OPENCLASSROOMS – PARCOURS DATA ANALYST
PROJET 7

PRÉVOIR DES REVENUS

JÉRÔME CHALAIN

PLAN

Contexte

Mission 1
Résumé des données utilisées

Mission 2
Distribution des revenus
- diversité des pays
- courbes de Lorenz
- indices de Gini

Mission 3
Distributions conditionnelles
- coefficients d'élasticité
- Génération de la classe $c_{(i,parent)}$

Mission 4
Modélisation

Explication du revenu des individus en fonction de plusieurs variables explicatives : le pays, l'indice de Gini de ce pays, la classe de revenus des parents, etc.

Conclusion

CONTEXTE

CONTEXTE

- Banque présente dans de nombreux pays.
- **Objectif :** cibler de nouveaux prospects, plus particulièrement les jeunes en âge d'ouvrir leur tout premier compte bancaire et, parmi ceux-ci, les plus susceptibles d'avoir, plus tard dans leur vie, de hauts revenus.
- **Mission :** créer un modèle permettant de déterminer le revenu potentiel d'une personne.
- **Proposition :** réaliser une régression linéaire avec 3 variables :
  1. le revenu des parents ;
  2. le revenu moyen du pays dans lequel habite le prospect ;
  3. l'indice de Gini calculé sur les revenus des habitants du pays en question.

MISSION 1
RÉSUMÉ DES DONNÉES UTILISÉES

RÉSUMÉ DES DONNÉES UTILISÉES

Les données sources :

1. distributions de revenus des populations de certains pays (source *World Income Distribution* via OCR) → table income

2. Populations des pays du monde (source FAO) → table pop

3. indices de Gini* estimés par la Banque mondiale → table gini

** L'indice de Gini est un indicateur synthétique permettant de rendre compte du niveau d'inégalité pour une variable et sur une population donnée. Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême). Cf. insee.fr*

RÉSUMÉ DES DONNÉES UTILISÉES

Table income : présentation

```
inc.head()
```

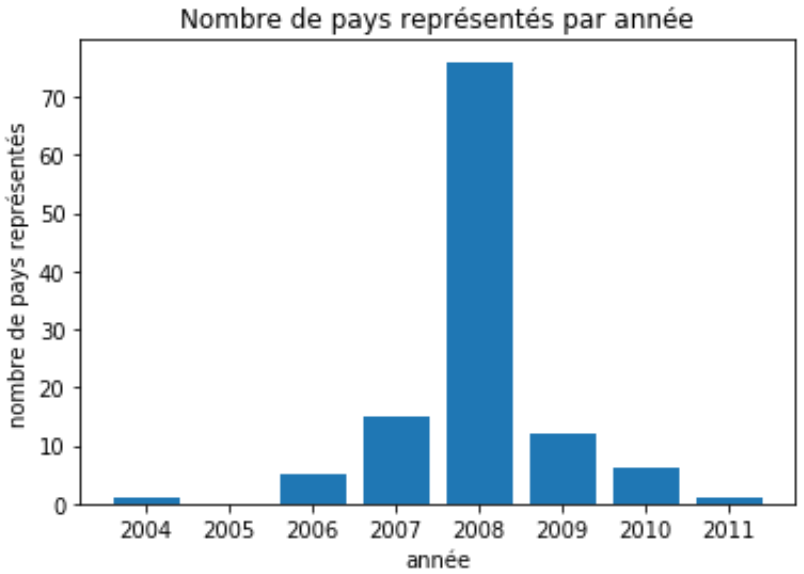
	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.0

116 pays, 100 lignes/pays.
Nous avons les codes des pays mais pas leurs noms.

```
inc.groupby('country_code').nunique().sum()
```

```
country_code       116
year               116
quantile         11599
nb_quantiles       116
income           11599
gdpppp             114
```

Il manque visiblement 1 ligne quantile (et donc *income*) et 2 lignes gdpppp*.



7 années (de 2004 à 2011, rien en 2005).
L'année 2008 est la plus représentée.

Gdp ppp
gross domestic product (based on) purchasing power parity.
En français : produit intérieur brut en parité du pouvoir d'achat (PIB PPA). Permet de pouvoir faire des comparaisons entre pays sans distorsion due au taux de change.

RÉSUMÉ DES DONNÉES UTILISÉES

Table income : traitements

- Quantile manquant : qtl 41 de la Lituanie. On l'évalue en faisant la moyenne avec les quantiles précédent et suivant.

6239	LTU	2008	40	100	4868.4507	17571.0	Lituanie
6240	LTU	2008	42	100	4895.8306	17571.0	Lituanie

```
(float(inc2.iloc[6239]['income']) +  float(inc2.iloc[6240]['income'])) / 2
```

```
4882.14065
```

- Utilisation de quantiles (centiles) = bonne méthode car compromis entre réduction de la taille de l'échantillon et préservation de l'information. Permet notamment d'effectuer des comparaisons relativement précises entre pays.

RÉSUMÉ DES DONNÉES UTILISÉES

Table income : traitements

- gdp ppp manquants : Palestine et Kosovo → complété avec des données de la Banque mondiale.

```
nb_pays = inc.groupby('country_code').nunique()
nb_pays[nb_pays['gdpppp']==0]
```

country_code	country_code	year	quantile	nb_quantiles	income	gdpppp
PSE	1	1	100	1	100	0
XKX	1	1	100	1	100	0

```
inc.loc[inc['country_code'] == 'PSE', 'gdpppp'] = 3695.024
inc.loc[inc['country_code'] == 'XKX', 'gdpppp'] = 7249.409
```

- Les valeurs gdpppp pour les Fidji sont incohérentes → remplacées par des données de la Banque mondiale.

Fidji	2008.0	50.5	100.0	2098.730882	4300332.0

```
inc2.loc[inc2['country_code'] == 'FJI', 'gdpppp'] = 7078.627
```

RÉSUMÉ DES DONNÉES UTILISÉES

Table income : jointure avec la table population

- Jointure via le nom et l'année.

Pour cela, création d'une table temporaire avec les noms des pays et leurs codes ISO.

```
inc3 = pd.merge(inc2, pop, on = ['country_name', 'year'], how = 'left')
```

- 20 pays ont leur population non renseignée.

Cause : les noms sont différents selon les tables.
Ex : Biélorussie != Bélarus, Egypte != Égypte
→ Mise à jour via une fonction.

```
# Fonction pour mettre à jour les noms des pays

def corr_country_name(name):

    if (name == 'Bélarus'):
       return 'Biélorussie'

    elif (name == 'Bolivie (État plurinational de)'):
       return 'Bolivie'

    else:
       return name
```

(extrait de la fonction)

	country_code	year	quantile	nb_quantiles	income	gdpppp	country_name	population
0	ALB	2008	1	100	728.89795	7297.0	Albanie	3002678.0

RÉSUMÉ DES DONNÉES UTILISÉES

Table gini

- Source : Banque mondiale
- 264 lignes, 65 colonnes.
- Table avec de nombreuses données manquantes.

```
gini.head()
```

	Country Name	Country Code	Indicator Name	Indicator Code	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969	1970	1971	1972	1973	1974	1975	1976
0	Aruba	ABW	GINI index (World Bank estimate)	SI.POV.GINI	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

(extrait de la table Gini)

RÉSUMÉ DES DONNÉES UTILISÉES

Table gini : jointure avec la table income + population

- Calcul de la moyenne de l'indice de Gini par pays…

```
# Calcul des moyennes des indices pour chaque pays
gini['mean_gini'] = gini.mean(axis=1)
```

- …puis jointure

```
# Jointure pour ajouter la moyenne de l'indice de Gini par pays
inc4g = pd.merge(inc4g, gini3, left_on = 'country_code', right_on = 'Country Code', how = 'left')
inc4g.drop('Country Code', axis = 1, inplace = True)
```

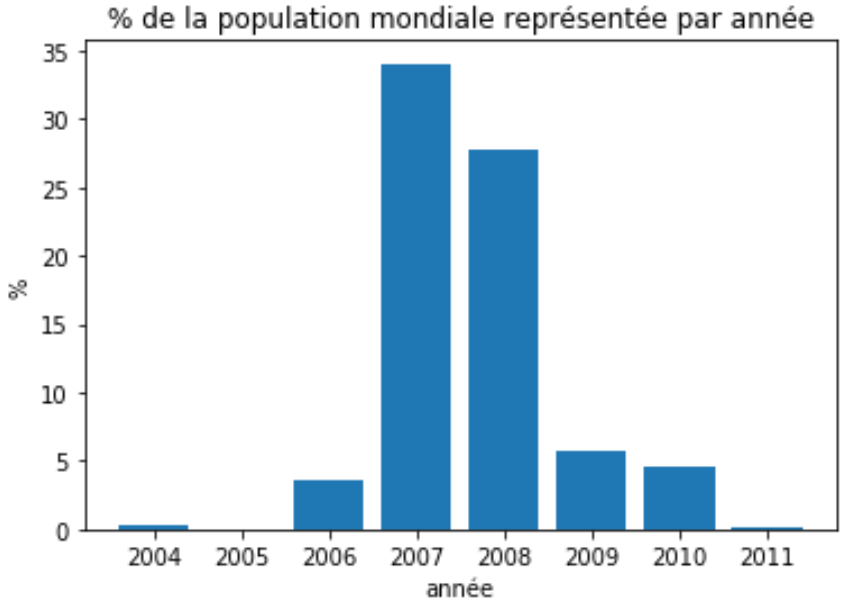
- Cambodge et Taiwan n'ont pas d'indice de Gini.
  Pas de source d'info sûre => suppression de ces lignes.

RÉSUMÉ DES DONNÉES UTILISÉES

Résultat final et population représentée par l'analyse

	country_code	year	quantile	nb_quantiles	income	gdpppp	country_name	population	mean_gini
0	ALB	2008	1	100	728.89795	7297.0	Albanie	3002678.0	31.411111

11 400 lignes, 9 colonnes



Population représentée : un peu plus de **75%** de la population mondiale recensée par la FAO (un peu + de 6 milliards de personnes).

MISSION 2
DISTRIBUTION DES REVENUS

DISTRIBUTION DES REVENUS

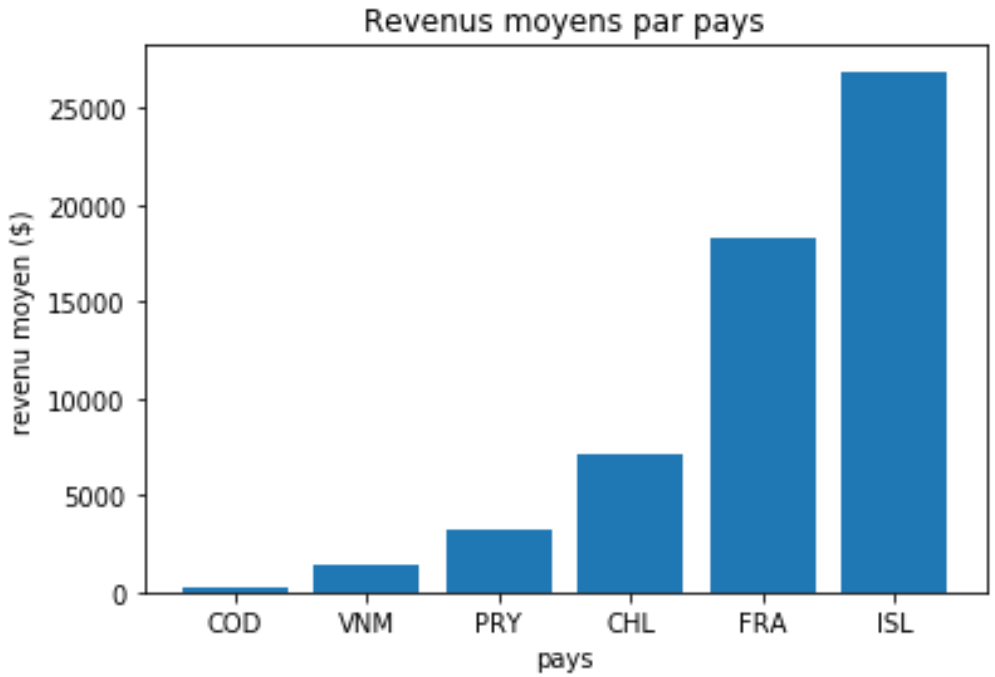
Diversité des pays en termes de distribution de revenus : choix de 6 pays représentatifs

```
# Je créé un df pour avoir la moyenne des revenus par pays, je le classe par income
inc_dr  = inc4.groupby(['country_name', 'country_code'])['income'].mean().reset_index()
inc_dr.sort_values('income').head()
```

	country_name	country_code	income
23	Congo République démocratique du	COD	276.016044

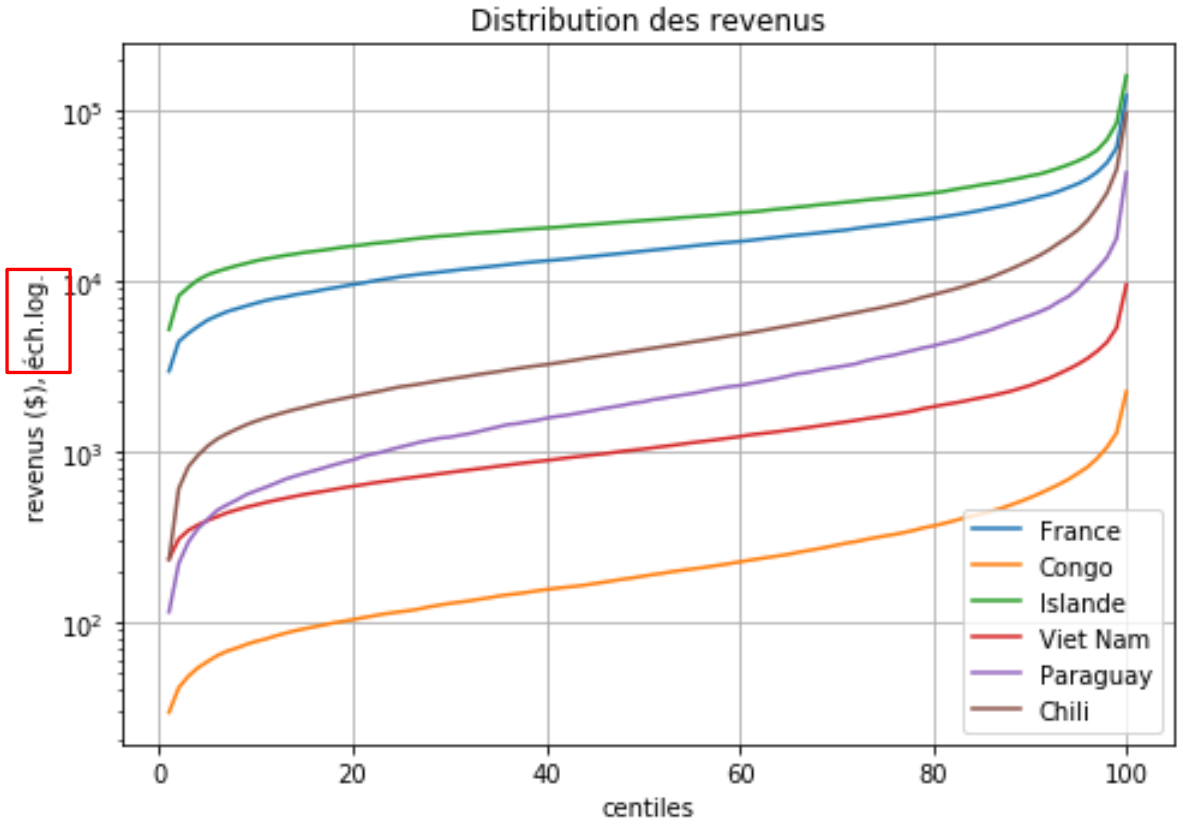
Pays retenus pour la comparaison :

- Les pays aux 2 extrêmes de la liste : Congo et Islande
- Ceux situés au niveau des quartiles : Viêt Nam, Paraguay, Chili
- La France pour avoir un point de comparaison connu



DISTRIBUTION DES REVENUS

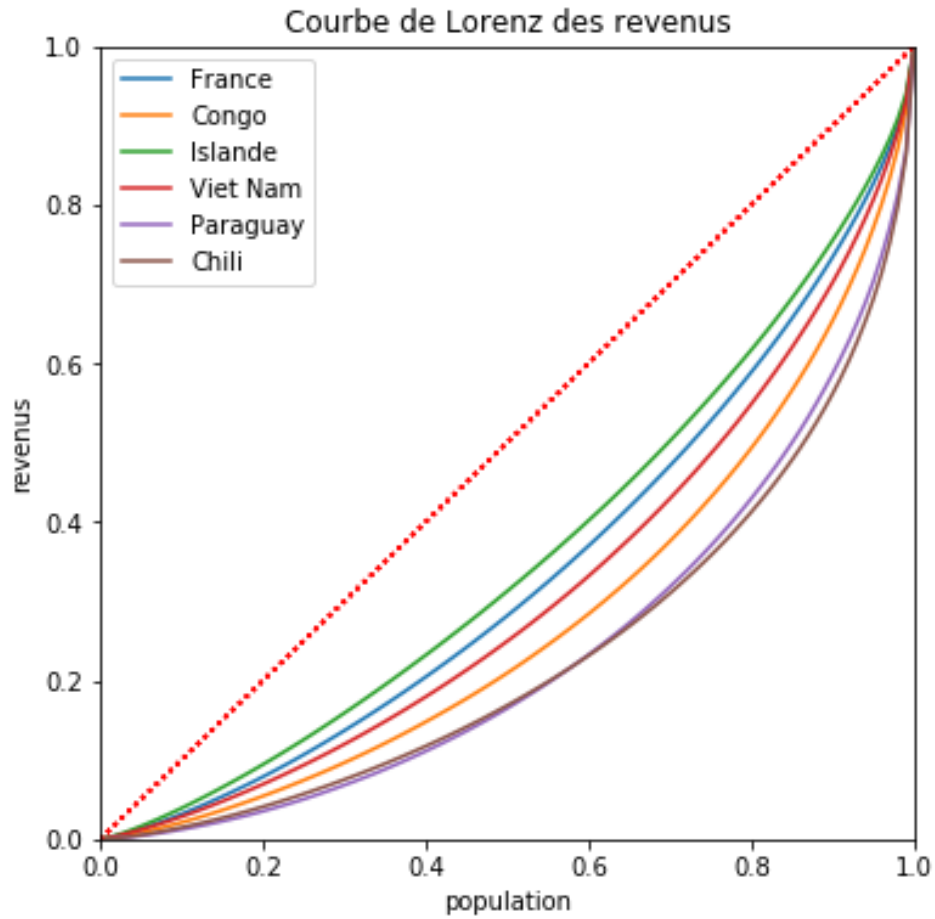
Représentation du revenu moyen de chacune des classes de revenus des pays choisis



L'Islande est visiblement le pays le plus riche du groupe et aussi le plus égalitaire.

DISTRIBUTION DES REVENUS

Courbes de Lorenz de chacun des pays choisis



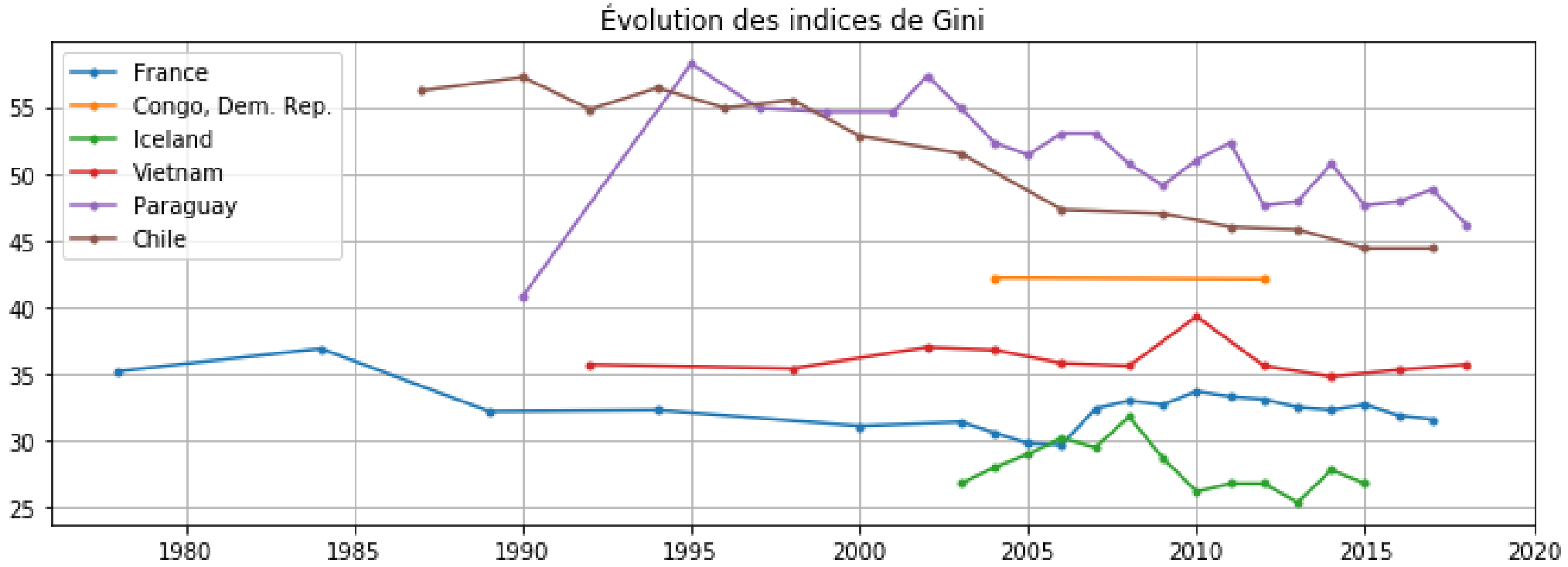
Les courbes les plus proches de la diagonale représentent les pays les plus égalitaires.

Ici : l'Islande, suivie de la France.

DISTRIBUTION DES REVENUS

Évolution de l'indice de Gini au fil des ans

Visiblement, le Paraguay et le surtout le Chili évoluent vers plus d'égalité.



Rappel : l'indice de Gini permet de rendre compte du niveau d'inégalité.

DISTRIBUTION DES REVENUS

Classement des pays par indice de Gini

La France se situe à la place n° 32 sur les 114 pays représentés.
Gini France = 0.32
Gini monde = 0.38
=> La France fait partie des pays les plus égalitaires.

Indices les plus bas (pays les plus égalitaires)

country_name	mean_gini
Slovénie	24.961111
République tchèque	25.944444
Slovaquie	26.160000
Danemark	26.405263
Finlande	26.768421

Indices les plus hauts (pays les plus inégalitaires)

country_name	mean_gini
Afrique du Sud	61.714286
Brésil	56.847059
Guatemala	55.000000
Swaziland	54.925000
Honduras	54.641379

MISSION 3
DISTRIBUTIONS CONDITIONNELLES

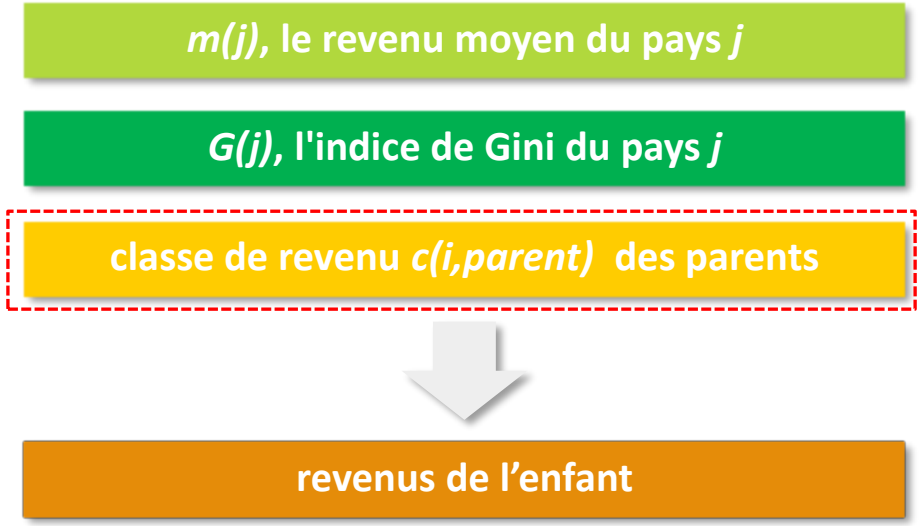
Coefficients d'élasticité

Point d'étape :

Nous avons à disposition 2 des 3 variables explicatives souhaitées :
1. ***m(j)***, le revenu moyen du pays *j*
2. ***G(j)***, l'indice de Gini du pays *j*

Il nous manque donc, pour un individu *i*, la classe de revenu ***c(i,parent)*** de ses parents.

Nous allons simuler cette information grâce à un coefficient $\rho(j)$ (propre à chaque pays) mesurant une corrélation entre le revenu de l'individu *i* et le revenu de ses parents…



Coefficients d’élasticité

…ce coefficient sera ici appelé coefficient d'élasticité. Il mesure la mobilité intergénérationnelle du revenu.

Précisions :

L'élasticité intergénérationnelle des revenus permet d'estimer la variation en % du revenu d'un enfant lorsque les revenus des parents augmentent de 1%.

Une parfaite mobilité sociale (= aucune influence de la situation familiale sur le revenu des enfants) implique donc une élasticité égale à 0. Plus l'élasticité est grande, plus la mobilité sociale est faible.

Coefficients d'élasticité

Mathématiquement, ce coefficient est déterminé par une régression linéaire simple dans laquelle le logarithme du revenu de l'enfant *Y(child)* est une fonction du logarithme du revenu des parents *Y(parent)* :

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

Pour chaque pays, nous allons utiliser une génération aléatoire de la classe de revenu des parents *c(i,parent)*, à partir de ces seules deux informations :

- le coefficient d'élasticité *ρ(j)*
- la classe de revenu de l'enfant *c(i,child)*.

Coefficients d'élasticité

Sources :

- Les coefficients (« *IGEincome* ») donnés par la Banque mondiale, dans GDIM dataset
- Les estimations provenant de multiples études, extrapolées à différentes régions du monde, regroupées dans le fichier *elasticity.txt*

Méthode suivie pour avoir les coefficients d'élasticité par pays :

Application du coef du pays quand il est connu (précis), sinon application du coef de la région (moins précis).

DISTRIBUTIONS CONDITIONNELLES

Coefficients d'élasticité

Après traitement, nous obtenons le tableau suivant :

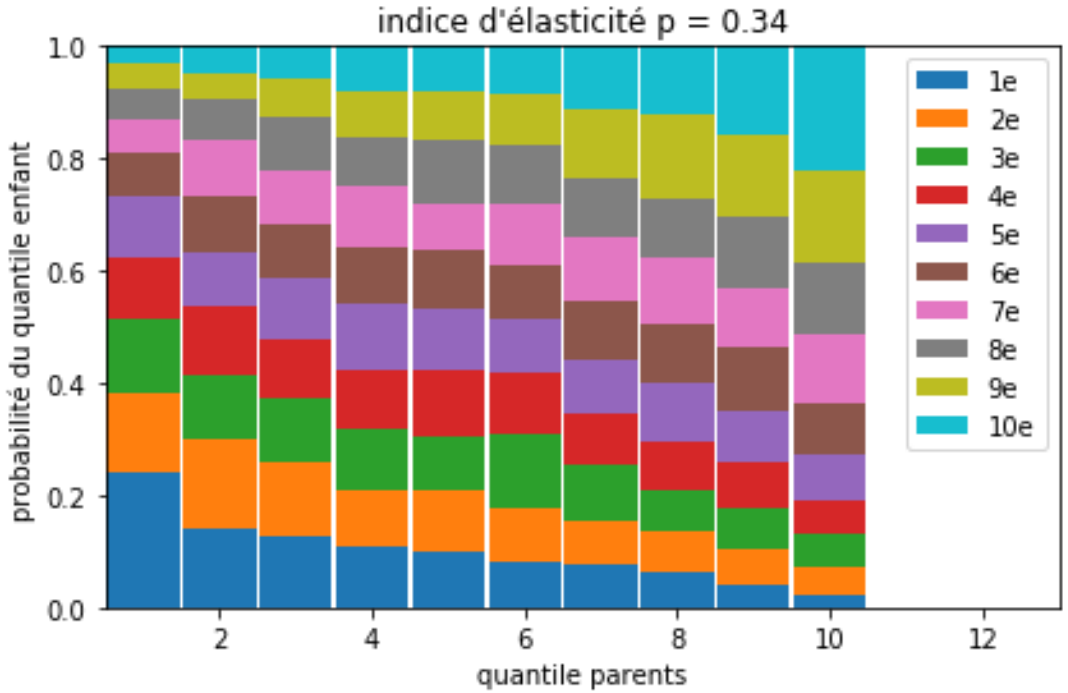
	country_code	quantile	nb_quantiles	income	gdpppp	mean_gini	coef_elast
0	ALB	1	100	728.89795	7297.0	31.411111	0.815874
1	ALB	2	100	916.66235	7297.0	31.411111	0.815874

Très faible mobilité sociale.
Pour info, le coefficient d'élasticité de la France est de 0,34

Attention :
Syrie et Kosovo ont été retirés (pas de source d'info sûre pour compléter les données manquantes).

Représentation des distributions conditionnelles

On utilise le code fourni par OCR.

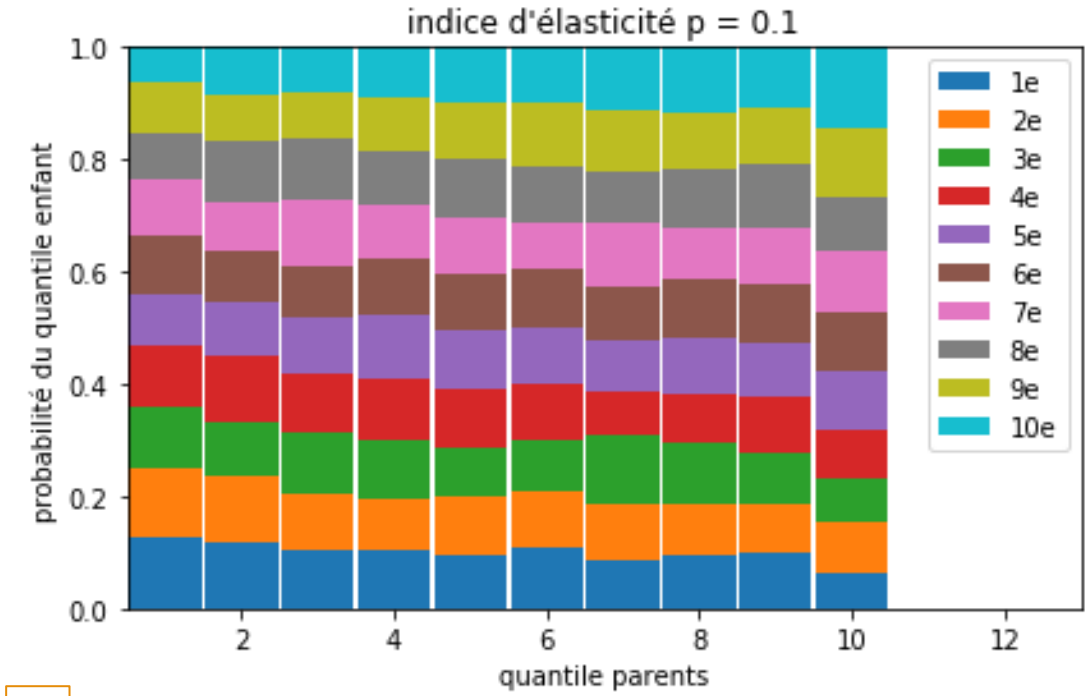


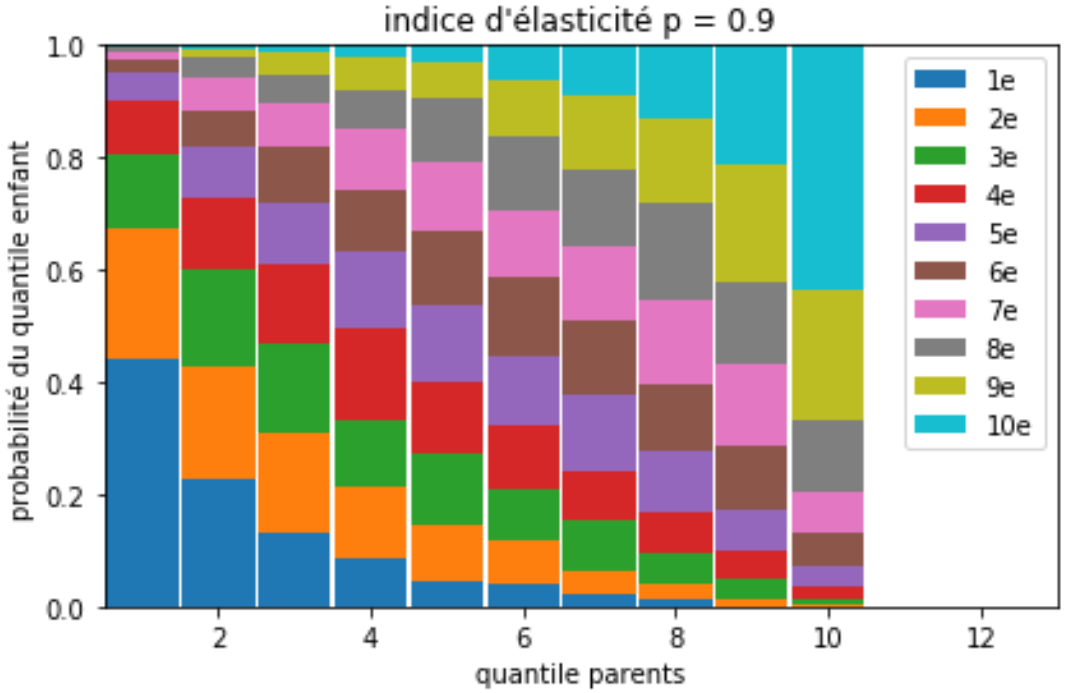
Exemple pour une population segmentée en 10 classes avec p(j) = 0,34 (France).

Montre que notre code fonctionne !

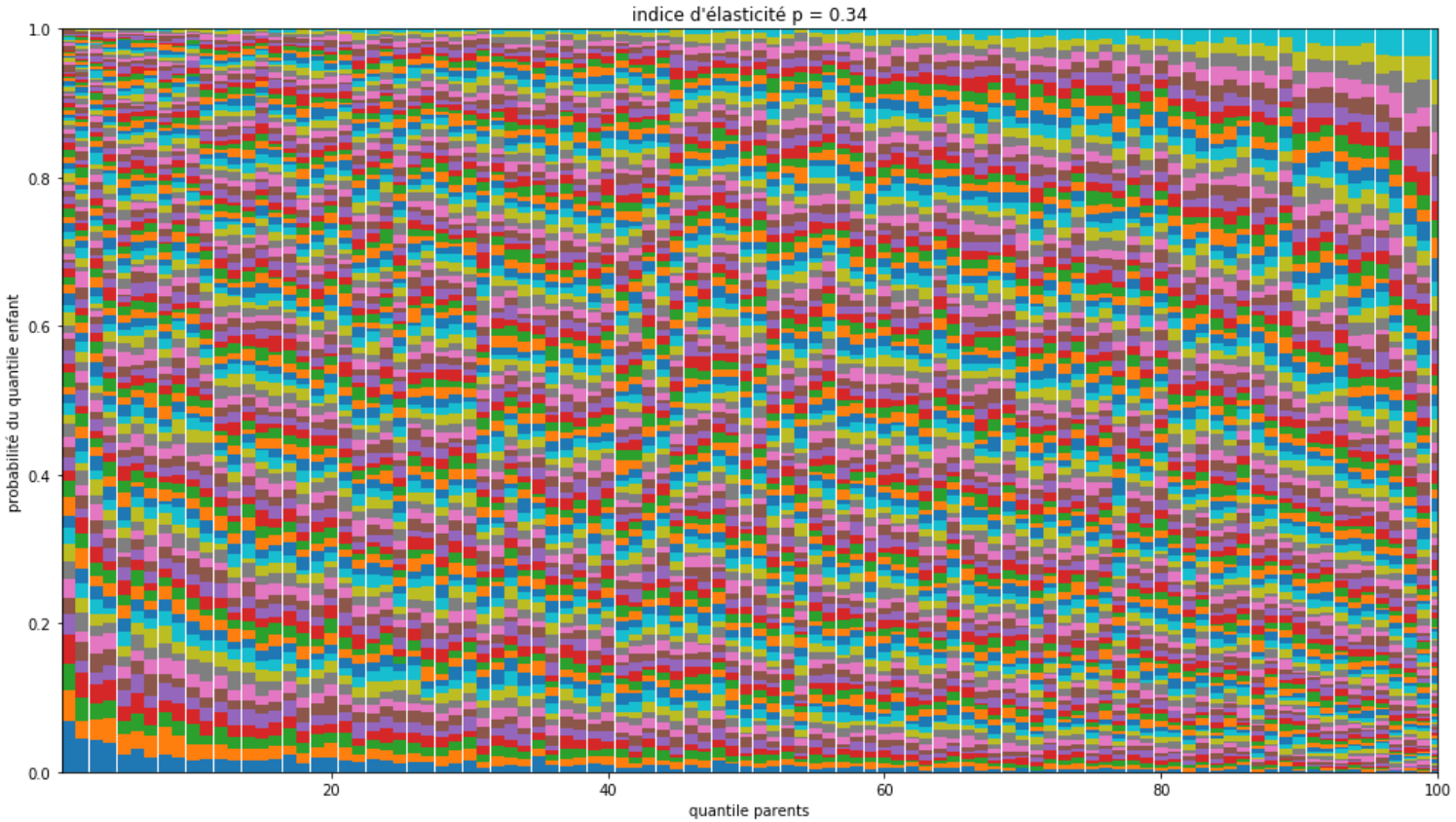
Représentation des distributions conditionnelles

Représentations avec des pays très égalitaires et des pays très inégalitaires :





Représentation des distributions conditionnelles



Exemple pour une population segmentée en centiles avec p(j) = 0,34 (France).

DISTRIBUTIONS CONDITIONNELLES

Génération du dataset pour la mission 4

Création de 499 « clones » → la taille du nouvel échantillon est donc 500 fois plus grande que celui de la *World Income Distribution*.

Pour chaque pays, il y a maintenant 500 individus. Nous attribuons aux 500 individus leurs classes conformément aux distributions trouvées précédemment.

Cela donne 50 000 lignes/pays (500 x 100).
Taille du *dataframe* : 5 600 000 lignes
(500 x 100 x nb de pays).

```
for country in country_list:
    # On va chercher le coef d'élasticité pour chaque pays
    pj = inc5.loc[inc5['country_code'] == country,'coef_elast'].iloc[0]
    # nombre de classes de revenu
    nb_quantiles = 100
    # taille de l'échantillon
    n = 50_000
    y_child, y_parents = generate_incomes(n, pj)
    sample = compute_quantiles(y_child, y_parents, nb_quantiles)
    cd = conditional_distributions(sample, nb_quantiles)
    # Pour chaque classe de revenu enfant...
    for c_i_child in range(100):
        # ...pour chaque classe de revenu parent...
        for c_i_parent in range(100):
            # ...on calcule la proba de la classe parent connaissant la classe enfant
            p = proba_cond(c_i_parent, c_i_child, cd)
            # On ajoute chaque nouvelle ligne à la précédente avec la méthode extend
            list_prob.extend([c_i_parent + 1] * (int(p * 500)))
```

MISSION 4
MODÉLISATION

MODÉLISATION

Objectif :

Expliquer le revenu des individus en fonction de plusieurs variables explicatives :

- le pays de l'individu
- l'indice de Gini de ce pays
- la classe de revenus des parents
- etc.

ANOVA

Y a-t-il une différence de revenu entre les pays ? ANOVA : income ~ country

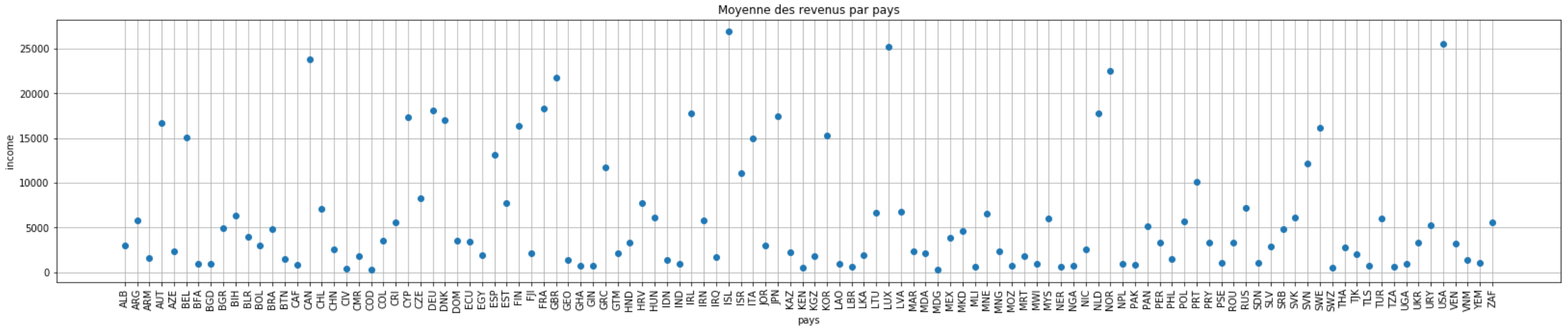
ANOVA ?

- **an**alysis **o**f **va**riance / analyse de la variance
- permet d'étudier l'impact d'une (ou plusieurs) variable qualitative (dans notre cas : *country*) sur une variable quantitative (*income*).
- L'ANOVA vise à comparer des moyennes sur plusieurs échantillons et à répondre à la question : y a-t-il des différences significatives entre les groupes ?

ANOVA

Y a-t-il une différence de revenu entre les pays ? ANOVA : income ~ country

Graphiquement, on observe des moyennes différentes en fonction des pays.



ANOVA

Y a-t-il une différence de revenu entre les pays ? ANOVA : income ~ country

Statistiquement :

- La p-valeur du test de Fisher est très en dessous du seuil de 5% → rejet de H0 (égalité des moyennes) => le pays a un effet sur les revenus.

	sum_sq	df	F	PR(>F)
country_code	4.928326e+11	111.0	97.54556	0.0
Residual	5.046871e+11	11088.0	NaN	NaN

- Plus précisément :
  η^2 = 0,49 → le pays explique 49% de la variance du revenu.
  Nous pouvons aller plus loin.

RÉGRESSION LINÉAIRE

income ~ income_mean + gini

variables explicatives = revenu moyen du pays de l'individu + indice de Gini du pays de l'individu

Résultats

- L'explication de la variance n'est pas meilleur qu'avec l'ANOVA.
- R^2 = 0,49
- La p-valeur de l'indice de Gini n'est pas significative (elle est proche de 1).

OLS Regression Results

Dep. Variable:	income_q	**R-squared:**	0.494
Model:	OLS	**Adj. R-squared:**	0.494
Method:	Least Squares	**F-statistic:**	5467.
Date:	Mon, 29 Jun 2020	**Prob (F-statistic):**	0.00
Time:	19:02:16	**Log-Likelihood:**	-1.1458e+05
No. Observations:	11200	**AIC:**	2.292e+05
Df Residuals:	11197	**BIC:**	2.292e+05
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-1.381e-11	357.813	-3.86e-14	1.000	-701.377	701.377
mean_gini	1.421e-13	8.406	1.69e-14	1.000	-16.477	16.477
income_mean	1.0000	0.010	95.885	0.000	0.980	1.020

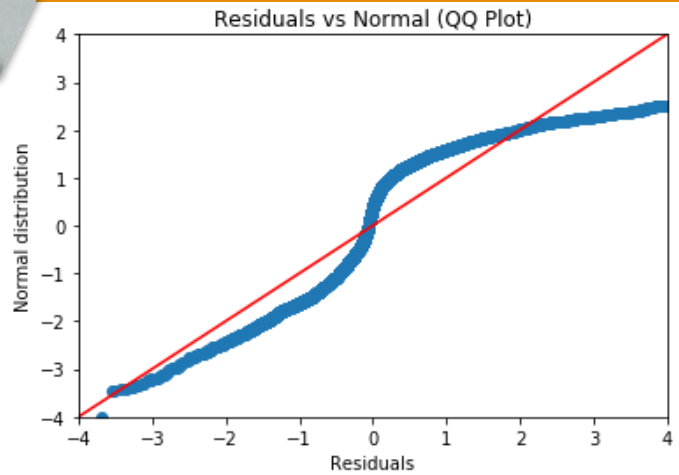
Omnibus:	14173.515	**Durbin-Watson:**	0.686
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	4100801.951
Skew:	6.758	**Prob(JB):**	0.00
Kurtosis:	95.762	**Cond. No.**	5.08e+04

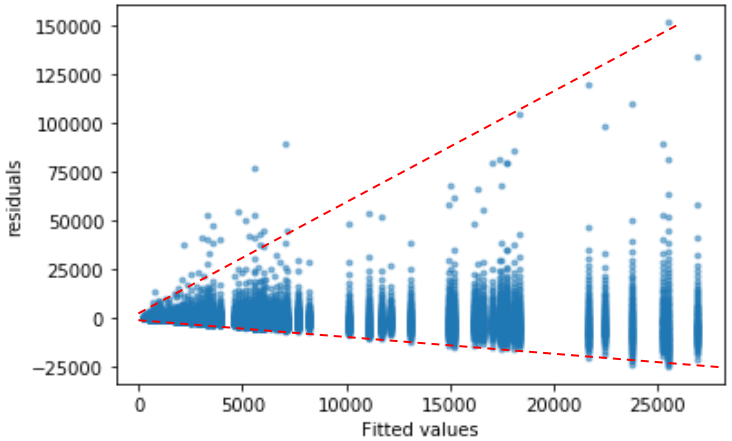
RÉGRESSION LINÉAIRE

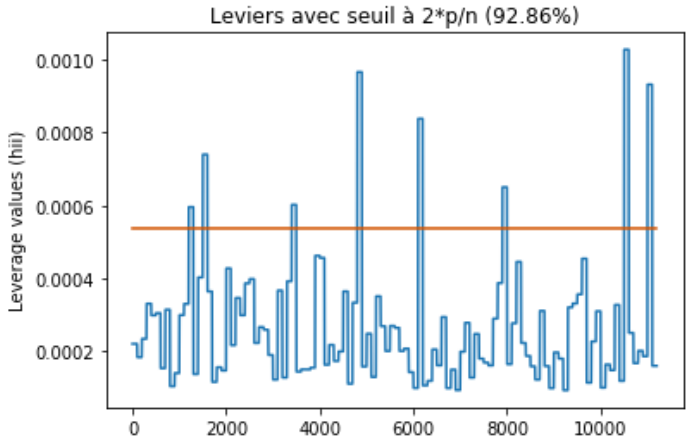
income ~ income_mean + gini

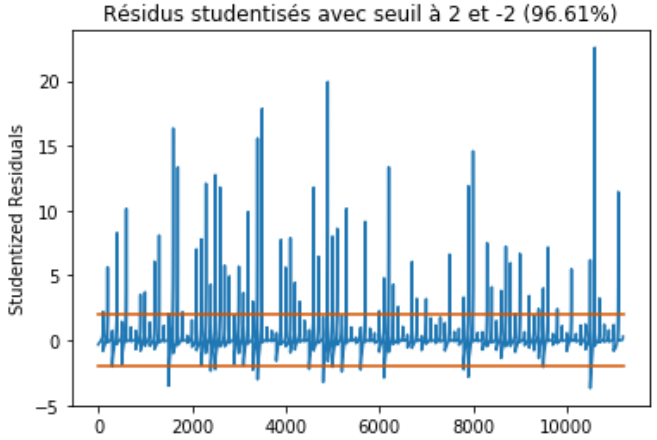
Analyse

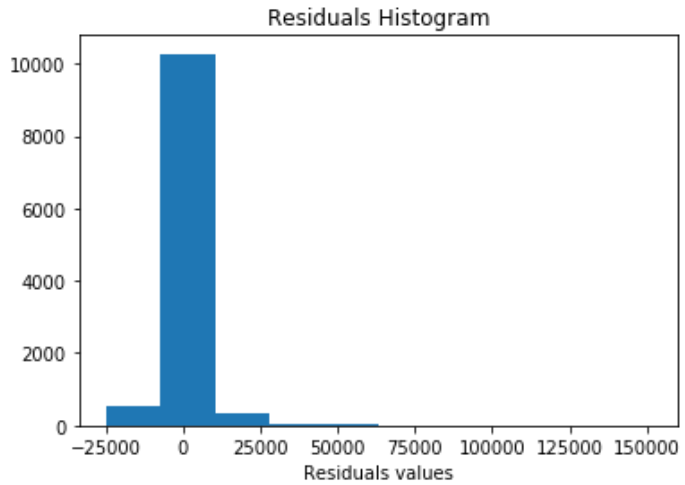
- Les résidus ne suivent pas une loi normale.
- Ils semblent dépendants de la valeur prédite (hétéroscédasticité)
- 7% des individus sont marginaux.
- Les données ne sont pas linéaires, notre modèle prédit mal.

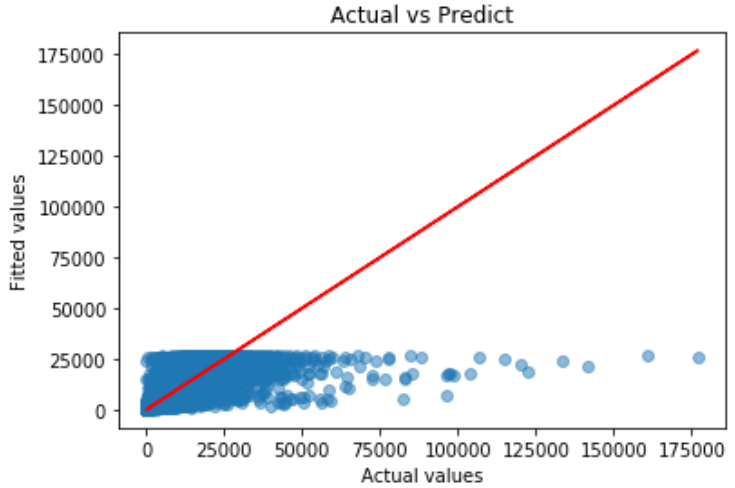


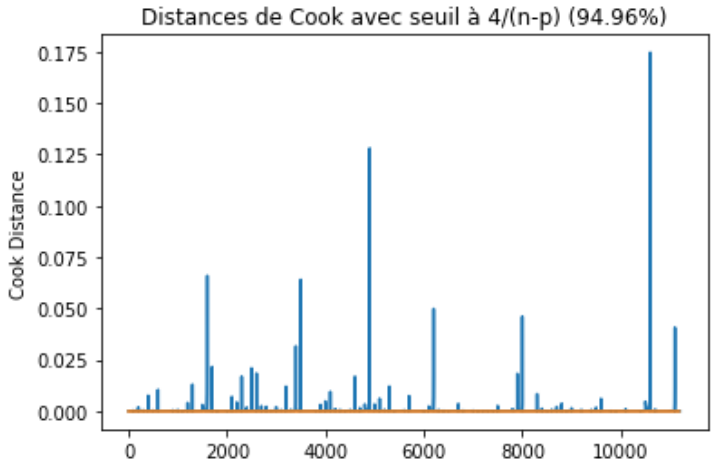












RÉGRESSION LINÉAIRE

log(income) ~ log(income_mean) + gini

Résultats :

- L'explication de la variance est bien meilleure.
- Toutes les p-valeurs sont faibles, y compris pour l'indice de Gini.
- **R^2** = 0,72 (meilleur)
- Le modèle explique 72% de la variance. Les 28% restants peuvent être expliqués par les autres facteurs non considérés dans le modèle (efforts, chance, etc).

OLS Regression Results

Dep. Variable:	income_q	**R-squared:**	0.724
Model:	OLS	**Adj. R-squared:**	0.724
Method:	Least Squares	**F-statistic:**	1.469e+04
Date:	Mon, 29 Jun 2020	**Prob (F-statistic):**	0.00
Time:	21:25:57	**Log-Likelihood:**	-12311.
No. Observations:	11200	**AIC:**	2.463e+04
Df Residuals:	11197	**BIC:**	2.465e+04
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.5117	0.070	7.261	0.000	0.374	0.650
mean_gini	-0.0167	0.001	-18.969	0.000	-0.018	-0.015
income_mean	0.9826	0.006	154.960	0.000	0.970	0.995

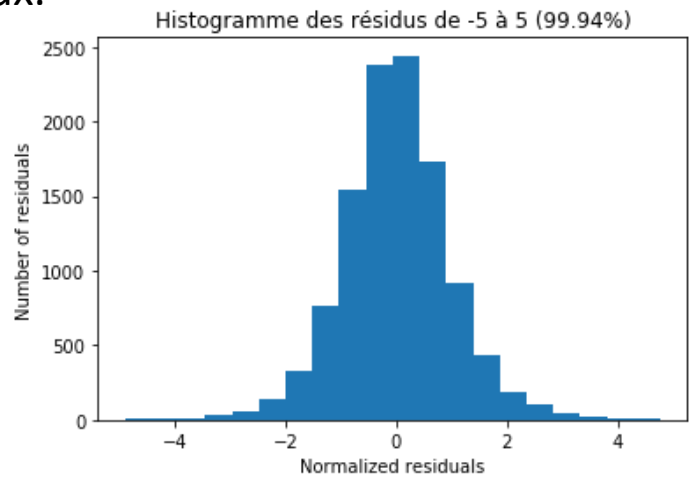
Omnibus:	753.353	**Durbin-Watson:**	0.386
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	3457.146
Skew:	-0.139	**Prob(JB):**	0.00
Kurtosis:	5.708	**Cond. No.**	412.

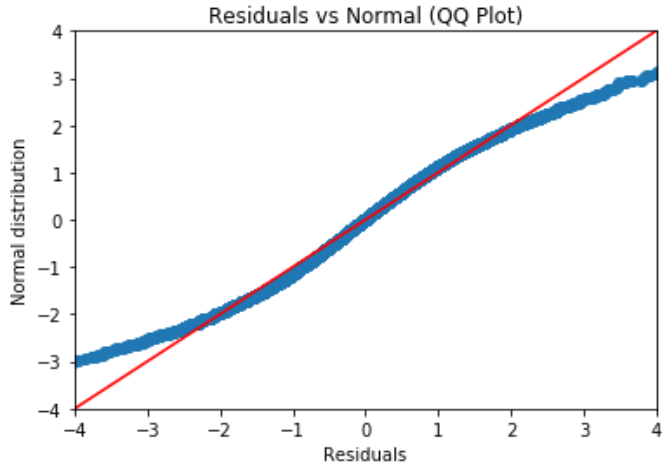
RÉGRESSION LINÉAIRE

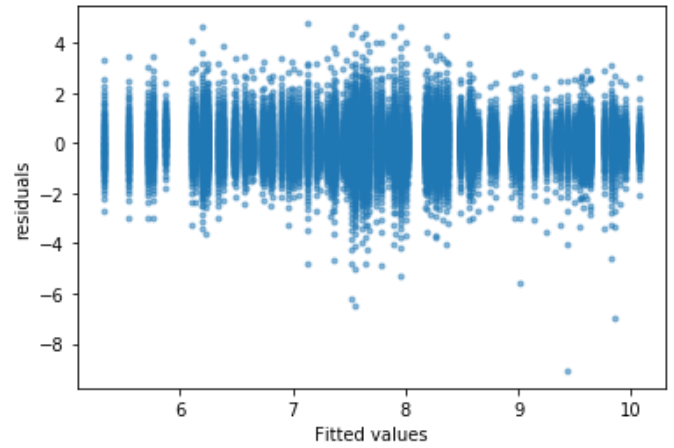
log(income) ~ log(income_mean) + gini

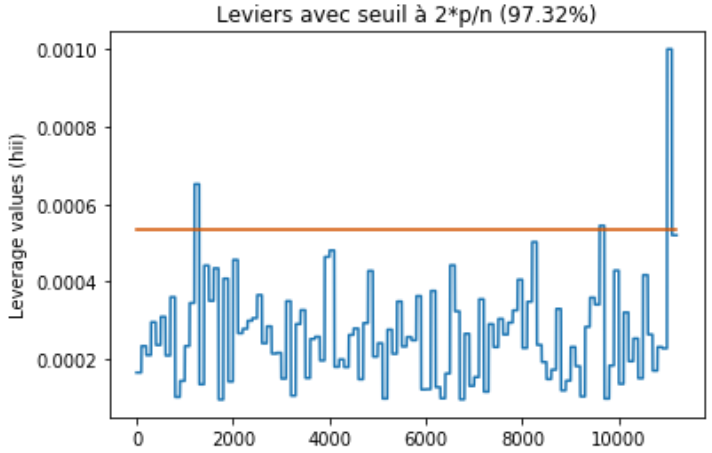
Analyse

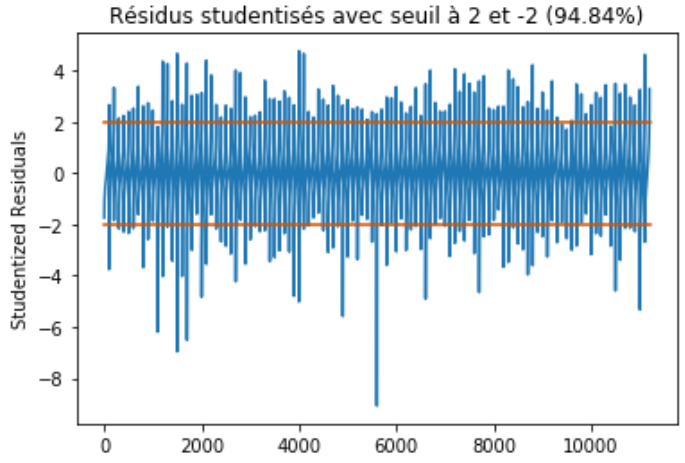
- Les résidus suivent une loi normale.
- Ils ont une variance constante (homoscédasticité).
- Il reste seulement 2% d'individus marginaux.

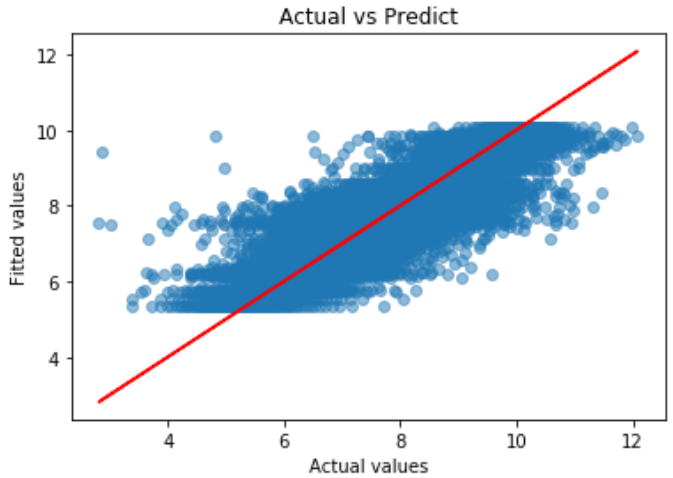


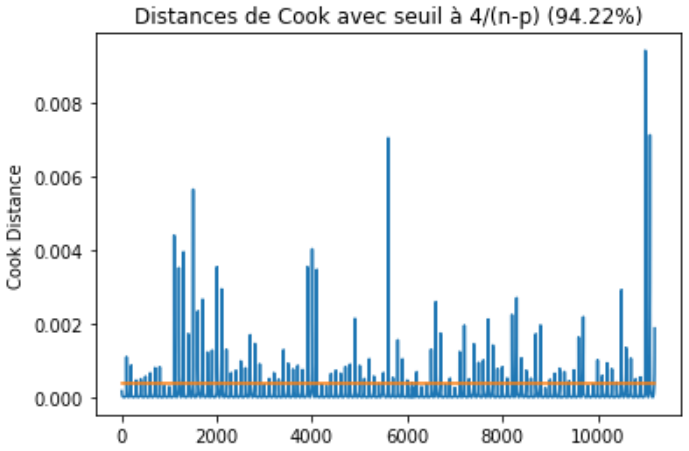












RÉGRESSION LINÉAIRE

Proposition faite au départ du projet :

log(income) ~ log(income_mean) + gini + c_parent

variables explicatives = revenu moyen du pays de l'individu + indice de Gini du pays de l'individu + revenu des parents

Résultats :

- L'explication de la variance est encore meilleure. **R^2** = 0,73.
- Il reste 27 % non expliqués. Le reste peut être expliqué par d'autres facteurs comme par exemple le lieu d'habitation, le niveau d'études, les efforts, la chance, etc.
- Le coef de Gini est négatif => plus il augmente, plus le revenu diminue.

OLS Regression Results

Dep. Variable:	income_q	R-squared:	0.732
Model:	OLS	Adj. R-squared:	0.732
Method:	Least Squares	F-statistic:	5.090e+06
Date:	Mon, 29 Jun 2020	Prob (F-statistic):	0.00
Time:	23:33:07	Log-Likelihood:	-6.0776e+06
No. Observations:	5600000	AIC:	1.216e+07
Df Residuals:	5599996	BIC:	1.216e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.3011	0.003	95.505	0.000	0.295	0.307
mean_gini	-0.0167	3.89e-05	-429.971	0.000	-0.017	-0.017
classe_parent	0.0042	1.05e-05	397.231	0.000	0.004	0.004
income_mean	0.9826	0.000	3513.955	0.000	0.982	0.983

Omnibus:	396746.449	Durbin-Watson:	0.398
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1913090.410
Skew:	-0.155	Prob(JB):	0.00
Kurtosis:	5.847	Cond. No.	713.

CONCLUSION

CONCLUSION

Rappel de la mission : créer un modèle permettant de déterminer le revenu potentiel d'une personne.

En ajoutant la classe de revenu des parents, nous obtenons un modèle suffisamment pertinent de manière à prévoir les revenus enfants.

Théoriquement, ce modèle pourrait encore être amélioré à condition d'ajouter les facteurs non considérés tels que les efforts, la chance, etc.


MERCI
DE VOTRE
ATTENTION !