Effectuez une prédiction de revenus

Créer un modèle de détermination du revenu potentiel d'une personne

Rappel de la mission

Objectif

- Cibler de nouveaux clients :
  ↳ les enfants des clients actuels ➔ ceux dont la probabilité d'avoir des revenus élevés est la plus grande

Méthode :

- Créer un modèle de détermination du niveau de revenu potentiel d'une personne

Données disponibles :

- Données relatives au pays de naissance (revenu moyen, indice de Gini…)
- Données relatives au revenu des parents

Sommaire

Mission 1

Présentation des données

1. Nettoyage des données

1.1. Dataset "income" (source interne)

	country	year_survey	quantile	nb_quantiles	income	gdpppp
0	ALB	2008	1	100	728.89795	7297.00000
1	ALB	2008	2	100	916.66235	7297.00000
2	ALB	2008	3	100	1010.91600	7297.00000

Table de correspondance code_pays / pays

```
code_corr_pays = pd.read_csv("data/inputs/sql-pays.csv")
code_corr_pays.head(2)
```

	id	code	alpha2	alpha3	nom_pays_fr	nomp_pays_en
0	1	4	AF	AFG	Afghanistan	Afghanistan
1	2	8	AL	ALB	Albanie	Albania

- Valeur aberrante du PIB/hab des îles Fidgi

```
inc_temp = income_data[['country', 'year_survey', 'gdpppp']].drop_duplicates().sort_values(by='gdpppp', ascending=False)
```

	country	year_survey	gdpppp
3200	FJI	2008	4300332.0
6299	LUX	2008	73127.0

- Manque le PIB/hab pour 2 pays

```
# on cherche les pays correspondant aux NaN
print(income_data[income_data['gdpppp'].isna()].country.unique())
```

```
['XKX' 'PSE']
```

→ Kosovo et Territoires Palestiniens Occupés

On remplace les valeurs manquantes et aberrantes par les valeurs trouvées sur le site de la Banque Mondiale

- Manque un centile pour un pays

```
# on cherche le pays qui n'a pas 100 entrées
for i in l_country:
    if income_data.groupby('country').size()[i] != 100:
        print(i)
        pays_cible = i
```

```
LTU
```

→ Lituanie

```
# liste de réf
l_ref=np.arange(1, 101, 1)

# liste pays_cible
df_pays_cible = income_data[income_data['country'] == pays_cible]
l_pays_cible = df_pays_cible['quantile'].tolist()

# comparaison des 2 listes
q_manquant = set(list(l_ref)) - set(list(l_pays_cible))
q_manquant = list(q_manquant)[0]

print(f" Ainsi, il manque le quantile {q_manquant} au pays {pays_cible}")
```

```
Ainsi, il manque le quantile 41 au pays LTU
```

On estime la valeur manquante grâce à la méthode "interpolate"

quantile	income
39	4802.36800
40	4868.45070
41	4882.14065
42	4895.83060
43	4950.63800

1. Nettoyage des données

1.2. Dataset "indices_gini"

	code_pays	pays	2004	2006	2007	2008	2009	2010	2011
0	ALB	Albanie	NaN	NaN	NaN	0.300	NaN	NaN	NaN
1	ARG	Argentine	0.484	0.463	0.462	0.449	0.437	0.436	0.426
2	ARM	Arménie	0.375	0.297	0.312	0.292	0.280	0.300	0.294

```
1 l_chine = ["HKG", "MAC", "CHN", "TWN"]
2 income_data[income_data["code_pays"].isin(l_chine)].drop_duplicates("code_pays")
```

	code_pays	annee	centile	income	gdp
1700	CHN	2007	1	16.719418	5712.0

Dans "income", Chine et provinces regroupés dans un seul pays

Dans "indices_gini", présence de Chine et de Taïwan

- Un pays de moins que dans dataset "income"

"Income" : 115 pays
"Gini" : 116 pays

```
1 pays_gini_manquant = set(list(income_data.country)) - set(list(gini_data.code_pays))
2 pays_gini_manquant
```

```
{'TWN'}    Taïwan
```

On supprime Taïwan de "indices_gini"

1.3. Dataset "population"

- Intégration des données démographiques à partir des données de la FAO

Traitements réalisés :

- Cas du Kosovo et de la Serbie traités différemment :
  - Banque mondiale différencie le Kosovo de la Serbie
  - FAO intègre le Kosovo à la Serbie

Solution retenue :
→ différencier le Kosovo de la Serbie

- Données du Soudan manquantes

2. Synthèse des données

Dataframe de synthèse

	code_pays	pays	annee	pop	gini	gdp	income_avg	centile	income
9213	RUS	Fédération de Russie	2008	143248764.0	0.4167	14766.0	7156.770709	14	2512.7940
2756	DOM	République Dominicaine	2008	9458075.0	0.5025	7505.0	3558.402105	57	2560.6123
8380	PAK	Pakistan	2008	171648986.0	0.2999	2335.0	887.839279	81	1114.1371

Taille du df
Absence de valeurs nulles

```
df.shape
# verif de presence de valeurs nulles
df.isna().sum()
```

executed in 11ms, finished 11:23:42 2021-06-15

```
(11500, 9)

code_pays     0
pays          0
annee         0
pop           0
gini          0
gdp           0
income_avg    0
centile       0
income        0
dtype: int64
```

- Notion de centile

  Les revenus sont distribués par centiles de la population

	pays	annee	centile	income
	Fédération de Russie	2008	14	2512.7940

Le 1% de la pop russe appartenant au 14e centile gagne en moyenne 2512 $PPA en 2008

Résumé des données

	nb_pays_income	nb_pays_gini	nb_pays_total	pop_mondiale	pop_analyse	part_pop_analyse
2004	1	65	115	6 461 159 389	5 965 985 908	92.34%
2006	5	68	115	6 623 517 833	6 108 503 125	92.22%
2007	15	67	115	6 705 946 610	6 179 609 328	92.15%
2008	75	68	115	6 789 088 686	6 251 081 605	92.08%
2009	12	70	115	6 872 767 093	6 322 742 818	92.0%
2010	6	71	115	6 956 823 603	6 394 448 759	91.92%
2011	1	66	115	7 041 194 301	6 466 129 243	91.83%

Le nombre total de pays correpond au nombre total de pays du dataset 'income' (115 pays différents)

✓ Ainsi, 115 pays différents sont représentés sur la durée totale de l'analyse, ce qui représente, en prenant les données démographiques de 2008 :

- 47.13% du nombre mondial de pays,
- 92.08% de la population mondiale.

Mission 2

Diversité mondiale de la répartition des revenus

1. Indice de Gini

Définition

→ indicateur synthétique permettant de rendre compte du niveau d'inégalité pour une variable, ici les revenus, et sur une population donnée.
→ Il varie entre 0 (égalité parfaite) et 1 (inégalité extrême).

Traitement des données

df gini_data :

	code_pays	pays	2004	2006	2007	2008	2009	2010	2011
26	DNK	Danemark	0.249	0.259	0.262	0.252	0.267	0.272	0.273
22	CRI	Costa Rica	0.483	0.494	0.493	0.486	0.506	0.482	0.487

Vérification que tous les pays ont au moins 1 valeur

```
# nb d'années représentées
nb_annees = len(gini_data.columns[2:])
# nb de NaN par lignes (donc nb d'année manquante par pays)
nb_nan_gini_pays=gini_data.isna().sum(axis=1)
# verif si nb de NaN par pays = nb d'années représentées
pays_aucun_gini = nb_nan_gini_pays[nb_nan_gini_pays.values == nb_annees]

nb_pays_aucun_gini = pays_aucun_gini.shape[0]
if nb_pays_aucun_gini>0:
    pays_sans_gini = gini_data.loc[gini_data.index.isin(pays_aucun_gini.index), 'pays'].to_list()
    pays = ", ".join(pays_sans_gini)
    display(HTML(str(nb_pays_aucun_gini)+" pays sans valeurs de Gini : "+pays))
```

7 pays sans valeurs de Gini : Ghana, Kenya, Cambodge, Monténégro, Serbie, République Arabe Syrienne, Yémen

On calcule pour ces pays la valeur de Gini pour l'année correspondant à leurs données "income"

Sélection de 6 pays

Objectif : montrer la diversité des situations au niveau mondial

Méthode utilisée :

- on sélectionne uniquement des pays avec un indice de Gini pour toutes les années de l'étude
- on les choisit de la façon suivante :
  - Pays avec l'indice de Gini moyen min et max
  - Pays avec le PIB/hab min et max
  - Pays avec le rapport "1% des plus riches" / "1% des plus pauvres" min et max

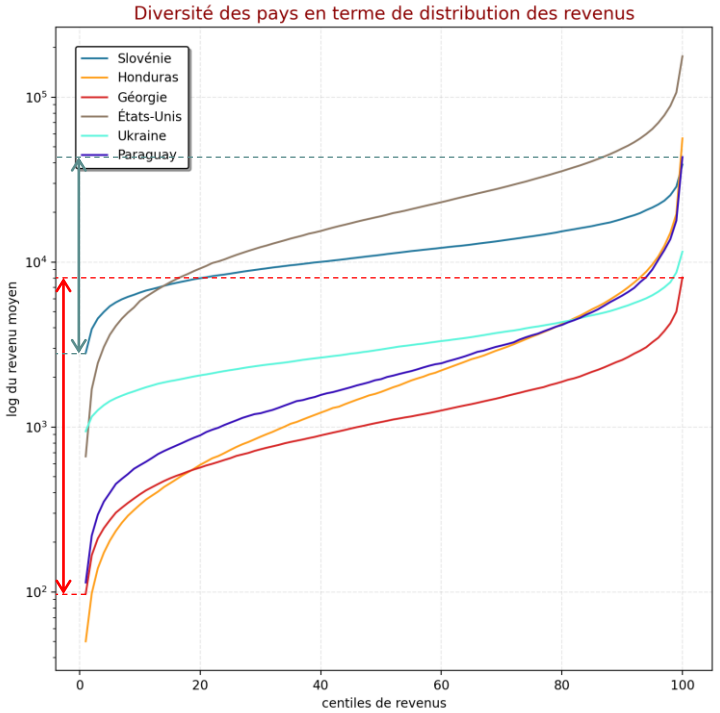
gini_avg min ---------- **Slovénie**
gini_avg max ---------- **Honduras**
pib/hab min ---------- **Géorgie**
pib/hab max ---------- **États-Unis**
rapport riches-pauvres min ----- **Ukraine**
rapport riches-pauvres min ----- **Paraguay**

2. Diversité des situations : échelle logarithmique

Data des 6 pays sélectionnés

	code_pays	pays	pop	gini	gdp	income_avg	income_min_c1	income_max_c100
0	GEO	Géorgie	4 142 654	0.3901	4 516 $	1 363 $	97 $	8 057 $
1	HND	Honduras	7 980 955	0.6017	3 628 $	3 296 $	50 $	56 265 $
2	PRY	Paraguay	6 081 296	0.5251	4 347 $	3 278 $	114 $	43 296 $
3	SVN	Slovénie	2 023 052	0.2307	27 197 $	12 106 $	2 814 $	39 012 $
4	UKR	Ukraine	46 158 711	0.2551	6 721 $	3 349 $	942 $	11 564 $
5	USA	États-Unis	303 486 012	0.4318	43 261 $	25 503 $	663 $	176 928 $

Représentation graphique grâce aux logs



Intérêt de l'échelle logarithmique pour rendre compte des inégalités de revenus

➔ **comparer** les **répartitions** des revenus de plusieurs pays **indépendamment de leur niveau**

↳ ex : comparaison Slovénie et Géorgie:

- les montants ne sont pas comparables (le centile max de la Géorgie n'est que de 8000$) mais la mesure des distance montre que l'écart entre les 1% plus riches et les 1% plus pauvres est plus important en Géorgie qu'en Slovénie

→ Echelle intéressante pour mesurer des écarts relatifs plutôt que absolus

Caractéristique de l'échelle logarithmique

→ 2 graduations dont le rapport vaut 10 sont à distances constantes

3. Courbe de Lorenz

Représentation graphique avec la courbe de Lorenz

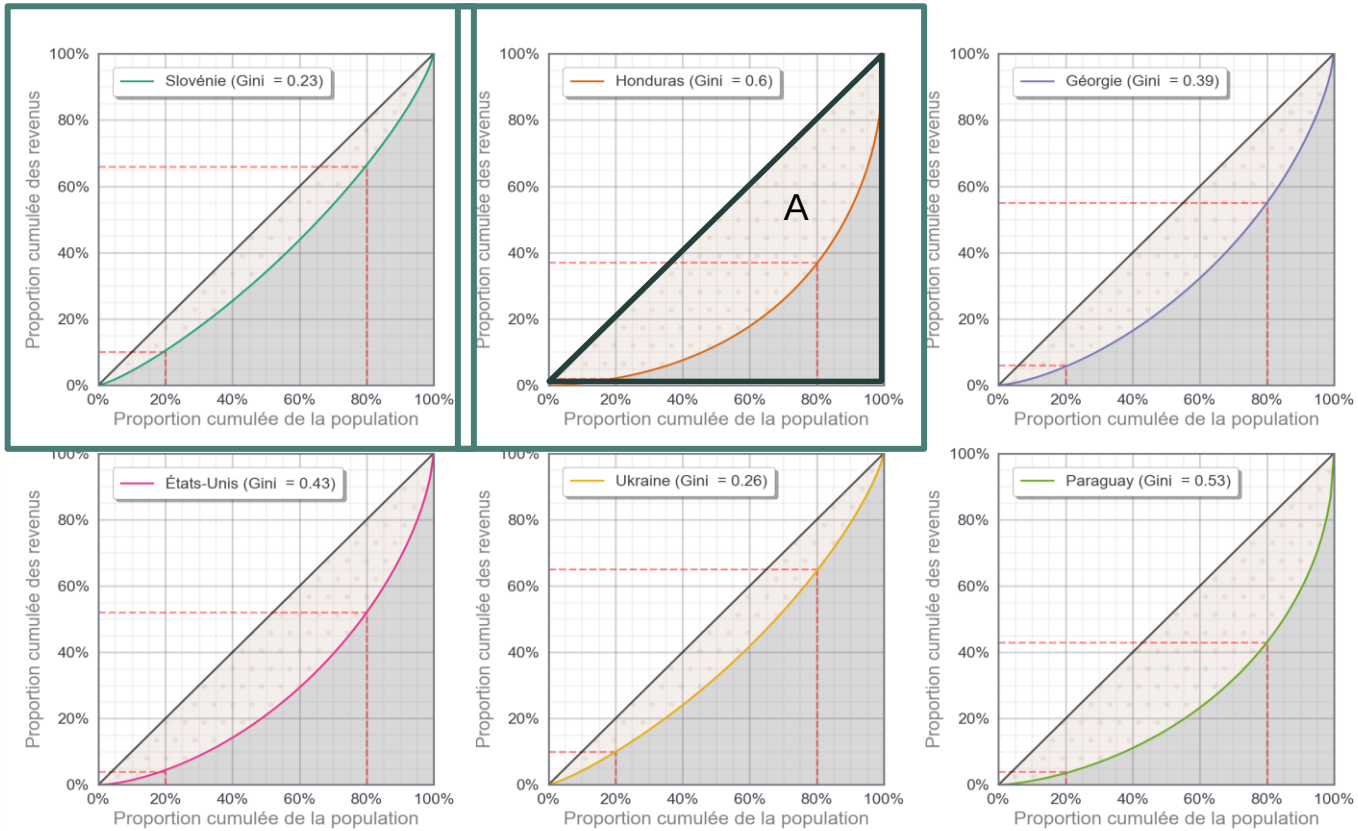
- ✓ Courbe qui permet de visualiser la fonction de répartition qui associe à chaque quantile de la population la part des revenus captée par celle-ci

- ✓ Lecture du graphique :
  - Plus la courbe de Lorenz est proche de la diagonale, plus la répartition est égalitaire
    Ex : la Slovénie
  - A l'inverse, plus elle est éloignée de la diagonale, plus la répartition est inégalitaire
    Ex : Honduras

- ✓ Exemple détaillé : [les Etats-Unis]

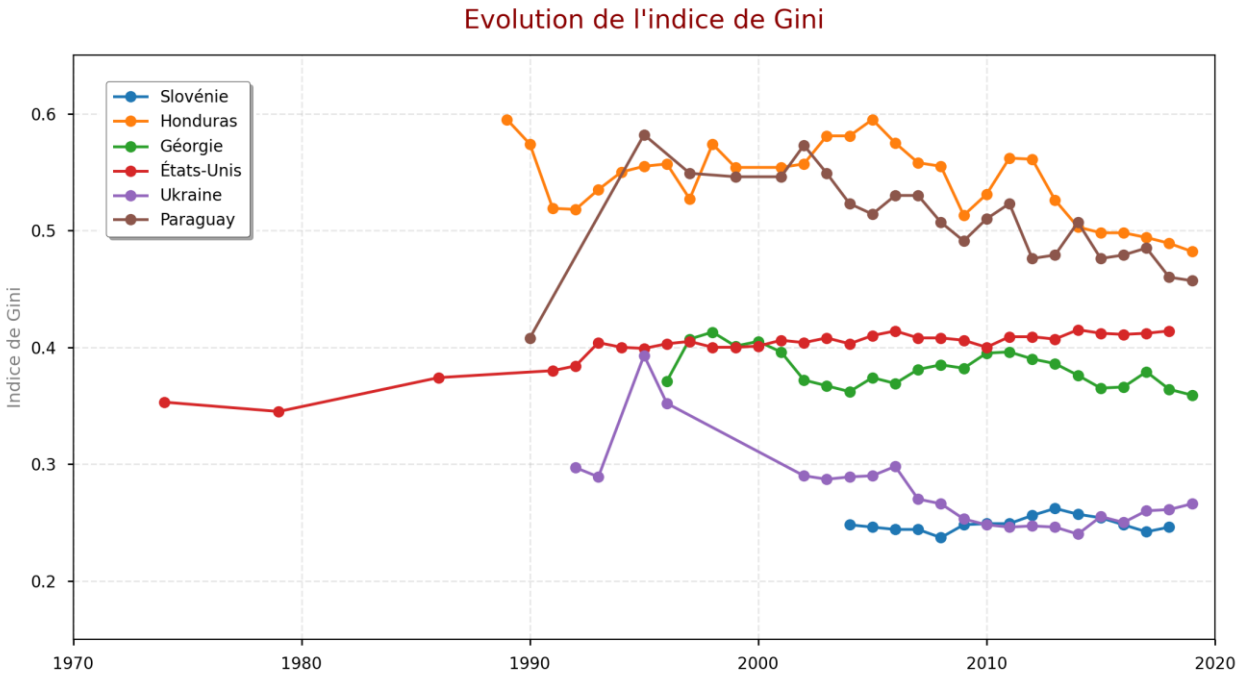
- ✓ Relation avec l'indice de Gini

  Indice de Gini est égal au rapport entre :
  - l'aire de la surface A (surface entre la courbe de Lorenz et la diagonale)
  - et l'aire du triangle rectangle



4. Evolution de l'indice de Gini

Evolution temporelle pour les pays sélectionnés



Evolution différente selon les pays

- L'indice des Etats-Unis est particulièrement stable dans le temps, en particulier depuis 1992

- L'indice du Paraguay et de l'Ukraine a fortement augmenté au milieu des années 90 Il passe ainsi de 0,4 en 1990 à 0,58 en 1995 (contexte politique particulier)

=> L'explication de l'évolution des indices de Gini par pays dépasse le cadre de notre analyse

5. Classement des pays par indice de Gini

```python
# Calcul de l'indice de Gini moyen
gini_data['gini_avg']=gini_data.mean(axis=1)

# on trie le df gini_data par la colonne gini_avg
gini_comp = gini_data.sort_values(by='gini_avg')

# on fait un reset_index() pour mettre à jour l'index (permettra d(avoir le rang de chaque pays)
gini_comp = gini_comp.reset_index()

# calcul de la moyenne mondiale
gini_avg_monde = gini_comp['gini_avg'].mean()
gini_avg_monde = round(gini_avg_monde,2)

# les 5 pays avec le gini le plus faible
pays_gini_faible = gini_comp.nsmallest(5, 'gini_avg').pays
pays_gini_faible = list(pays_gini_faible)

# les 5 pays avec le gini le plus élevé
pays_gini_eleve = gini_comp.nlargest(5, 'gini_avg').pays
pays_gini_eleve = list(pays_gini_eleve)

# position de la France
val_gini_fr = gini_comp.loc[gini_comp['code_pays'] == 'FRA', 'gini_avg'].values[0]
val_gini_fr = round(val_gini_fr, 2)
pos_fr = (gini_comp[gini_comp['code_pays'] == 'FRA'].index.values[0]) + 1
```

Enseignements du classement des pays par l'indice de Gini

- **Gini mondial moyen : 0,38**

- Les 5 pays les plus égalitaires (donc avec l'indice de Gini le plus faible) :

  ↳ **Slovénie, Danemark, Slovaquie, République Tchèque, Azerbaïdjan**

- Les 5 pays les plus inégalitaires (donc avec l'indice de Gini le plus élevé) :

  ↳ **Afrique du Sud, République Centrafricaine, Honduras, Guatemala, Brésil**

- **La France se classe en 35e position,** avec un Gini moyen de 0,32

Mission 3

Détermination des classes de revenus "parents"

Contextualisation

Rappel de la mission

Objectif : Construire un modèle permettant de déterminer la classe des revenus enfants potentielle, à partir de plusieurs variables explicatives :

- variables relatives au pays de naissance : revenu moyen et indice de Gini
- variable relative à la classe de revenu des parents

Lors de la mission 2, nous avons préparé les variables liées au pays

➔ nous allons maintenant préparer la variable relative à la classe de revenu des parents

➔ il faut donc associer à chaque *c_i_enfant* la classe revenu parent *c_i_parents* correspondante ;

Cette classe parent sera différente en fonction des pays :

↳ elle dépend de la **mobilité intergénérationnelle des revenus**

La mobilité intergénérationnelle des revenus

✓ Elle se mesure grâce au coefficient d'élasticité P_j pour le pays j ➔ $ln(Y_{child}) = \alpha + \rho_j ln(Y_{parent}) + \epsilon$

✓ Valeur comprise entre 0 et 1 : plus P_j est élevée, moins la mobilité intergénérationnelle des revenus est importante

✓ On récupère la valeur de ce coefficient pour 65 pays
et on l'estime pour les 50 autres pays en fonction de leur zone géographique

	code_pays	pays	annee	pop	gini	gdp	income_avg	centile	income	p
7461	MRT	Mauritanie	2008	3296238.0	0.4043	2226.734400	1798.609083	62	1624.8164	0.6600

Distributions conditionnelles et création de l'échantillon

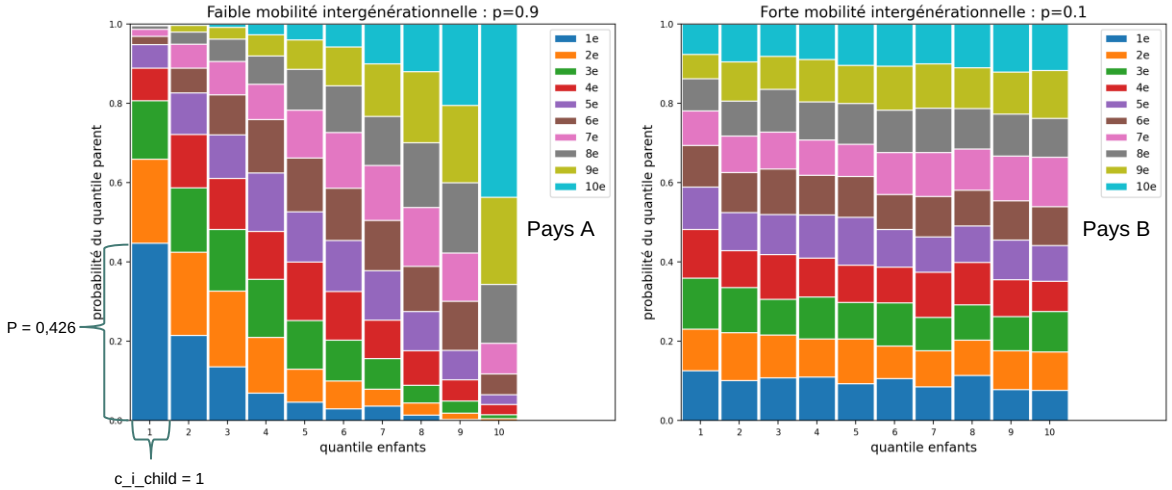
Distributions conditionnelles

- On va déterminer pour chaque pays les distributions conditionnelles de *c_i_parents* pour une génération aléatoire

  ➔ On aura ainsi 100*100 probabilités conditionnelles par pays

- Afin de créer un modèle robuste, on va créer 500 individus pour chaque classe de revenus enfants

Exemple de représentation graphique des probabilités conditionnelles



Rq : répartition des revenus par déciles pour plus de lisibilité

Soient 2 pays, pays A et pays B

✓ Pays A :

Sachant que :

- la classe de revenus d'un enfant est 1
- et que le coef. d'élasticité est de 0.9 (donc faible mobilité intergénérationnelle),

alors la probabilité que la classe de revenu des parents soit aussi égale à 1 est de :

P(c_i_parent = 1 | c_i_child = 1, pj = 0.9) = 0.426

Mission 4

Modèles explicatifs des revenus "enfants"

1. ANOVA à 1 facteur explicatif

Rappel : objectif de créer des modèles pour expliquer les variations de revenus enfants

	code_pays	pays	pop	gini	gdp	y_child_avg	y_child	elas	c_i_parent	ln_y_child	ln_y_child_avg
3838668	MYS	Malaisie	27 236 006	0.4682	13163.0	6006.34	7752.15	0.54	51	8.955726	8.700571
642649	BRA	Brésil	192 030 362	0.5445	9559.0	4807.48	7732.09	0.64	57	8.953134	8.477929

- `code_pays` : code international iso-3
- `pays` : nom du pays en français
- `pop` : nombre d'habitants du pays
- `gini` : indice de gini du pays
- `gdp` : PIB/hab en $PPA
- `y_child_avg` : revenu moyen des enfants
- `y_child` : revenu enfant par centile c_i_parent
- `elas` : coef d'élasticité du pays
- `c_i_parent` : centile de revenu parent
- `ln_y_child` : logarithme de y_child
- `ln_y_child_avg` : logarithme de y_child_avg

1er modèle : ANOVA avec une variable explicative : le pays de naissance

- Objectif : expliquer le revenu des individus en fonction du pays

  ↳ on veut mesurer l'influence de la variable explicative "pays" (variable catégorielle) sur la variable à expliquer "revenus" (variable continue)

- Principe de l'analyse de la variance (ANOVA) :

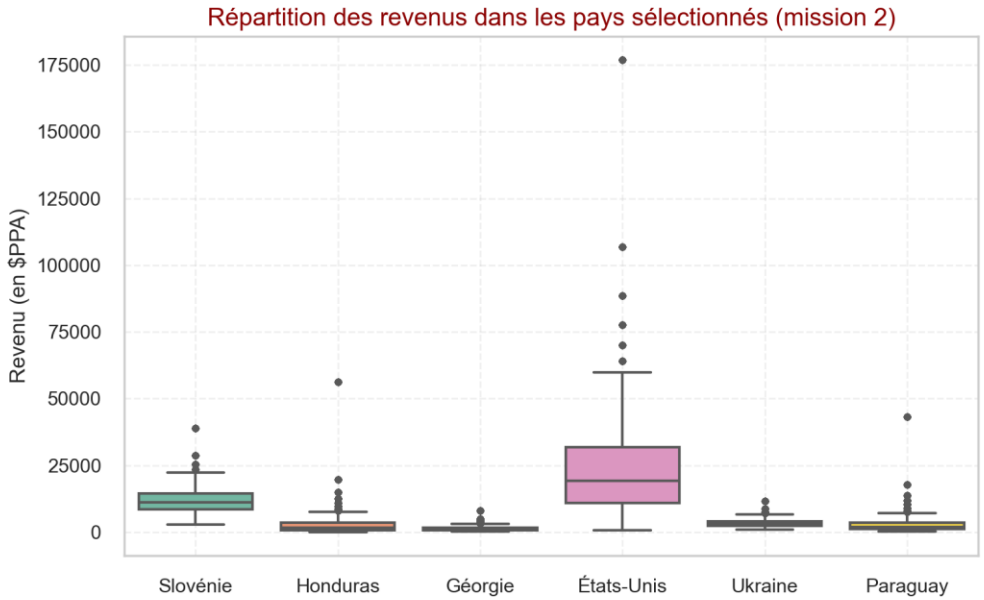
  ↳ déterminer si les moyennes de plusieurs groupes sont différentes grâce à la décomposition de la variance

  SCT = SCE + SCR => variation totale = variation expliquée par le modèle + variation résiduelle

  => eta² (η^2) = $\frac{SCE}{SCT}$ => part de la variation expliquée par le modèle

1. ANOVA à 1 facteur explicatif

Représentation graphique des 6 pays sélectionnés lors de la mission 2



Les revenus paraissent bien différents en fonction des pays

↳ réalisons une ANOVA pour confirmer cette hypothèse

1. ANOVA à 1 facteur explicatif

Réalisation de l'ANOVA et analyse des résultats

- Utilisation de la bibliothèque Python "statsmodels"

```
anova_pays = ols('y_child ~ code_pays', data=dcf).fit()
```

- Test de Fischer

Hypothèse nulle H0 : La moyenne des revenus des différents pays est égale

Hypothèse alternative H1: Tous les pays n'ont pas la même moyenne des revenus

si α <= 0.05 : on rejette Ho au profit de l'alternative H1 :
→ les revenus sont différents selon les pays

si α > 0.05 : on accepte H0 :
→les revenus sont les mêmes quelque soit le pays

	df	sum_sq	mean_sq	F	PR(>F)
code_pays	114.0	2.496311e+14	2.189746e+12	49864.239021	0.0
Residual	5749885.0	2.525014e+14	4.391416e+07	NaN	NaN

P_value proche de 0, donc < 0,05 → on ne peut pas accepter H0

↳ le pays a bien une influence sur les montants de revenus

1. ANOVA à 1 facteur explicatif

- Performance du modèle

  Calcul du η^2 qui indique la part de la variance des revenus expliquée par la variable "pays"

	df	sum_sq	mean_sq	F	PR(>F)	EtaSq
code_pays	114.0	2.496311e+14	2.189746e+12	49864.239021	0.0	0.497142
Residual	5749885.0	2.525014e+14	4.391416e+07	NaN	NaN	NaN

$\eta^2 = 0,5$

↳ La variable pays explique donc près de 50% de la variance des revenus

Conditions d'application de l'ANOVA

Test de Fischer repose sur les hypothèses suivantes

- normalité des distributions
- homoscédasticité des variances

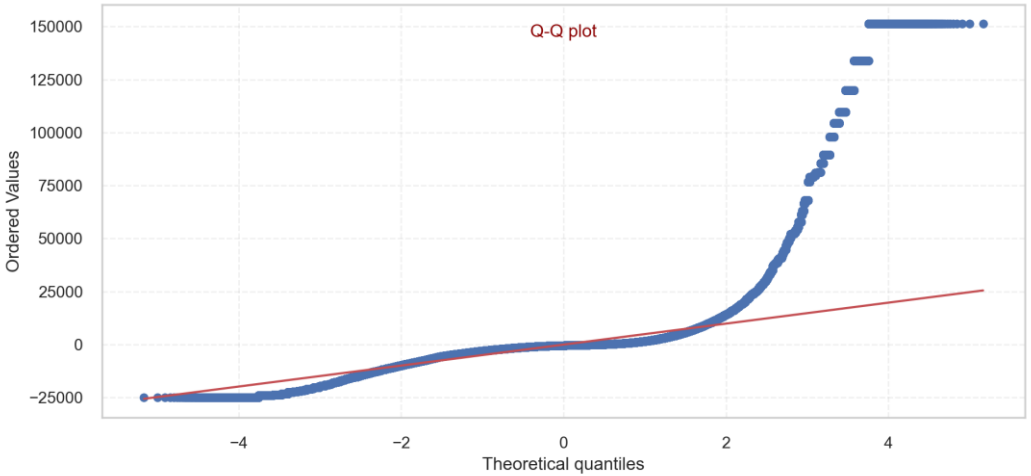
Pour vérifier que notre modèle respecte ces hypothèses, nous allons tester les **résidus**

Cf. les différences entre valeurs observées et valeurs estimées par le modèle de régression

→ cf. les erreurs observées

1. ANOVA à 1 facteur explicatif

- Test de normalité des résidus



Test de normalité de Kolmogorov-Smirnov

H0 : distribution normale des résidus
H1 : résidus ne suivent pas une distribution normale

```
import statsmodels.stats.diagnostic as smd
stat, p = smd.kstest_normal(anova_pays.resid, dist='norm')
```

p = 0,000999 donc < 0,05

→ on rejette H0
→ donc les résidus ne suivent pas une loi normale

- Test d'homoscédasticité des variances : test de Levene

H0 : homoscédasticité des résidus
H1 : hétéroscédasticité des résidus

```
import pingouin as pg
res_anova_homo = pg.homoscedasticity(dcf, dv='y_child', group='code_pays', method='levene')
display(res_anova_homo)
```

	W	pval	equal_var
levene	12753.68435	0.0	False

p = 0 donc < 0,05

→ on rejette H0
→ donc les variances ne sont pas constantes

1. ANOVA à 1 facteur explicatif

Ainsi, les conditions de normalité et d'homoscédasticité des résidus ne sont pas respectées

Mais l'ANOVA est un modèle robuste

→ si la taille des échantillons est suffisamment importante (ici chaque modalité de la variable pays est composée de 100 individus différents), les résultats restent interprétables en l'absence du respect strict des hypothèses d'application

Résultats de l'ANOVA

$\eta^2 \simeq 0.5$ → la variable pays permet d'expliquer près de 50% de la variance des revenus

Remarque :

- en effectuant l'ANOVA sur le log des revenus, on obtient un η^2 de 0.73

→ le pays de naissance explique alors près de 73% de la variance des revenus

2. Régression linéaire à 2 variables explicatives

2ème modèle avec deux variables explicatives : le revenu moyen et l'indice de Gini

Choix du modèle : linéaire ou logarithmique

- Critère de choix

  → on va choisir le modèle le plus performant, c'est-à-dire celui dont le R^2 est le plus élevé

- Equations des modèles

  * dans le cadre linéaire : $$\text{y\_child}_j = \beta_0 + \beta_1 \, \text{y\_child\_avg}_j + \beta_2 \, \text{gini}_j + \epsilon_j$$

  * dans le cadre logarithmique : $$\log(\text{y\_child})_j = \beta_0 + \beta_1 \log(\text{y\_child\_avg})_j + \beta_2 \, \text{gini}_j + \epsilon_j$$

avec :

- β_0 ---------- la constante,
- y_child_j ---------- le revenu des individus enfants du pays j (en PPA)
- $\log(\text{y\_child})_j$ ---------- le revenu des individus enfants du pays j exprimé en log
- y_child_avg_j ---------- le revenu enfant moyen du pays j (en PPA)
- $\log(\text{y\_child\_avg})_j$ ------ le revenu enfant moyen du pays j exprimé en log
- gini_j ---------- l'indice de Gini du pays j
- ϵ_j ---------- l'erreur du modèle du pays j (les résidus)

2. Régression linéaire à 2 variables explicatives

- Modèle linéaire

A) Définition du modèle

```
# on régresse y_child en fonction de 'y_child_avg' et de 'gini'
reg_multi = ols('y_child ~ y_child_avg+gini', data=dcf).fit()
```

B) Tests du modèle

Test global : **test de Fischer**

H0 : $\beta_1 = \beta_2 = 0$ → tous les coef sont nuls
→ le modèle n'est pas significatif

H1 : $\beta_1 \neq 0$ et ou $\beta_2 \neq 0$ → au moins 1 coef est non nul
→ le modèle est significatif

F-statistic:	2.842e+06
Prob (F-statistic):	0.00

p = 0 donc < 0,05

→ on rejette H0
→ **le modèle est significatif**

Test de significativité des variables explicatives : **test de Student**

Variable "revenu_moyen"	Variable "indice de Gini"
H0 : $\beta_1 = 0$	H0 : $\beta_2 = 0$
H1 : $\beta_1 \neq 0$	H1 : $\beta_2 \neq 0$

Si on rejette H0, alors la variable est significative

	coef	std err	t	P>|t|
const	-0.0022	13.945	-0.000	1.000
y_child_avg	1.0000	0.000	2233.299	0.000
gini	0.0049	32.934	0.000	1.000

Variable **"revenu_moyen"**
p = 0 : variable **significative**

Variable "**indice de Gini**"
p = 1 : variable **non significative**

C) Performance du modèle

R-squared:	0.497
Adj. R-squared:	0.497

$R^2 \simeq 0,5$
R^2 correspond au η^2 de l'ANOVA

Or seule la variable "revenu_moyen" est ici significative
→ résultat est cohérent

Ainsi, seule une variable est significative avec le modèle linéaire → effectuons une nouvelle régression en utilisant les logarithmes

2. Régression linéaire à 2 variables explicatives

- Modèle logarithmique

A) Définition du modèle

```
# on régresse y_child en fonction de 'y_child_avg' et de 'gini'
reg_multi = ols('y_child ~ y_child_avg+gini', data=dcf).fit()
```

B) Tests du modèle

Test global : **test de Fischer**

H0 : $\beta_1 = \beta_2 = 0$ → tous les coef sont nuls
→ le modèle n'est pas significatif

H1 : $\beta_1 \neq 0$ et ou $\beta_2 \neq 0$ → au moins 1 coef est non nul
→ le modèle est significatif

F-statistic:	7.623e+06
Prob (F-statistic):	0.00

p = 0 donc < 0,05

→ on rejette H0
→ **le modèle est significatif**

Test de significativité des variables explicatives : **test de Student**

Variable "revenu_moyen"	Variable "indice de Gini"
H0 : $\beta_1 = 0$	H0 : $\beta_2 = 0$
H1 : $\beta_1 \neq 0$	H1 : $\beta_2 \neq 0$

Si on rejette H0, alors la variable est significative

	coef	std err	t	P>|t|
Intercept	0.4648	0.003	162.145	0.000
ln_y_child_avg	0.9861	0.000	3618.416	0.000
gini	-1.6350	0.003	-470.506	0.000

Variable **"log(revenu_moyen)"**
p = 0 : variable **significative**

Variable "**indice de Gini**"
p = 1 : variable **significative**

C) Performance du modèle

R-squared:	0.726
Adj. R-squared:	0.726

$R^2 \simeq 0{,}73$

donc près de 73% de la variance totale est expliquée par les 2 variables explicatives

2. Régression linéaire à 2 variables explicatives

Résultats régression à 2 variables explicatives

Modèle linéaire	Modèle logarithmique
- 1 seule variable est significative	- les 2 variables sont significatives
- $R^2 = 0,5$	- $R^2 = 0,73$

➔ On retient donc le modèle logarithmique pour la suite de l'analyse

Analyse des valeurs atypiques et influentes

- Valeurs atypiques sur les variables explicatives : le levier

→ pour l'observation i, le levier exprime la distance entre :
- ce "point" i
- le centre de gravité du nuage de points des variables explicatives

→ seuil critique : levier > $2 * \frac{p+1}{n}$ avec p : nombre de variables explicatives
n: taille de l'échantillon

2. Régression linéaire à 2 variables explicatives

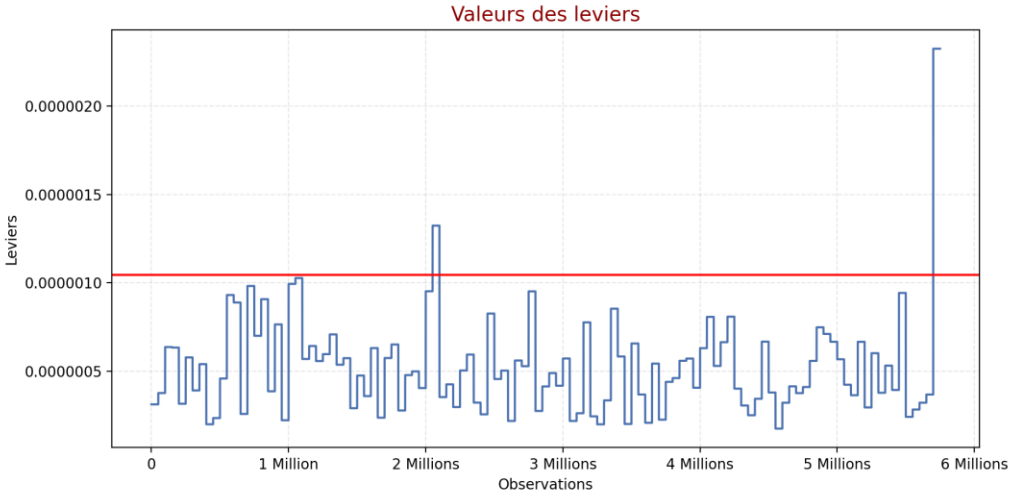
→ On intègre les leviers dans notre df

```python
# creation d'une instance d'Influence
infl = reg_multi_ln.get_influence()

# insert des leviers dans le df analyses
analyses['levier'] = infl..hat_matrix_diag

# seuil des leviers
seuil_levier = 2*((p+1)/n)
print("Seuil des leviers : {:.10f}".format(float(seuil_levier)))
```

```
Seuil des leviers : 0.0000010435
```

Remarque :

- il n'y a qu'une seule valeur de levier par pays
(car un seul revenu moyen et un seul indice de Gini par pays)



❑ Pays avec un levier important :

```python
# pays ont des leviers importants
pays_leviers = analyses.groupby('pays')['levier'].max().sort_values(ascending=False)
synthese_res['pays_leviers_importants'] = pays_leviers
pays_leviers.head(10).apply(lambda x: '{:.8f}'.format(float(x)))
```

Afrique du Sud	0.00000232
Honduras	0.00000132
Colombie	0.00000103
République Démocratique du Congo	0.00000099
République Centrafricaine	0.00000098
Guatemala	0.00000095
Kenya	0.00000095
États-Unis	0.00000094
Bolivie	0.00000093
Chili	0.00000091

❑ Pays avec un levier supérieur au seuil critique

Seuil des leviers = 0.0000010435

donc 2 pays avec des valeurs de levier atypiques :

- Afrique du sud
- Honduras

2. Régression linéaire à 2 variables explicatives

- Valeurs atypiques sur la variable à expliquer : le résidus studentisé

→ Le résidu standardisé (ou résidu studentisé interne) représente l'importance du résidu observé : $\hat{\varepsilon}_i = y_i - \hat{y}_i$

Or si par hypothèse la variance de l'erreur est constante, ce n'est pas le cas pour le résidu

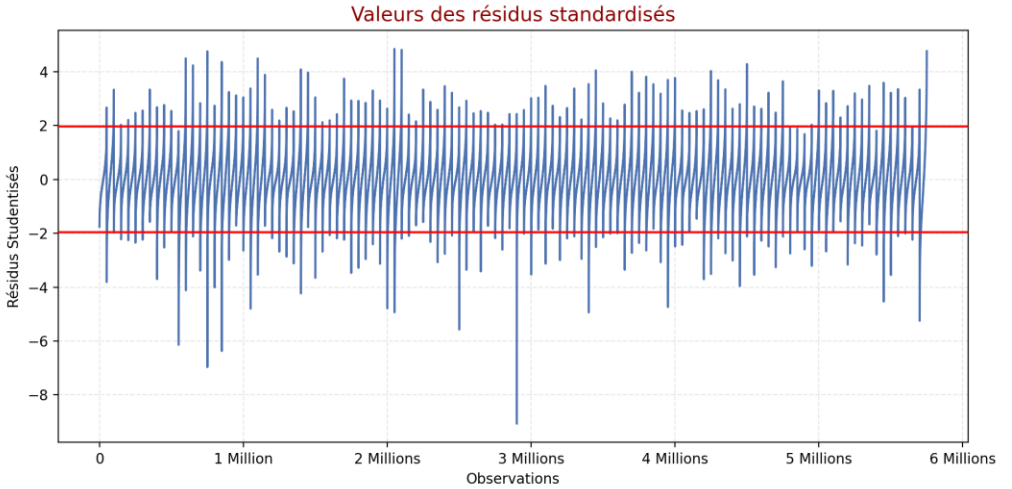
→ il faut donc normaliser le résidu par son écart-type pour rendre les écarts comparables d'une observation à l'autre

→ seuil critique : $|t_i| > t_{1-\frac{\alpha}{2}}\ (n - p - 1)$

On intègre les résidus standardisés dans notre df

```
# creation d'une instance d'Influence
infl = reg_multi_ln.get_influence()
# insert des résidus studentisé dans le df analyses
analyses['res_stud'] = infl.resid_studentized_internal

# seuil résidus studentisés
seuil_stud = t.ppf(1-alpha/2,n-p-1)
print('Seuil des résidus standardisés : {:.4f}'.format(float(seuil_stud)))
```

```
Seuil des résidus standardisés : 1.9600
```



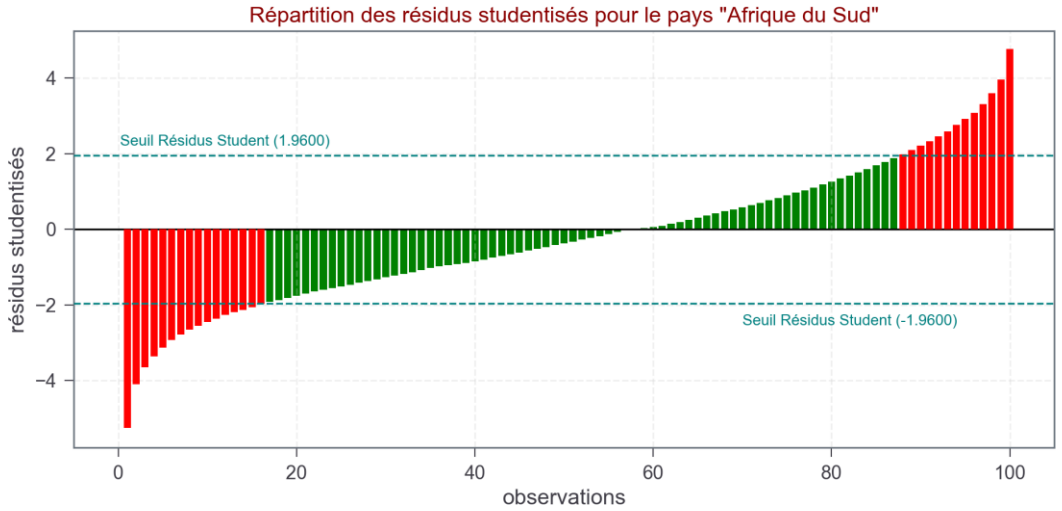
2. Régression linéaire à 2 variables explicatives

Nombre de résidus dans la région critique :

→ 316500, soit 5.5% du dataset

113 pays concernés (affichage des 10 pays présentant le plus de valeurs hors seuil)

Afrique du Sud	14500
Honduras	11000
Bolivie	10000
Colombie	8500
Panama	8000
Brésil	7500
Guatemala	7000
République Centrafricaine	7000
Paraguay	6500
Nicaragua	6500



- Valeurs influentes

→ influence d'une observation s'effectue à l'aide de la distance de Cook

Pour évaluer l'influence d'un point i sur la régression, on compare les prédictions obtenues avec :

- le modèle complet
- le modèle sans le point i

} si la différence est élevée, alors le point i joue un rôle important dans l'estimation des coefficients

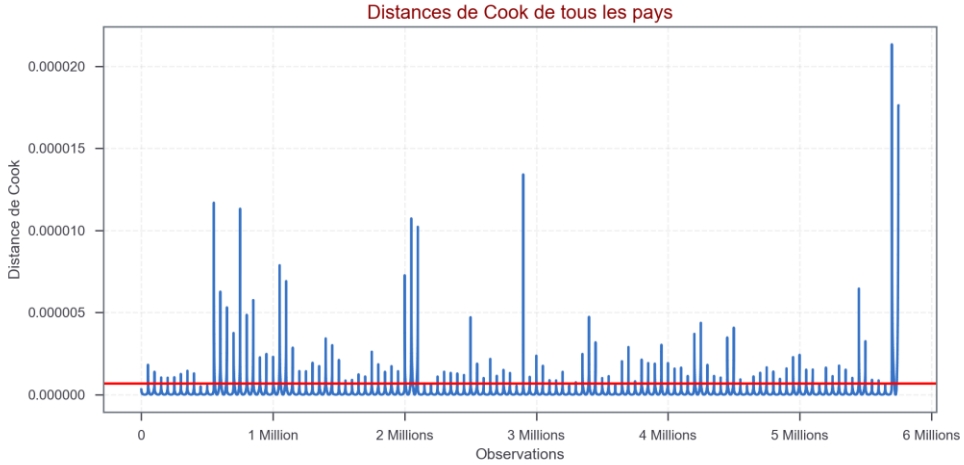
→ seuil critique : on considère une observation influente lorsque $\mathcal{D}_i > \frac{4}{n-p}$

2. Régression linéaire à 2 variables explicatives

On intègre les distance de Cook dans notre df

```
# creation d'une instance d'Influence
infl = reg_multi_ln.get_influence()

# cacul de la distance de cook de chaque observation
# avec cooks[0] : la distance
# et cooks[1] : p_value associée
cooks = infl.cooks_distance
dis_cooks_l = cooks[0]

# insert de la distance de cook dans le df analyses
analyses['dis_cooks'] = dis_cooks_l

# seuil distance de cook
seuil_cooks = 4/(n-p)
print('Seuil de la distance de Cook : {:.8f}'.format(float(seuil_cooks)))
```

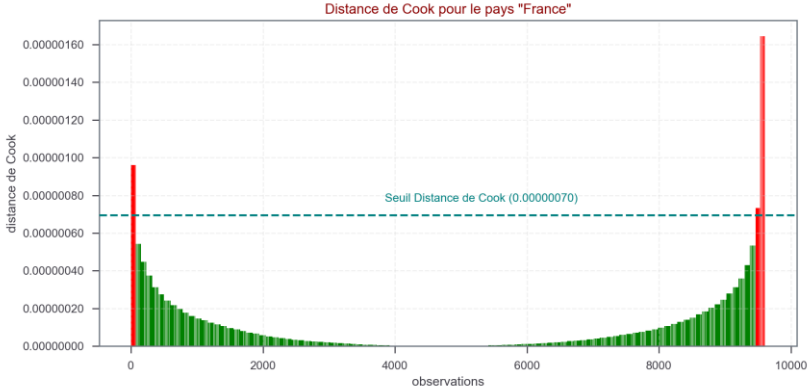
```
Seuil de la distance de Cook : 0.00000070
```



Présence de 325500 points influents, soit 5.66% du dataset

97 pays concernés

Afrique du Sud	30500
Honduras	22000
Colombie	15500
Bolivie	15500
Brésil	13000
République Centrafricaine	13000
Guatemala	12500
Panama	12000
Chili	10000
États-Unis	10000



2. Régression linéaire à 2 variables explicatives

- Synthèse concernant la détection des outliers

→ création d'un dataframe récapitulatif indiquant le nombre et le type d'outliers par pays, ordonné par ordre décroissant du nombre total d'outliers

code_pays	pays	pop	gini	gdp	y_child_avg	elas	ln_y_child_avg	levier	res_stud	dis_cooks	somme_outliers
ZAF	Afrique du Sud	49 779 471	0.6698	9602.00	5617.90	0.68	8.633714	50000	14500	30500	95000
HND	Honduras	7 980 955	0.6017	3628.00	3296.27	0.66	8.100546	50000	11000	22000	83000
BOL	Bolivie	9 721 454	0.5615	3950.00	3016.26	0.87	8.011774	0	10000	15500	25500
COL	Colombie	44 254 975	0.5693	8185.00	3547.01	1.10	8.173859	0	8500	15500	24000
BRA	Brésil	192 030 362	0.5445	9559.00	4807.48	0.64	8.477929	0	7500	13000	20500

✓ On peut vérifier que les pays avec le plus d'outliers sont ceux où l'indice de Gini est le plus élevé

- Traitement des outliers

→ pour un traitement efficace, il faudrait davantage d'informations, comme par exemple la composition du portefeuille clients par pays

→ en l'absence d'informations complémentaire, on va appliquer une correction "automatique" :

↳ on supprime les observations qui contiennent des valeurs atypiques (levier ou résidus standardisés) ET des valeurs influentes (distance de Cook)

Nombre d'outliers supprimés : 257500, soit 0.04% du dataset

Nombre de pays concernés : 97

	nb_valeurs	part_en_pct
Afrique du Sud	30500	11.84%
Honduras	22000	8.54%
Bolivie	10000	3.88%
Colombie	8500	3.30%
Panama	8000	3.11%

R-squared:	0.800	F-statistic:	1.098e+07
Adj. R-squared:	0.800	Prob (F-statistic):	0.00

	coef	std err	t	P>|t|
Intercept	0.3092	0.002	124.238	0.000
ln_y_child_avg	1.0003	0.000	4323.928	0.000
gini	-1.5312	0.003	-489.841	0.000

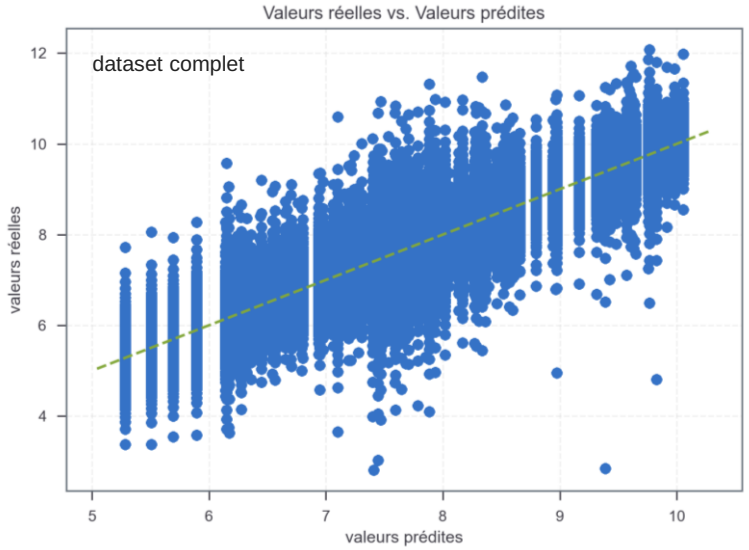
✓ En supprimant seulement 0.04% des valeurs initiales, **la variance des revenus expliquée par le modèle passe de 73% à 80%.**

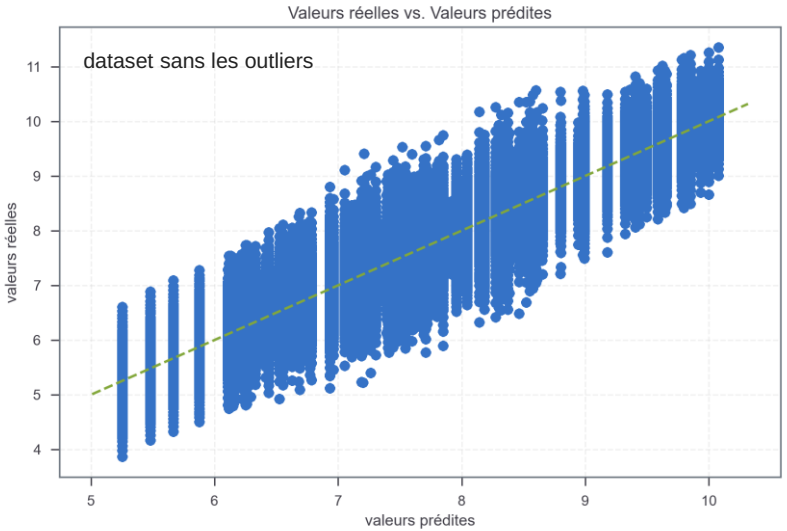
2. Régression linéaire à 2 variables explicatives

Hypothèses d'application du modèle linéaire

- Condition de linéarité : relation linéaire entre la variable à expliquer et les variables explicatives

→ Méthode graphique : si linéarité, alors les points de la droite "prédictions / valeurs réelles" devraient suivre la droite d'équation y=x





2. Régression linéaire à 2 variables explicatives

- Absence de colinéarité entre les variables explicatives

→ Plusieurs méthodes possibles : exemple du VIF

- le Facteur d'Inflation de la Variance (VIF)

↳ évaluer la liaison d'une variable explicative avec l'ensemble des autres variables

* critère strict : si VIF > 5 → présence de colinéarité

* critère plus permissif : si VIF > 10 → présence de colinéarité

→ Application à notre modèle (dataset complet) :

```
# variables explicatives
X = dcf[["ln_y_child_avg","gini"]]
# ajout de la constante
X = sm.add_constant(X)

# df facteurs VIF
vif_data = pd.DataFrame()
vif_data["feature"] = X.columns[1:]

# calcul du facteur VIF pour chaque variable
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(1,len(X.columns))]
```

	feature	VIF
0	ln_y_child_avg	1.073297
1	gini	1.073297

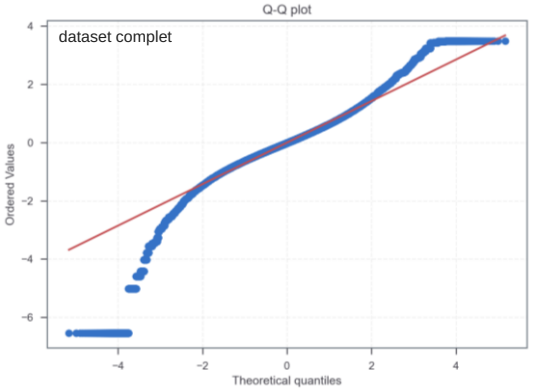
VIF < 5 pour les 2 variables

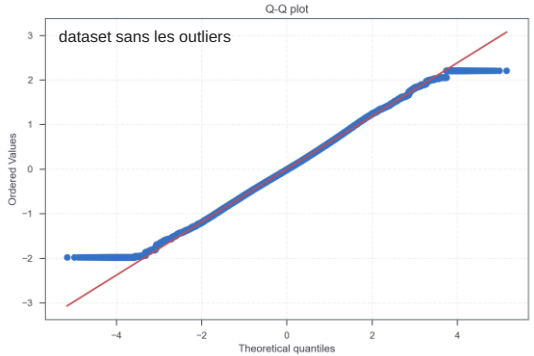
→ Absence de colinéarité entre les variables explicatives

2. Régression linéaire à 2 variables explicatives

- Normalité des distributions

→ Analyse des résidus : → test graphique : Q-Q plot + tests statistiques de normalité





Tests statistiques de normalité :

- test de Jarque-Bera
- test de Kolmogorov-Smirnov

H0 : Les résidus suivent une loi normale

H1 : Les résidus ne suivent pas une loi normale

si p_value < 0.05 : on rejette H0 :

→ les données ne suivent pas une loi normale

si p_value >= 0.05 : on accepte H0 :

→ les données suivent une loi normale

Dataset complet :

Jarque-Bera test:
* statistic: 1732751.0608
* p-value: 0.0

Kolmogorov-Smirnov test:
* statistic: 0.1079
* p-value: 0.0000

Dataset sans les outliers :

Jarque-Bera test:
* statistic: 5050.8948
* p-value: 0.0

Kolmogorov-Smirnov test:
* statistic: 0.1253
* p-value: 0.0000

Dans tous les cas, p_value proche de 0

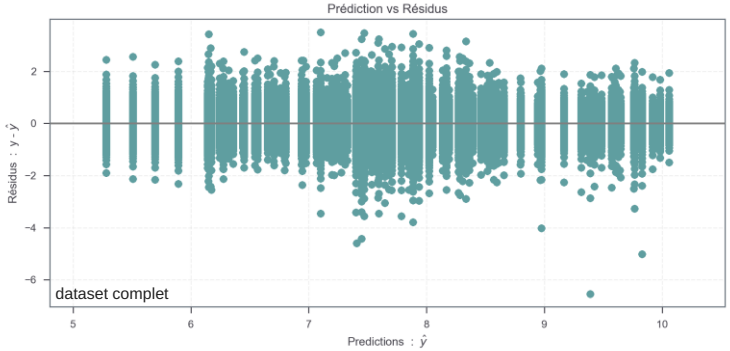
=> on rejette H0

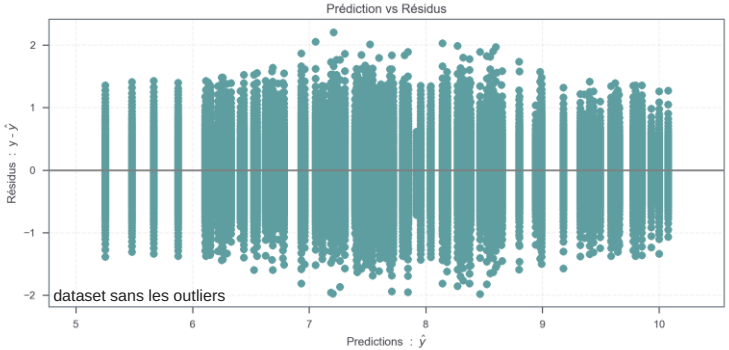
=> les distributions ne suivent pas une loi normale

2. Régression linéaire à 2 variables explicatives

- Homoscédasticité des résidus

→ Analyse des résidus : → test graphique : résidus / valeurs prédites + tests statistiques d'homoscédasticité





Tests statistiques d'homoscédasticité :
- Breusch Pagan

H0 : Homoscédasticité des résidus
H1 : Hétéroscédasticité des résidus

si p_value < 0.05 : on rejette H0 :
→ les variances ne sont pas constantes

si p_value >= 0.05 : on accepte H0 :
→ les variances sont constantes

Dataset complet :
* p-value: 0.0

Dataset sans les outliers :
* p-value: 0.0

Dans tous les cas, p_value proche de 0
=> on rejette H0
=> Hétéroscédasticité des résidus

Remarque concernant ces hypothèses

2 hypothèses ne sont pas respectées :
- normalité des résidus
- homoscédasticité des résidus

Mais le modèle de régression est un modèle dit "robuste"
→ lorsque l'échantillon est suffisamment important comme ici, le modèle est capable de supporter des écarts importants vis-à-vis du respect de ces 2 contraintes

→ les résultats restent interprétables

3. Régression linéaire à 3 variables explicatives

3ème modèle avec trois variables explicatives : le revenu moyen l'indice de Gini et la classe de revenus des parents

Choix du modèle : linéaire ou logarithmique

- Critère de choix

  → on va choisir le modèle le plus performant, c'est-à-dire celui dont le R^2 est le plus élevé

- Equations des modèles

  * dans le cadre linéaire : $$\text{y\_child}_j = \beta_0 + \beta_1 \, \text{y\_child\_avg}_j + \beta_2 \, \text{gini}_j + \beta_3 \, \text{c\_i\_parent}_j + \epsilon_j$$

  * dans le cadre logarithmique : $$\log(\text{y\_child})_j = \beta_0 + \beta_1 \log(\text{y\_child\_avg})_j + \beta_2 \, \text{gini}_j + \beta_3 \, \text{c\_i\_parent}_j + \epsilon_j$$

avec :

- β_0 ---------- la constante,
- y_child_j ---------- le revenu des individus enfants du pays j (en PPA)
- $\log(\text{y\_child})_j$ ---------- le revenu des individus enfants du pays j exprimé en log
- y_child_avg_j ---------- le revenu enfant moyen du pays j (en PPA)
- $\log(\text{y\_child\_avg})_j$ ---------- le revenu enfant moyen du pays j exprimé en log
- gini_j ---------- l'indice de Gini du pays j
- c_i_parent_j ---------- le centile de revenus parents
- ϵ_j ---------- l'erreur du modèle du pays j (les résidus)

✓ Modèle linéaire :
- Modèle globalement significatif
- variable 'gini' non significative
- $R^2 = 0{,}52$

✓ Modèle logarithmique
- Modèle globalement significatif
- Toutes les variables sont significatives
- $R^2 = 0{,}78$

On retient donc le **modèle logarithmique** qui **explique plus de 78% de la variance totale**

3. Régression linéaire à 3 variables explicatives

Hypothèses d'application du modèle linéaire

- Hypothèse de linéarité :
  - ↳ graphiquement, il semble y avoir un lien linéaire entre la variable à expliquer et les variables explicatives
- Absence de colinéarité entre les variables explicatives
- Non normalité des variables explicatives
- Hétéroscédasticité des résidus

Analyse des valeurs atypiques et influentes

Performance du modèle sans les outliers

Suppression des observations à la fois atypiques ET influentes

(257340 obs, soit 0.04% du dataset)

R-squared:	0.839
Adj. R-squared:	0.839

$R^2 = 0,84$

En supprimant 0.04% des valeurs initiales, **la variance des revenus expliquée** par le modèle **passe de 77.70% à 84%.**

3. Régression linéaire à 3 variables explicatives

Interpréter les coefficients d'une régression linéaire

✓ De manière générale, les coefficients s'interprètent différemment selon le modèle de la régression linéaire

Soit β_1 le coefficient de la variable explicative X_1 :

- modèle niveau-niveau : si X_1 augmente de 1 unité, alors y varie de β_1 unités
- modèle log-log : si X_1 augmente de 1% alors y varie de β_1 %
- modèle log-niveau : si X_1 augmente de 1 unité, alors y varie de (β_1 *100)%
- modèle niveau-log : si X_1 augmente de 1%, alors y varie de (β_1/100) unités

✓ On récupère l'équation de régression de notre modèle

```
# equation de notre modèle
res_reg_log['equation_lineaire']
```

$$\log(\text{y\_child})_j = -0.0995 + 0.9861 \log(\text{y\_child\_avg})_j - 1.6355\, \text{gini}_j + 0.0112\, \text{c\_i\_parent}_j + \epsilon_j$$

✓ Interprétons le coefficient associé à l'indice de Gini : β = -1,6355

- On se situe dans le cadre log-niveau

  → si augmentation de 0,1 de l'indice de Gini (donc aggravation des inégalités), alors les revenus diminuent de (-1,6355*10)%, soit 16,35%

⇨ Ainsi, toutes choses égales par ailleurs, vivre dans un pays avec un indice de gini plus élevé de 0,1 se traduit par des revenus diminués de 16.35%

Conclusion

	HYPOTHESE				Tests et Performance		
	Linéarité	Non-Colinéarité	Normalité	Homoscédasticité	Fischer	Student	R^2
Anova 1 facteur ("pays")							
- Modèle linéaire			False	False	True		49,70%
- Modèle log			False	False	True		72,70%
Régression 2 variables							
X=["y_child_avg", "gini"]	False	True	False	False	True	False	49,70%
X=["log(y_child_avg)", "gini"]	False	True	False	False	True	True	72,60%
X=["log(y_child_avg)", "gini"] - outliers	True	True	False	False	True	True	79,90%
Régression 3 variables							
X=["y_child_avg", "gini", "c_i_parent"]	False	True	False	False	True	False	53%
X=["log(y_child_avg)", "gini", "c_i_parent"]	False	True	False	False	True	True	77,70%
X=["log(y_child_avg)", "gini", "c_i_parent"] - outliers	True	True	False	False	True	True	84%

+ 5 pts

✓ Supériorité du modèle logarithmique (la variable "indice de Gini" n'est pas significative selon le test de Student pour le modèle linéaire)

✓ Pour le modèle logarithmique à 3 variables explicatives :

- en gardant toutes les données : le modèle explique 77.70% de la variation des revenus
- en supprimant les outliers : le modèle explique 84% de la variation des revenus

→ Résultat principal : **le pays de naissance explique l'essentiel des revenus enfants** :

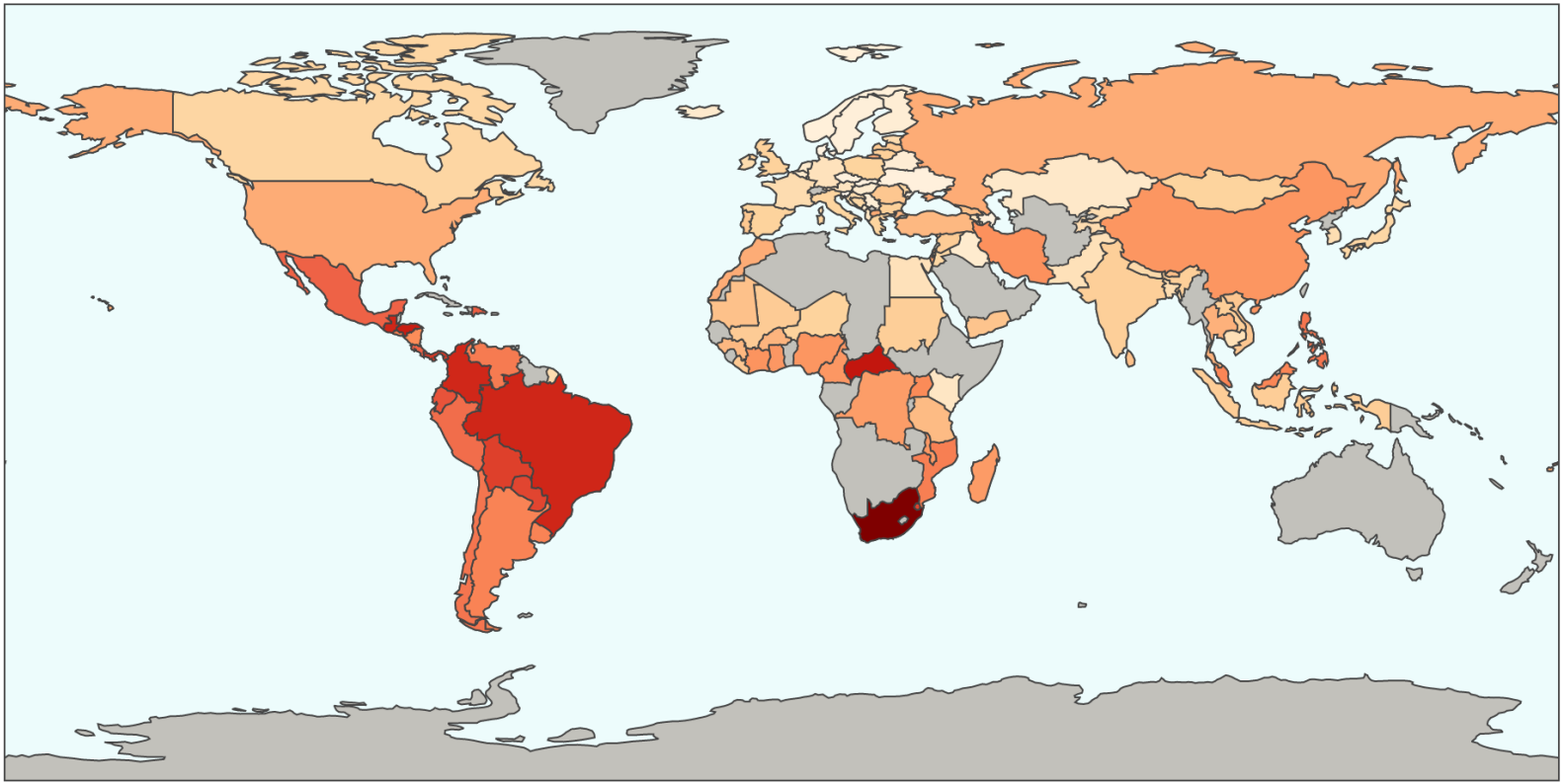
- l'ajout de la variable "indice de Gini" ne permet pas d'expliquer davantage le niveau des revenus d'une personne qu'avec la seule variable "pays"

- l'ajout des classes de revenus parents permet de gagner seulement 5 points de performance au modèle de régression

MERCI POUR VOTRE ATTENTION

QUESTIONS ?

Indices de Gini moyens (2004-2011)



Version interactive

Courbe de Lorenz des Etats-Unis

Courbe de Lorenz permet de mettre en évidence la répartition très inégalitaire des revenus aux Etats-Unis :

→ Les 20% les plus riches se partagent près de la moitié des revenus

→ Alors que les 20% les plus pauvres se partagent seulement 4,5% des revenus annuels