Mardi 13 juillet 2021

Prédiction de revenus

Analyse et modélisation

- ➔ Banque internationale
- ➔ Cibler les jeunes en âge d'ouvrir un compte
  - ◆ Enfants des clients actuels
- ➔ Prédire le revenu potentiel des enfants d'après le revenu de leur parents
- ➔ Peu de données à disposition :
  - ◆ Revenu des clients actuels (parents)
  - ◆ Revenu moyen du pays
  - ◆ Indice de Gini du pays

Type de modélisation adaptée

Régression linéaire

Prédire une quantitative avec une ou des quantitatives

Sommaire

Données

01

Jeu de données

Données fournies par la Banque mondiale

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728,89795	7297
1	ALB	2008	2	100	916,66235	7297
2	ALB	2008	3	100	1010,916	7297
3	ALB	2008	4	100	1086,9078	7297
4	ALB	2008	5	100	1132,6997	7297
...	...	...	...	...	...	...
11594	COD	2008	96	100	810,6233	303,19305
11595	COD	2008	97	100	911,7834	303,19305
11596	COD	2008	98	100	1057,8074	303,19305
11597	COD	2008	99	100	1286,6029	303,19305
11598	COD	2008	100	100	2243,1226	303,19305

11599 rows × 6 columns

➔ 6 variables

- Pays
- Année de l'étude
- Quantile (classe) de revenu
- Nombre de quantiles
- Revenu moyen du quantile (unité : PPP)
- PIB à parité du pouvoir d'achat (PPA) du pays

➔ Aucun doublon

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 11599 entries, 0 to 11598
Data columns (total 6 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   country       11599 non-null  object
 1   year_survey   11599 non-null  int64
 2   quantile      11599 non-null  int64
 3   nb_quantiles  11599 non-null  int64
 4   income        11599 non-null  object4
 5   gdpppp        11399 non-null  object4
dtypes: int64(3), object(3)
memory usage: 543.8+ KB
```

- ➔ 200 valeurs manquantes dans *gdpppp*
- ➔ Conversion de *income* et *gdpppp* en float

- ➔ 116 pays différents pour 11599 lignes
  Une ligne est manquante
- ➔ Études de 2004 à 2011
- ➔ Quantiles de 1 à 100
- ➔ Nombre de quantiles unique (100)
  Inutile : suppression de la variable
- ➔ Revenus moyens de 16 à 176928
- ➔ PPA de 3 à 4.3

	country	year_survey	quantile	nb_quantiles	income	gdpppp
count	11599	11599.000000	11599.000000	11599.0	11599.000000	1.139900e+04
unique	116	NaN	NaN	NaN	NaN	NaN
top	HUN	NaN	NaN	NaN	NaN	NaN
freq	100	NaN	NaN	NaN	NaN	NaN
mean	NaN	2007.982757	50.500819	100.0	6069.224260	5.022128e+04
std	NaN	0.909633	28.868424	0.0	9414.185972	4.000688e+05
min	NaN	2004.000000	1.000000	100.0	16.719418	3.031931e+02
25%	NaN	2008.000000	25.500000	100.0	900.685515	2.576000e+03
50%	NaN	2008.000000	51.000000	100.0	2403.244900	7.560000e+03
75%	NaN	2008.000000	75.500000	100.0	7515.420900	1.877300e+04
max	NaN	2011.000000	100.000000	100.0	176928.550000	4.300332e+06

df.describe()

Ligne manquante

	country country	year_survey year_survey	quantile quantile	income income	gdpppp gdpppp
0	ALB	2008	1	728.89795	7297.00000
1	ALB	2008	2	916.66235	7297.00000
2	ALB	2008	3	1010.91600	7297.00000
3	ALB	2008	4	1086.90780	7297.00000
4	ALB	2008	5	1132.69970	7297.00000
...	...	...	...	...	...
11595	COD	2008	97	911.78340	303.19305
11596	COD	2008	98	1057.80740	303.19305
11597	COD	2008	99	1286.60290	303.19305
11598	COD	2008	100	2243.12260	303.19305
11599	LTU	2008	41	4882.14065	17571.00000

11600 rows × 5 columns
100 rows × 5 columns 119 rows × 2 columns

Les 100 classes de revenus par pays Moyenne des quantiles par pays France

➔ 100 classes de revenus par pays

➔ La Lituanie n'a que 99 lignes
Moyenne des quantiles différentes des autres pays

➔ Imputation du quantile manquant (41)
Moyenne des quantiles 40 et 42

- ➔ 2 pays sans PPA : Kosovo et Palestine
- ➔ 1.7% des données
- ➔ Pays écartés du dataset (114 restants)

	country	year_survey	quantile	income	gdpppp
5800	XKX	2008	1	437.89370	NaN
5801	XKX	2008	2	508.17133	NaN
5802	XKX	2008	3	591.82820	NaN
5803	XKX	2008	4	668.00000	NaN
5804	XKX	2008	5	730.40220	NaN
...	...	...	...	...	...
11294	PSE	2009	96	2763.88480	NaN
11295	PSE	2009	97	3077.83330	NaN
11296	PSE	2009	98	3449.22240	NaN
11297	PSE	2009	99	4165.99700	NaN
11298	PSE	2009	100	6343.87550	NaN

200 rows × 5 columns

Variables supplémentaires

Indice de Gini
(calculé, pas de valeurs manquantes)

Nom du pays
(données de population)

Population de l'année d'étude
(data.worlbank.org)
1 pays écarté (Taïwan)

	country	year_survey	quantile	income	gdpppp	gini	Country Name	population	region	IGE
0	ALB	2008	1	728.89795	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874
1	ALB	2008	2	916.66235	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874
2	ALB	2008	3	1010.91600	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874
3	ALB	2008	4	1086.90780	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874
4	ALB	2008	5	1132.69970	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874
...	...	...	...	...	...	...	...	...	...	...
11294	COD	2008	96	810.62330	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703
11295	COD	2008	97	911.78340	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703
11296	COD	2008	98	1057.80740	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703
11297	COD	2008	99	1286.60290	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703
11298	COD	2008	100	2243.12260	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703

6500 rows × 10 columns

Coefficient d'élasticité
Mobilité des revenus
(48 pays écartés)

Région
(données IGE)

49 pays écartés (65 au final)

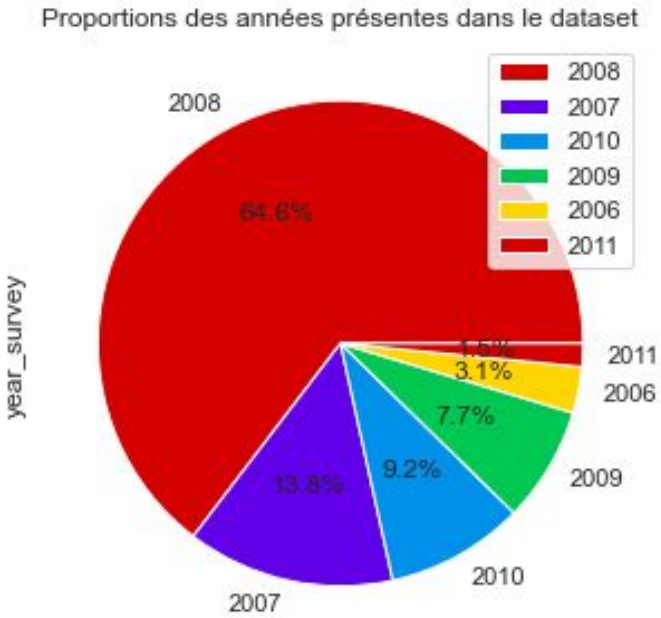
Analyse

02

65 pays répartis en 7 régions :

- ➔ **Europe & Central Asia** : ['Albania' 'Belarus' 'Bosnia and Herzegovina' 'Kazakhstan 'Kyrgyz Republic' 'North Macedonia' 'Romania' 'Russian Federation']
- ➔ **High income** : ['Austria' 'Belgium' 'Canada' 'Chile' 'Croatia' 'Cyprus' 'Czech Republic' 'Denmark' 'Finland' 'France' 'Germany' 'Greece' 'Ireland' 'Italy' 'Japan' 'Korea, Rep.' 'Latvia' 'Luxembourg' 'Netherlands' 'Norway' 'Portugal' 'Slovak Republic' 'Slovenia' 'Spain' 'Sweden' 'United Kingdom' 'United States']
- ➔ **South Asia** : ['Bangladesh' 'India' 'Nepal' 'Pakistan']
- ➔ **Latin America & Caribbean** : ['Bolivia' 'Brazil' 'Colombia' 'Ecuador' 'Guatemala' 'Panama' 'Peru']
- ➔ **East Asia & Pacific** : ['China' 'Malaysia' 'Mongolia' 'Timor-Leste' 'Vietnam']
- ➔ **Sub-Saharan Africa** : ['Congo, Dem. Rep.' 'Ghana' 'Guinea' 'Kenya' 'Madagascar' 'Malawi' 'Mali' 'Nigeria' 'South Africa' 'Tanzania' 'Uganda']
- ➔ **Middle East & North Africa** : ['Egypt, Arab Rep.' 'Jordan' 'Morocco']

- ➔ 6 années différentes
- ➔ Aucun pays n'a d'indicateur pour plusieurs années



	year_survey	population
0	2006	6623518000
1	2007	6705947000
2	2008	6789089000
3	2009	6872767000
4	2010	6956824000
5	2011	7041194000

Source : Wikipedia

➔ Estimation de la population mondiale annuelle

- Fréquence d'apparition de chaque année dans le dataset
- Multipliée par la population de l'année
- Divisée par le nombre d'années présentes

➔ Population mondiale : 6.7 milliards

➔ Échantillon : 5 milliards d'habitants

75% de la population mondiale annuelle estimée

Diversité des distributions de revenus

➔ Panel de 5 pays

France, Slovénie, Suède, USA, Panama

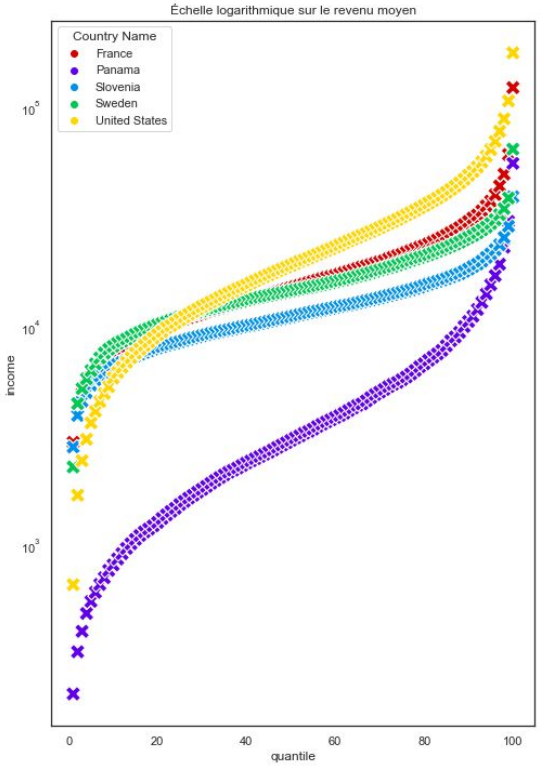
➔ Revenu moyen sur une échelle logarithmique

Représente sur un petit espace une large gamme de valeurs

➔ **USA**
- les plus forts revenus
- parmi les revenus les plus faibles

➔ **France**
- Les classes les plus basses ont parmi les plus hauts revenus entre pays
- Les classes les plus hautes ont parmi les plus hauts revenus entre pays

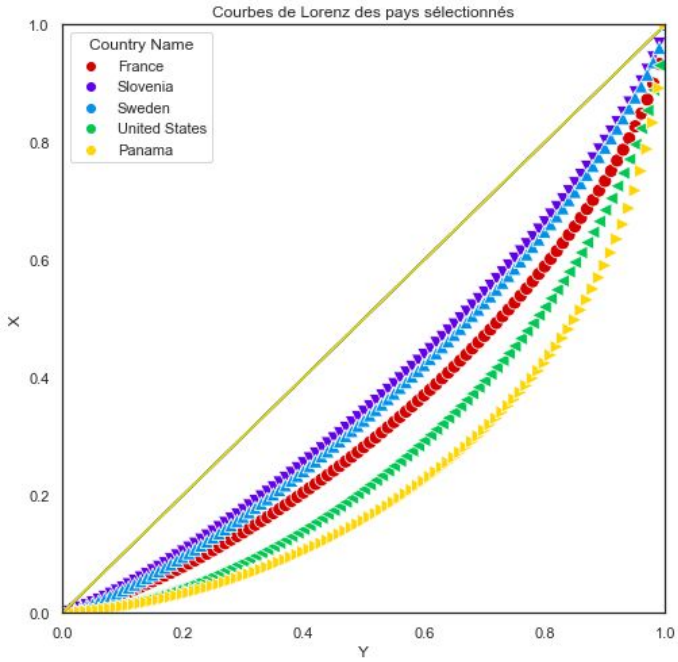
➔ **Panama** : la plus forte amplitude



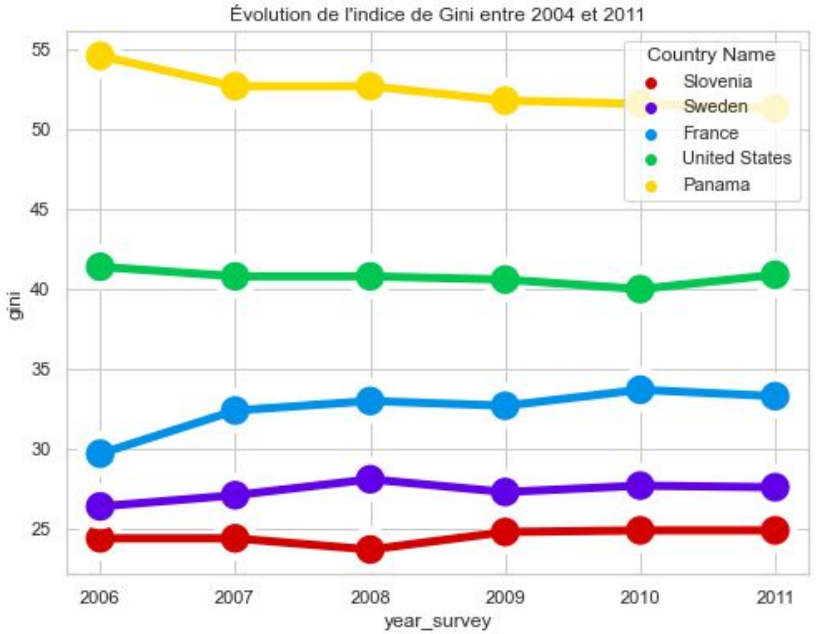
- ➔ Représente les inégalités de revenus

  Indice de Gini = aire entre la diagonale et la courbe

- ➔ **Panama** et **USA** : les plus fortes inégalités de revenus
- ➔ France : inégalité moyenne



Évolution de l'indice de Gini



➔ Indice de Gini

- **0 : égalité parfaite**, toutes les classes ont le même revenu
- **1 : inégalité parfaite**, une seule classe concentre tous les revenus

➔ Indice moyen : 37%

➔ **France** : 33%

- 26ème pays le + égalitaire
- 40ème pays le plus inégalitaire

	Country Name	gini
0	Slovenia	0.230731
1	Slovak Republic	0.247219
2	Czech Republic	0.252864
3	Sweden	0.254887
4	Denmark	0.259871

5 pays les plus égalitaires

	Country Name	gini
62	Guatemala	0.568293
63	Colombia	0.569271
64	South Africa	0.669779

5 pays les plus inégalitaires

65 rows × 2 columns

Génération de réalisations

03

Procédé

➔ Variable *income* = revenu enfant à prédire

- On suppose que les revenus du dataset sont ceux des enfants
- Quantiles initiaux = classe enfants

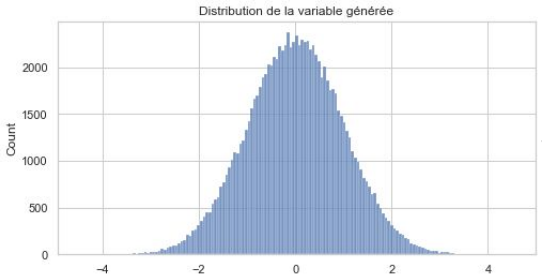
➔ Création de 499 clones de chaque individu

➔ Génération du quantile des parents

- Probabilités conditionnelles à partir de la classe enfant
- D'après le coefficient d'élasticité
- Potentielle variable explicative pour la modélisation

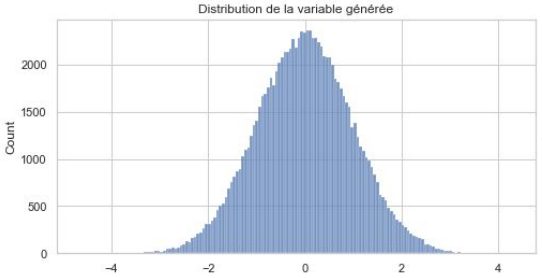
Revenus des parents

ln_y_parent



Résidus

ln_y_parent



2 variables de *n* individus, distribuées selon une loi normale

ln_y_child = IGE * ln_y_parents + residus

- ➔ Calcul des revenus enfant
  - y_child
  - D'après le coefficient d'élasticité
  - Ordre de grandeur qui ne reflétera pas la réalité
- ➔ Passage de ln_y_parent et ln_y_child à l'exponentielle

  Inverse du logarithme
- ➔ Découpage des revenus parent et enfant en 100 quantiles

	y_child	y_parents	c_i_child	c_i_parent
0	1.857582	0.865283	70	45
1	2.052173	0.837190	73	43
2	0.616849	1.233678	35	59
3	15.273448	7.242267	99	98
4	5.313596	1.306459	92	61
...	...	...	...	...
99995	1.254662	3.297281	58	89
99996	0.036814	0.055489	1	1
99997	0.517663	0.224645	30	7
99998	0.457658	0.163294	27	4
99999	0.582007	0.513286	34	26

100000 rows × 4 columns

Génération de la classe parents

- ➔ Nouvel échantillon : 499 clones de chaque individu
  Échantillon 500 fois plus grand
- ➔ Jointure de distributions normales
  - De même longueur que l'échantillon
  - **ln_y_parent**
  - **résidus**
- ➔ Attribution des classes parent aux 500 clones
  Avec le bon coefficient d'élasticité
- ➔ Découpage en 100 quantiles
  On ne garde que la classe parent

country	year_survey	quantile	income	gdpppp	gini	Country Name	population	region	IGE	income_mean	y_child	y_parents
ALB	2008	1	728.89795	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	12.873686	3.489411

year_survey	quantile	income	gdpppp	gini	Country Name	population	region	IGE	income_mean	y_child	y_parents	c_i_child	c_i_parent
2008	1	728.89795	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	12.873686	3.489411	99	90
2008	2	916.66235	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	1.952593	2.004095	73	76
2008	3	1010.91600	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	1.762282	1.443572	70	65
2008	4	1086.90780	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	1.398173	2.587095	62	83
2008	5	1132.69970	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	4.104722	0.661380	90	34
...	...	...	...	...	...	...	...	...	...	...	...	...	...
2008	96	810.62330	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	0.751855	0.836155	41	43
2008	97	911.78340	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	10.851580	1.053216	98	53
2008	98	1057.80740	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	1.548413	1.407393	65	64
2008	99	1286.60290	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	1.657550	2.009538	68	76
2008	100	2243.12260	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	1.301112	0.185311	60	5

si **P(c_i_parent=8 sachant c_i_child=5 et IGE=0.9) = 0.03**
assigner **ci_parent=8** à 15 des 500 individus ayant ci_child=5
car **500 * 0.03 = 15**

Modélisation

04

Objectifs

➔ Prédire le revenu des individus en fonction de plusieurs explicatives

- Pays de l'individu
- Revenu moyen du pays
- Indice de Gini
- Classe de revenus des parents
- etc.

➔ **ANOVA** : *le pays est-il corrélé au revenu ?*

➔ **Régression linéaire** : *les variables prédisent elles correctement le revenu enfant ?*

➔ Analyse de la variance entre :

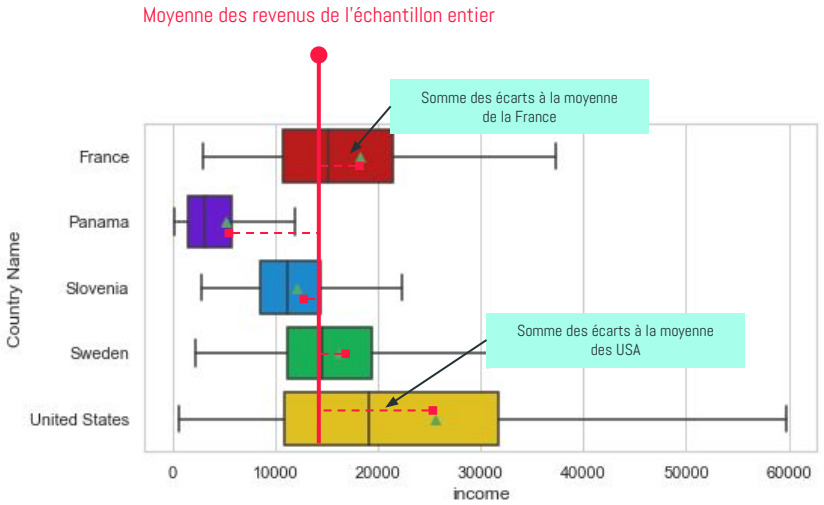
- **1 catégorielle explicative** : le pays
- **1 continue expliquée** : le revenu
- *Les moyennes de revenus sont-elles significativement différentes entre pays ?*

➔ Types d'ANOVA

- **Paramétrique** : remplit conditions (+ robuste)
- **Non paramétrique** : ne remplit pas toutes les conditions (- robuste)

➔ Conditions (paramètres)

- Normalité
- Homogénéité de la variance (homoscédasticité)
- Indépendance des individus (pas de doublons)



Test des paramètres

ANOVA non validée

Test d'ANOVA

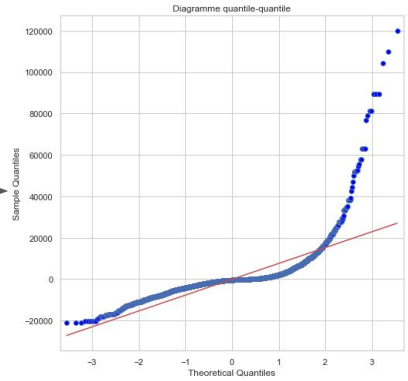
p-value : **0.0**
H0 rejetée : **Probable corrélation**

Test de Shapiro

p-value : **0.0**
H0 rejetée : **Distribution non normale**

Test de Levene

p-value : **0.0**
H0 rejetée : **Pas d'homoscédasticité**



Alternative

Test de Kruskal-Wallis

p-value : **0.0**
H0 rejetée : **Probable corrélation**

- → ANOVA non paramétrique
- → Quand normalité et homoscédasticité non respectées
- → Le revenu est probablement corrélé au pays

Fonction logarithme

- Régressions sur 2 versions des variables
  - Données normales
  - Données à l'échelle logarithmique
- Échelle logarithmique
  - Représente des nombres aux ordres de grandeur différents sur un même graphique
  - Données centrées et réduites
  - Réduit les outliers (réduit la marginalité des quantiles)
  - Réduit l'asymétrie positive
  - Normalise, lisse la distribution
  - Fonction inverse de l'exponentielle

income	gdpppp	gini	Country Name	population	region	IGE	income_mean	c_i_parent	income_ln	income_mean_ln	population_ln	gdpppp_ln
728.89795	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	4	6.591534	8.004643	14.896405	8.895219
916.66235	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	47	6.820739	8.004643	14.896405	8.895219
1010.91600	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	81	6.918612	8.004643	14.896405	8.895219
1086.90780	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	54	6.991092	8.004643	14.896405	8.895219
1132.69970	7297.00000	0.304624	Albania	2947314.0	Europe & Central Asia	0.815874	2994.829902	88	7.032359	8.004643	14.896405	8.895219
...	...	...	...	...	...	...	...	...	...	...	...	...
810.62330	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	31	6.697803	5.620459	17.916685	5.714370
911.78340	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	79	6.815402	5.620459	17.916685	5.714370
1057.80740	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	1	6.963954	5.620459	17.916685	5.714370
1286.60290	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	98	7.159761	5.620459	17.916685	5.714370
2243.12260	303.19305	0.443997	Congo, Dem. Rep.	60411195.0	Sub-Saharan Africa	0.707703	276.016044	18	7.715624	5.620459	17.916685	5.714370

➔ Calcul d'une variable quantitative d'après d'autres quantitatives

X = variable(s) indépendante(s) explicatives (features, inputs, paramètres)
y = variable dépendante expliquée

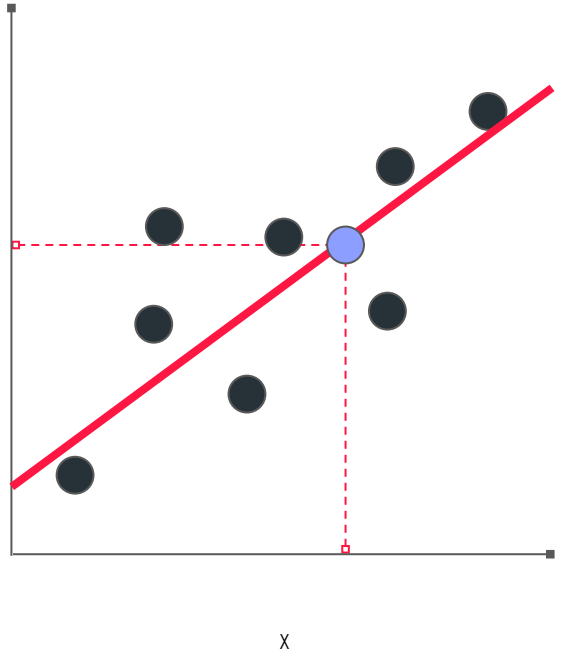
➔ Apprentissage supervisé

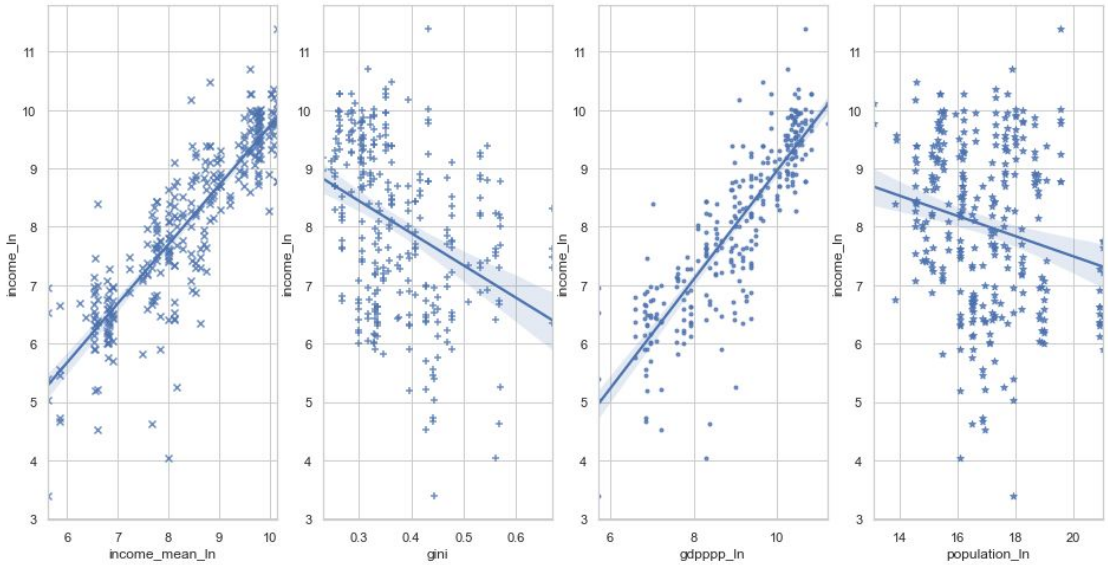
On connaît la valeur réelle de **y** pour chaque groupe d'explicatives

➔ Postulat : on peut aligner les points sur une droite

Relation linéaire décrivant le mieux la relation entre X et y
= f(**X**) = pente x **X** + constante (point sur l'ordonnée quand X = 0)

➔ Prédit des données continues





Régression entre le revenu et les variables quantitatives

Droites à la moindre somme des carrés des distances verticales aux valeurs (*R-squared*)

Prédicteurs (features) :

- revenu moyen au logarithme
- gini

OLS Regression Results

Dep. Variable:	income_ln	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	5.344e+06
Date:	Sun, 11 Jul 2021	Prob (F-statistic):	0.00
Time:	23:26:04	Log-Likelihood:	-3.5146e+06
No. Observations:	3250000	AIC:	7.029e+06
Df Residuals:	3249997	BIC:	7.029e+06
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4774	0.004	132.709	0.000	0.470	0.484
income_mean_ln	0.9860	0.000	2989.131	0.000	0.985	0.987
gini	-1.6758	0.004	-385.302	0.000	-1.684	-1.667

Omnibus:	293481.741	Durbin-Watson:	0.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1801877.265
Skew:	-0.199	Prob(JB):	0.00
Kurtosis:	6.626	Cond. No.	112.

Notes:
[1] Standard Errors assu… …fied.

Coefficients de régression

Y = *a* + *b* * X

- ➔ indique la force du lien linéaire entre l'explicative et la prédite
- ➔ chercher les coefficients de la fonction linéaire permettant de maximiser les prédictions (R2)
- ➔ Minimiser la fonction de perte et d'erreur

Variance expliquée

Doit idéalement s'approcher de 1

Colinéarité

Affecte la performance d'un modèle
Chercher la valeur la plus basse

Significativité des coefficients de régression

- ➔ p < 5% : la variable est significative
- ➔ p > 5% : la variable n'est pas intéressante

➔ Variance expliquée inférieure

- 0.46 contre 0.76

➔ Significativités nulles

- p-values de l'ordonnée et de *gini* à 1
- Coefficient de régression du revenu moyen à 1
- Le revenu moyen explique à lui seul le modèle (échelle des valeurs trop grande)

➔ Colinéarité plus forte

- 112 précédemment
- Avertissement du résumé

➔ Modèle moins performant qu'avec logarithme

OLS Regression Results

Dep. Variable:	income	R-squared:	0.468
Model:	OLS	Adj. R-squared:	0.468
Method:	Least Squares	F-statistic:	1.431e+06
Date:	Sun, 11 Jul 2021	Prob (F-statistic):	0.00
Time:	23:26:07	Log-Likelihood:	-3.3803e+07
No. Observations:	3250000	AIC:	6.761e+07
Df Residuals:	3249997	BIC:	6.761e+07
Df Model:	2		
Covariance Type:	nonrobust		

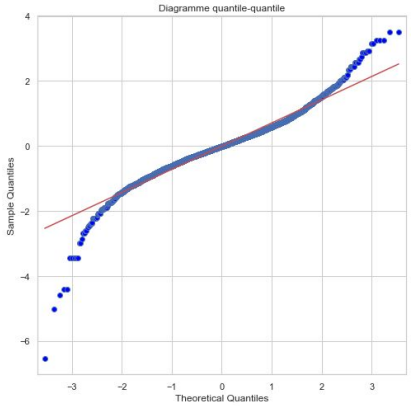
	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-2.846e-10	21.518	-1.32e-11	1.000	-42.175	42.175
income_mean	1.0000	0.001	1560.073	0.000	0.999	1.001
gini	3.738e-11	50.262	7.44e-13	1.000	-98.512	98.512

Omnibus:	3699598.503	Durbin-Watson:	0.683
Prob(Omnibus):	0.000	Jarque-Bera (JB):	602306325.511
Skew:	5.786	Prob(JB):	0.00
Kurtosis:	68.680	Cond. No.	1.34e+05

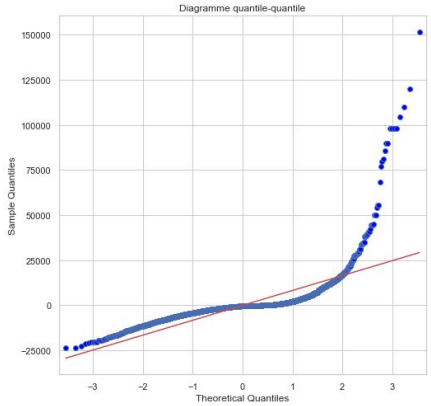
Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.34e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

Normalité des résidus

- ➔ Normalité des résidus requise pour valider la régression linéaire
- ➔ Distribution des résidus plus normale avec données au logarithme



Prédicteurs au logarithme



Prédicteurs sans logarithme

Modèle 3 : ajout de la classe parent (logarithme)

OLS Regression Results

Dep. Variable:	income_ln	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	3.562e+06
Date:	Sun, 11 Jul 2021	Prob (F-statistic):	0.00
Time:	23:26:09	Log-Likelihood:	-3.5146e+06
No. Observations:	3250000	AIC:	7.029e+06
Df Residuals:	3249996	BIC:	7.029e+06
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4776	0.004	130.403	0.000	0.470	0.485
income_mean_ln	0.9860	0.000	2989.130	0.000	0.985	0.987
gini	-1.6758	0.004	-385.301	0.000	-1.684	-1.667
c_i_parent	-5.39e-06	1.37e-05	-0.393	0.694	-3.23e-05	2.15e-05

Omnibus:	293481.479	Durbin-Watson:	0.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1801876.391
Skew:	-0.199	Prob(JB):	0.00
Kurtosis:	6.626	Cond. No.	771.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

➔ Variance expliquée identique

➔ Classe parent non significative

- p-value à 1
- La classe parent est inutile pour modéliser

➔ Colinéarité plus forte

- 112 précédemment
- Avertissement du résumé

➔ Modèle moins performant que sans la classe parent

Modèle 4 : ajout de la classe parent (sans logarithme)

OLS Regression Results

Dep. Variable:	income	R-squared:	0.468
Model:	OLS	Adj. R-squared:	0.468
Method:	Least Squares	F-statistic:	9.537e+05
Date:	Sun, 11 Jul 2021	Prob (F-statistic):	0.00
Time:	23:26:11	Log-Likelihood:	-3.3803e+07
No. Observations:	3250000	AIC:	6.761e+07
Df Residuals:	3249996	BIC:	6.761e+07
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-7.2579	22.858	-0.318	0.751	-52.059	37.543
income_mean	1.0000	0.001	1560.072	0.000	0.999	1.001
gini	-0.0245	50.262	-0.000	1.000	-98.537	98.487
c_i_parent	0.1440	0.153	0.941	0.347	-0.156	0.444

Omnibus:	3699594.930	Durbin-Watson:	0.683
Prob(Omnibus):	0.000	Jarque-Bera (JB):	602303763.247
Skew:	5.786	Prob(JB):	0.00
Kurtosis:	68.680	Cond. No.	1.34e+05

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.34e+05. This might indicate that there are
strong multicollinearity or other numerical problems.

➔ Variance expliquée identique

➔ Classe parent non significative

- p-value à 34%
- La classe parent est inutile pour modéliser

➔ Modèle moins performant

- que sans la classe parent
- qu'avec des variables au logarithme

Modèle 5 : toutes les variables

- ➔ Toutes les variables en features, au logarithme
- ➔ Pas plus de variance expliquée
  - ◆ Toutes les variables sont pourtant significatives
- ➔ Plus forte colinéarité
- ➔ Modèle moins performant

OLS Regression Results

Dep. Variable:	income_ln	R-squared:	0.767
Model:	OLS	Adj. R-squared:	0.767
Method:	Least Squares	F-statistic:	2.138e+06
Date:	Sun, 11 Jul 2021	Prob (F-statistic):	0.00
Time:	23:26:13	Log-Likelihood:	-3.5145e+06
No. Observations:	3250000	AIC:	7.029e+06
Df Residuals:	3249994	BIC:	7.029e+06
Df Model:	5		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.4515	0.006	73.444	0.000	0.439	0.464
income_mean_ln	1.0059	0.001	734.127	0.000	1.003	1.009
gini	-1.7067	0.006	-302.031	0.000	-1.718	-1.696
gdpppp_ln	-0.0186	0.001	-14.507	0.000	-0.021	-0.016
population_ln	0.0021	0.000	7.564	0.000	0.002	0.003
IGE	0.0083	0.002	3.599	0.000	0.004	0.013

Omnibus:	294219.681	Durbin-Watson:	0.407
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1812734.810
Skew:	-0.199	Prob(JB):	0.00
Kurtosis:	6.637	Cond. No.	333.

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Résumé

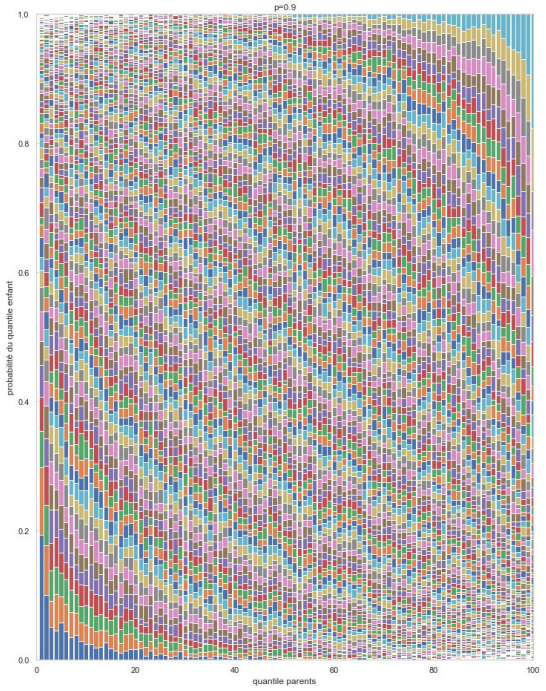
- ➔ Un indice de Gini plus fort fait baisser la prédiction de revenu
- ➔ Facteurs non expliqués par le modèle : niveau d'études, sexe, hasard, etc.

Modèle

- ➔ 65 pays (75% de la population mondiale)
- ➔ Corrélés au revenu moyen par pays
- ➔ Régression linéaire pour prédire le revenu enfant
  - ◆ Feature 1 : Revenu moyen (au logarithme) du pays
  - ◆ Feature 2 : Indice de Gini du pays
- ➔ 76% de variance expliquée

➔ Distribution conditionnelle de la classe parent pour chaque classe enfant

- Si 6 individus ont c_i_child = 5 et c_i_parent = 8
- Et si 200 individus sur 20.000 ont c_i_child = 5
- Alors probabilité d'avoir c_i_parent = 8 sachant c_i_child = 5 et sachant le coefficient d'élasticité = 6 / 200



Faible mobilité des revenus (IGE = 0.9)