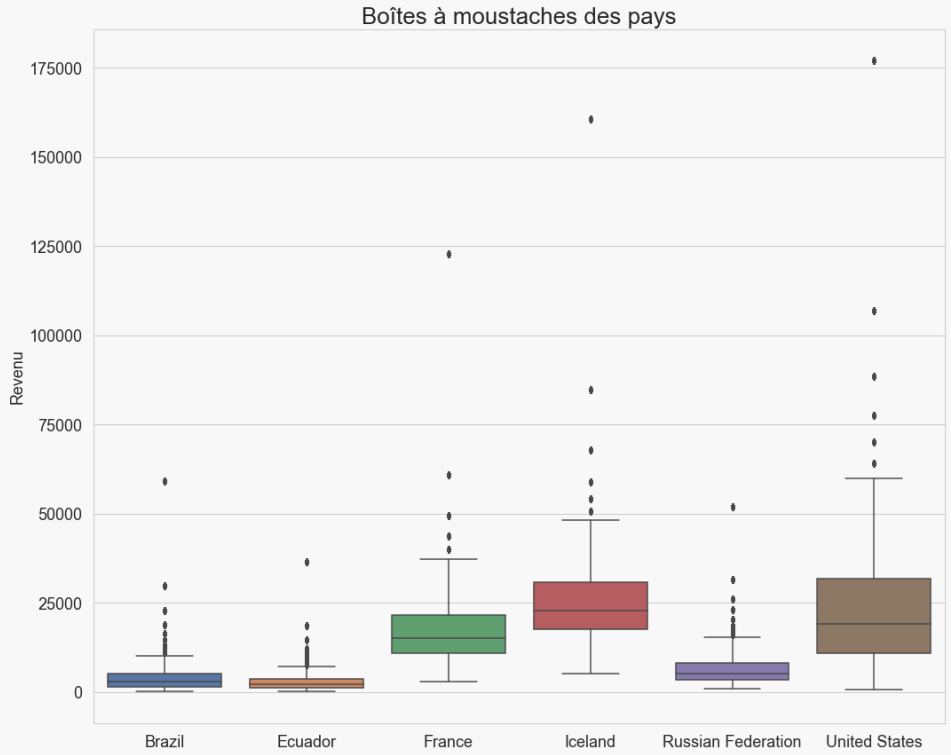
```
for pays in pays_list:
    pj=data_final.loc[data_final['country']==pays,'pj'].iloc[0]
    proba=loop(n,nb_quantiles,pj)

    for i in range(100):
        proba_i=proba[i]
        quantiles_p=np.random.choice(element, 500, p = proba_i)
        quantiles_total.append(quantiles_p)
quantiles_parents_final=[j for quantile in quantiles_total for j in quantile]
data_final=pd.concat([data_final,pd.Series(quantiles_parents_final)],axis=1)
```

Génération des probabilités avec le coefficient d'élasticité puis des quantiles parents

Données nécessaires pour la mission 4 :

	country	country_name	y_child	pj	Gj	ln_y_child	c_i_parent
0	ALB	Albania	728.89795	0.815874	0.321410	6.591534	7
1	ALB	Albania	728.89795	0.815874	0.321410	6.591534	11
2	ALB	Albania	728.89795	0.815874	0.321410	6.591534	2
3	ALB	Albania	728.89795	0.815874	0.321410	6.591534	12
4	ALB	Albania	728.89795	0.815874	0.321410	6.591534	42
...	...	...	...	...	...	...	...
5799995	ZAF	South Africa	82408.55000	0.677000	0.682949	11.319444	89
5799996	ZAF	South Africa	82408.55000	0.677000	0.682949	11.319444	97
5799997	ZAF	South Africa	82408.55000	0.677000	0.682949	11.319444	100
5799998	ZAF	South Africa	82408.55000	0.677000	0.682949	11.319444	38
5799999	ZAF	South Africa	82408.55000	0.677000	0.682949	11.319444	17

Représentation des revenus en fonction du pays :



Les revenus sont plus élevés pour la France, l'Islande et les USA

Vérification de la loi normale :

Homoscédasticité : équivalence des variances

Test levene
p-value =0.0

H0 : Les variances de population sont égales

Ha : Au moins une des variances est différente

p value < 0.05 : Il y a une différence entre les variances dans la population

Normalité : distribution normale

Test shapiro
p-value =0.0

H0 : L'échantillon est normalement distribué

Ha : L'échantillon n'est pas normalement distribué

p value < 0.05 : L'échantillon ne suit pas une loi normale

Aucune des conditions de l'ANOVA paramétrique est remplie

Distribution équivalente entre les échantillons

Test Kruskal-Wallis
p-value =0.0

H0 : La médianes des populations sont égales

Ha : Les médianes des populations sont différentes

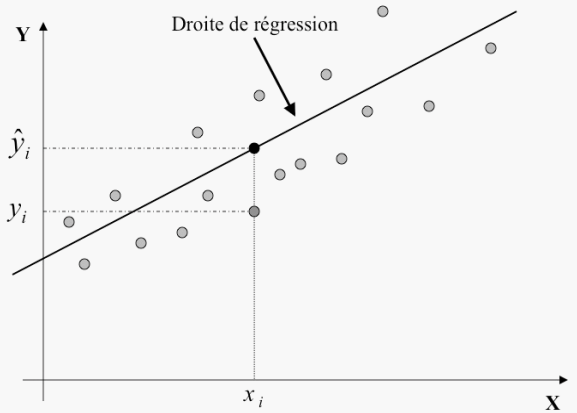
p value < 0.05 : Au moins un échantillon est différent

→ **ANOVA non paramétrique**

Corrélation entre les deux variables

Régression linéaire :

Méthode statistique de prédiction



Une variable (X) expliquée est modélisée par une fonction affine d'une autre variable (y)

Technique d'analyse

OLS : Ordinary Least Squares

OLS Regression Results

Dep. Variable:	y_child	R-squared:	0.496
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	2.858e+06
Date:	Wed, 21 Jul 2021	Prob (F-statistic):	0.00
Time:	06:56:07	Log-Likelihood:	-5.9310e+07
No. Observations:	5800000	AIC:	1.186e+08
Df Residuals:	5799997	BIC:	1.186e+08
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.192e-09	14.699	8.11e-11	1.000	-28.809	28.809
mj	1.0000	0.000	2234.874	0.000	0.999	1.001
Gj	2.237e-10	33.529	6.67e-12	1.000	-65.715	65.715

Omnibus:	7299082.828	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2103715018.034
Skew:	6.739	Prob(JB):	0.00
Kurtosis:	95.322	Cond. No.	1.18e+05

Comparaison de la différence entre les points du data set et les prédictions afin de mesurer l'erreur

Anova sur la régression linéaire entre le revenu et le pays :

OLS Regression Results

Dep. Variable:	y_child	**R-squared:**	0.492
Model:	OLS	**Adj. R-squared:**	0.492
Method:	Least Squares	**F-statistic:**	4887.
Date:	Thu, 22 Jul 2021	**Prob (F-statistic):**	0.00
Time:	07:18:21	**Log-Likelihood:**	-5.9362e+06
No. Observations:	580000	**AIC:**	1.187e+07
Df Residuals:	579884	**BIC:**	1.187e+07
Df Model:	115		
Covariance Type:	nonrobust		

Explique à 50% la variabilité des revenus à partir du pays

	sum_sq	df	F	PR(>F)	EtaSq
country_name	2.553966e+13	115.0	4886.995541	0.0	0.492171
Residual	2.635217e+13	579884.0	NaN	NaN	NaN

H0 : Les variables sont indépendantes

Ha : Les variables sont dépendantes

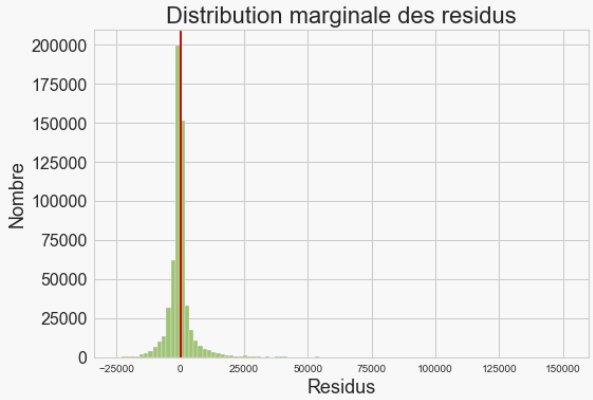
p value < 0.05 : Les moyennes sont significativement différentes entres les pays

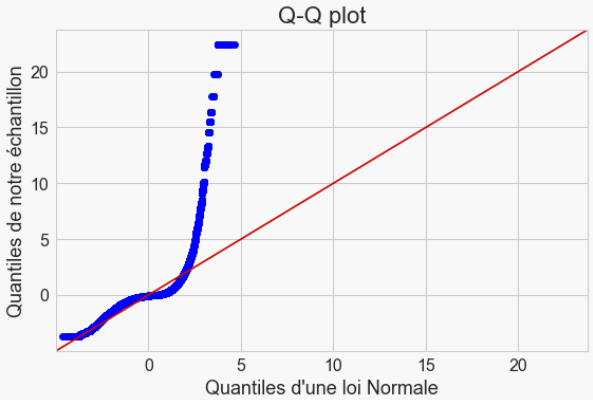
Dépendance du pays de l'individu et du revenu

Test sur les résidus :

Normalité : distribution normale

Résidus : différence entre la valeur observée et la valeur prédite du modèle





Test Kolmogorov-Smirnov
p-value =0.0

H0 : Les échantillons suivent une distribution normale

Ha : Les échantillons ne suivent pas une distribution normale

p value < 0.05 : Les échantillons ne suivent pas une loi normale

Les résidus ne suivent pas une loi normale

Test sur les résidus :

Test d'homoscédasticité

Test Goldfeld et Quandt
p-value =0.98

H0 : Les échantillons possèdent une variance égales

Ha : Les échantillons possèdent des variances différentes

p value > 0.05 : Il y a homoscédasticité des résidus

Test de corrélation

Test Durbin-Watson
r =2.00

H0 : Les résidus ne sont pas auto-corrélés

Ha : Les résidus sont auto-corrélés

p value = 2 : Il n'y a pas d'auto-corrélation

Régression linéaire de y_child sur le revenu moyen du pays et Gj

OLS Regression Results

Dep. Variable:	y_child	R-squared:	0.496
Model:	OLS	Adj. R-squared:	0.496
Method:	Least Squares	F-statistic:	2.858e+06
Date:	Thu, 22 Jul 2021	Prob (F-statistic):	0.00
Time:	07:18:31	Log-Likelihood:	-5.9310e+07
No. Observations:	5800000	AIC:	1.186e+08
Df Residuals:	5799997	BIC:	1.186e+08
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.192e-09	14.699	8.11e-11	1.000	-28.809	28.809
mj	1.0000	0.000	2234.874	0.000	0.999	1.001
Gj	2.237e-10	33.529	6.67e-12	1.000	-65.715	65.715

Omnibus:	7299082.828	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2103715018.034
Skew:	6.739	Prob(JB):	0.00
Kurtosis:	95.322	Cond. No	1.18e+05

Explique à 50% la variabilité des revenus enfants à partir du revenu moyen et le Gini

Coefficients de la fonction affine

Significativité : P-value < 0.05

Le modèle prend uniquement en compte le mj

Mj et Gj ajoute la même information au modèle

Indique la multicolinéarité : plusieurs variables donnent la même information

Ce modèle prédit le revenu moyen comme le revenu

Régression linéaire de ln(y_child) sur le ln(mj) et Gj :

OLS Regression Results

Dep. Variable:	ln_y_child	R-squared:	0.729
Model:	OLS	Adj. R-squared:	0.729
Method:	Least Squares	F-statistic:	7.793e+06
Date:	Thu, 22 Jul 2021	Prob (F-statistic):	0.00
Time:	07:18:46	Log-Likelihood:	-6.3181e+06
No. Observations:	5800000	AIC:	1.264e+07
Df Residuals:	5799997	BIC:	1.264e+07
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.4961	0.003	170.994	0.000	0.490	0.502
ln_mj	0.9864	0.000	3651.055	0.000	0.986	0.987
Gj	-1.6523	0.004	-471.886	0.000	-1.659	-1.645

Omnibus:	372790.535	Durbin-Watson:	0.001
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1752378.640
Skew:	-0.081	Prob(JB):	0.00
Kurtosis:	5.688	Cond. No.	116.

Explique à 73% la variabilité des revenus enfants à partir du revenu moyen et le Gini

Toutes les variables sont significatives

Plus de multicolinéarité : les variables apportent toute une information différente

Ce modèle prédit le revenu en prenant en compte toutes les variables

Régression linéaire de y_child sur le mj, Gj et le c_i_parent :

OLS Regression Results

Dep. Variable:	y_child	R-squared:	0.521
Model:	OLS	Adj. R-squared:	0.521
Method:	Least Squares	F-statistic:	2.106e+06
Date:	Thu, 22 Jul 2021	Prob (F-statistic):	0.00
Time:	07:19:00	Log-Likelihood:	-5.9162e+07
No. Observations:	5800000	AIC:	1.183e+08
Df Residuals:	5799996	BIC:	1.183e+08
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-2605.1437	15.089	-172.651	0.000	-2634.718	-2575.570
mj	0.9999	0.000	2292.355	0.000	0.999	1.001
Gj	-0.8288	32.685	-0.025	0.980	-64.890	63.232
c_i_parent	51.6097	0.094	550.820	0.000	51.426	51.793

Omnibus:	7407327.164	Durbin-Watson:	0.084
Prob(Omnibus):	0.000	Jarque-Bera (JB):	2321599296.992
Skew:	6.900	Prob(JB):	0.00
Kurtosis:	100.037	Cond. No.	1.18e+05

Explique à 52 % la variabilité des revenus enfants à partir des revenus moyen, le Gini et la classe des parents

Le modèle ne prend pas en compte le Gini

Les variables apportent la même information

Ce modèle prédit le revenu sans prendre en compte l'indice de Gini

Régression linéaire de ln(y_child) sur le ln(mj), Gj et le c_i_parent :

OLS Regression Results

Dep. Variable:	ln_y_child	R-squared:	0.781
Model:	OLS	Adj. R-squared:	0.781
Method:	Least Squares	F-statistic:	6.890e+06
Date:	Thu, 22 Jul 2021	Prob (F-statistic):	0.00
Time:	07:19:18	Log-Likelihood:	-5.6996e+06
No. Observations:	5800000	AIC:	1.140e+07
Df Residuals:	5799996	BIC:	1.140e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.0546	0.003	-20.603	0.000	-0.060	-0.049
ln_mj	0.9863	0.000	4061.510	0.000	0.986	0.987
Gj	-1.6522	0.003	-524.969	0.000	-1.658	-1.646
c_i_parent	0.0109	9.3e-06	1174.213	0.000	0.011	0.011

Omnibus:	390634.811	Durbin-Watson:	0.375
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1843285.013
Skew:	-0.128	Prob(JB):	0.00
Kurtosis:	5.750	Cond. No.	824.

Explique à 78 % la variabilité des revenus enfants à partir du revenu moyen, le Gini et la classe des parents

Toutes les variables sont significatives

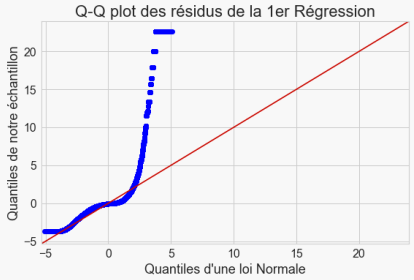
Plus de multicolinéarité : les variables apportent toute une information différente

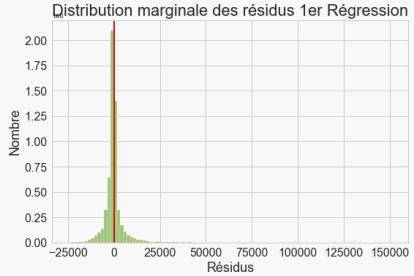
Ce modèle prédit le revenu en prenant en compte toutes les variables

Test sur les résidus des différentes régressions :

Normalité : distribution normale

1ère Régression

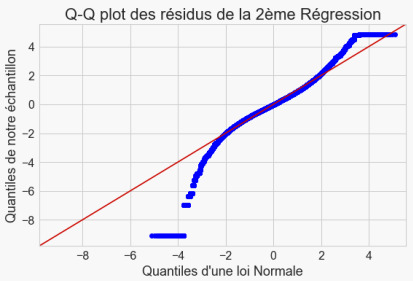


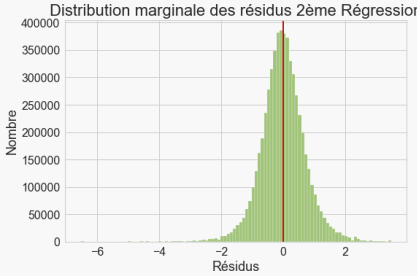


Test Jarque - Bera

p-value =0.0

2ème Régression

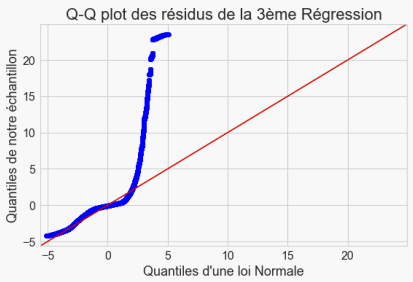


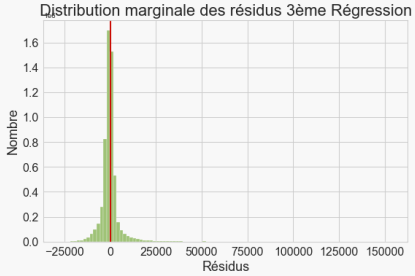


Test Jarque - Bera

p-value =0.0

3ème Régression

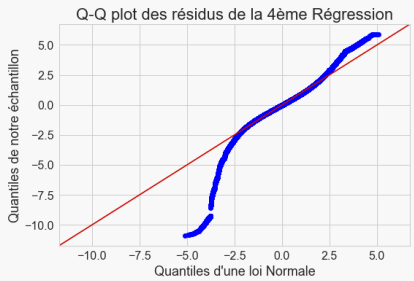


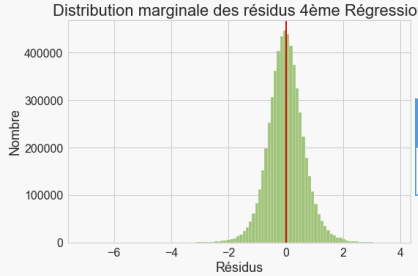


Test Jarque - Bera

p-value =0.0

4ème Régression





Test Jarque - Bera

p-value =0.0

H0 : Les échantillons suivent une distribution normale

Ha : Les échantillons ne suivent pas une distribution normale

p value < 0.05 : Les échantillons ne suivent pas une loi normale

Test sur les résidus des différentes régressions :

Test d'homoscédasticité

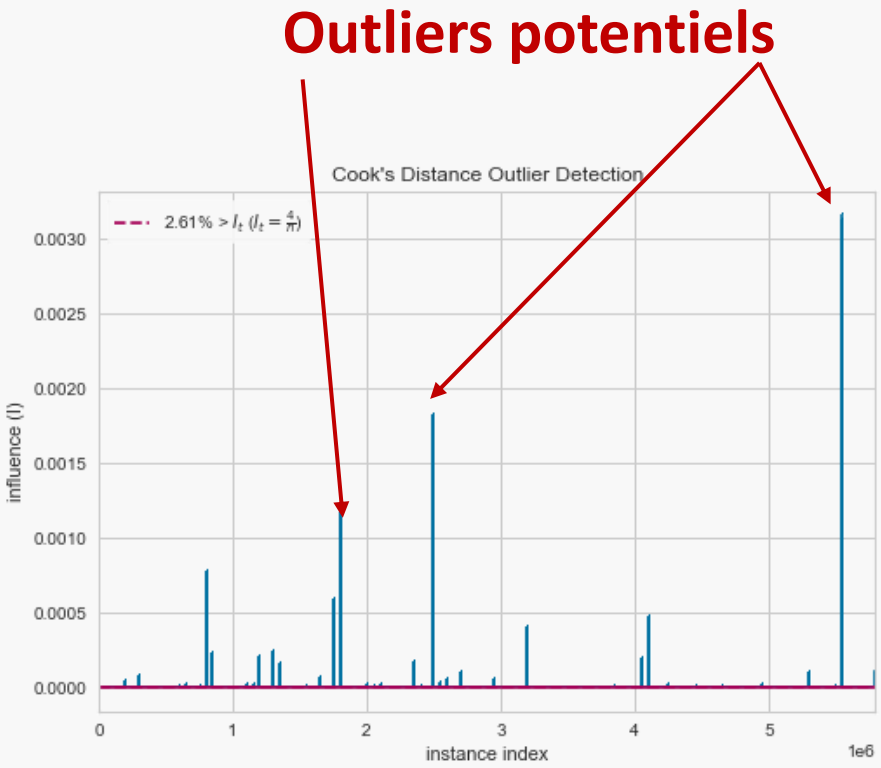
1ère Régression	2ème Régression	3ème Régression	4ème Régression
Test Goldfeld et Quandt	Test Goldfeld et Quandt	Test Goldfeld et Quandt	Test Goldfeld et Quandt
p-value =0.99	p-value =2 e-10	p-value =0.99	p-value =0.99

H0 : Les échantillons possèdent des variances égales

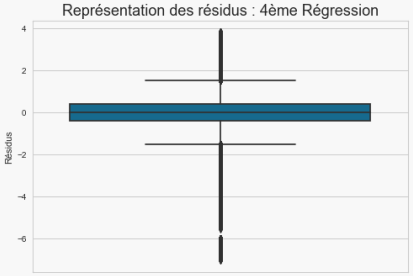
Ha : Les échantillons possèdent des variances différentes

Vérification que les erreurs restent constantes peu importe la valeur que prend la variable explicative

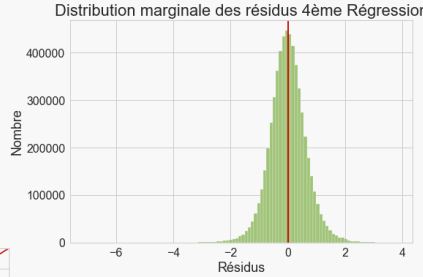
Suppression des outliers :

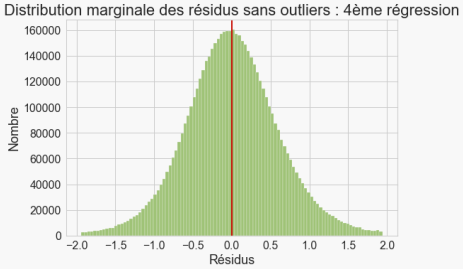


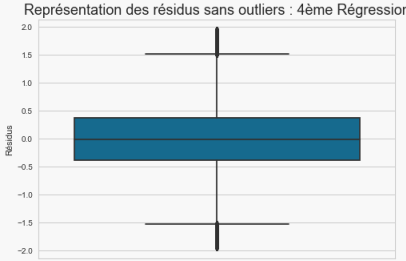
Suppression des outliers :

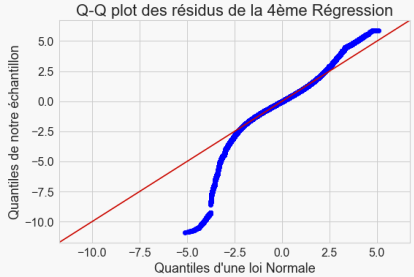


Détermination du Z-score : sélection des scores < 3

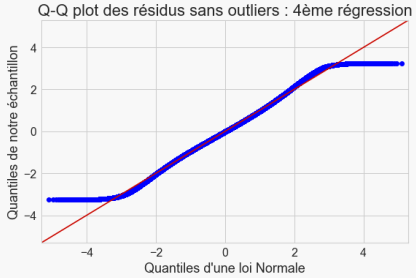








Amélioration de la normalité



	Variance totale expliquée	**Variable significative**	**Multicolinéarité**
1ère Régression	50 %	Mj	Oui
2ème Régression	73 %	Const, ln(Mj), Gj	Non
3ème Régression	52 %	Const, Mj, Gj, c_i_parent	Oui
4ème Régression	78 %	Const, ln(Mj), Gj, c_i_parent	Non

- La 4ème régression linéaire a permis **d'expliquer 78 % de la variabilité** des revenus
- L'expression en **logarithme** a permis d'**améliorer** le modèle et diminuer la **multicolinéarité**
- L'ajout de **la classe parent** a amélioré la **variance expliquée**

- L'ensemble des variables **ne permettent pas d'expliquer à 100 % la variabilité** de revenu

Le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu'il en défavorise?

	coef
const	-0.0546
ln_mj	0.9863
Gj	-1.6522
c_i_parent	0.0109

- Plus le **Gini augmente** plus il y a une **inégalité**
- Gini : **coefficient négatif**

- Plus il y a un **Gini important** plus les personnes **sont défavorisées**




Merci pour votre attention