

# PRÉDIRE DES REVENUS

Sylvain Friot

16/10/2019

# DONNÉES UTILISÉES

# Population

Source : **FAO**

**Doublon sur la Chine  
supprimé**

# Distribution des revenus

Centiles par pays

**World Income Distribution**

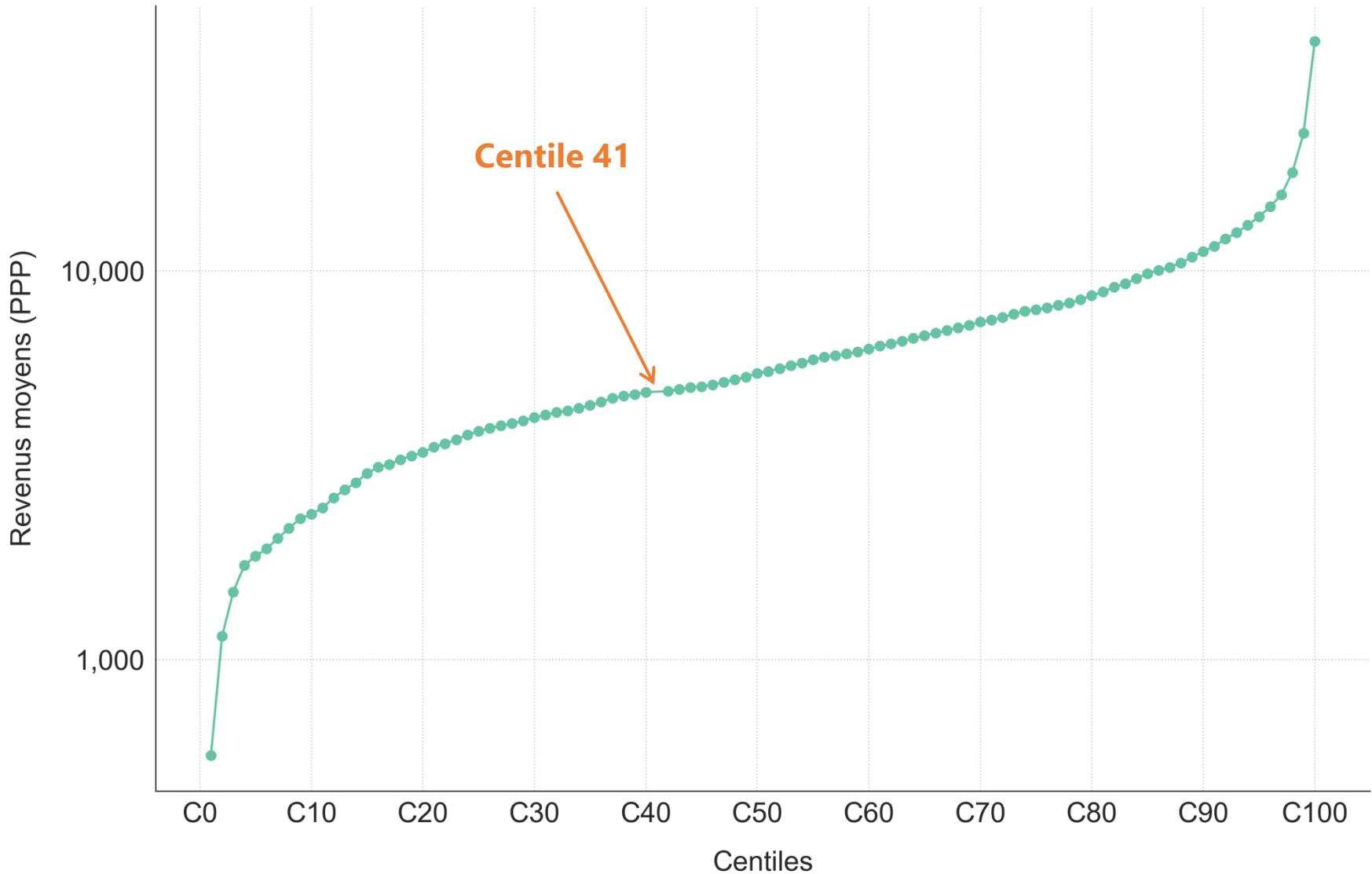
Concentration

**Banque Mondiale**



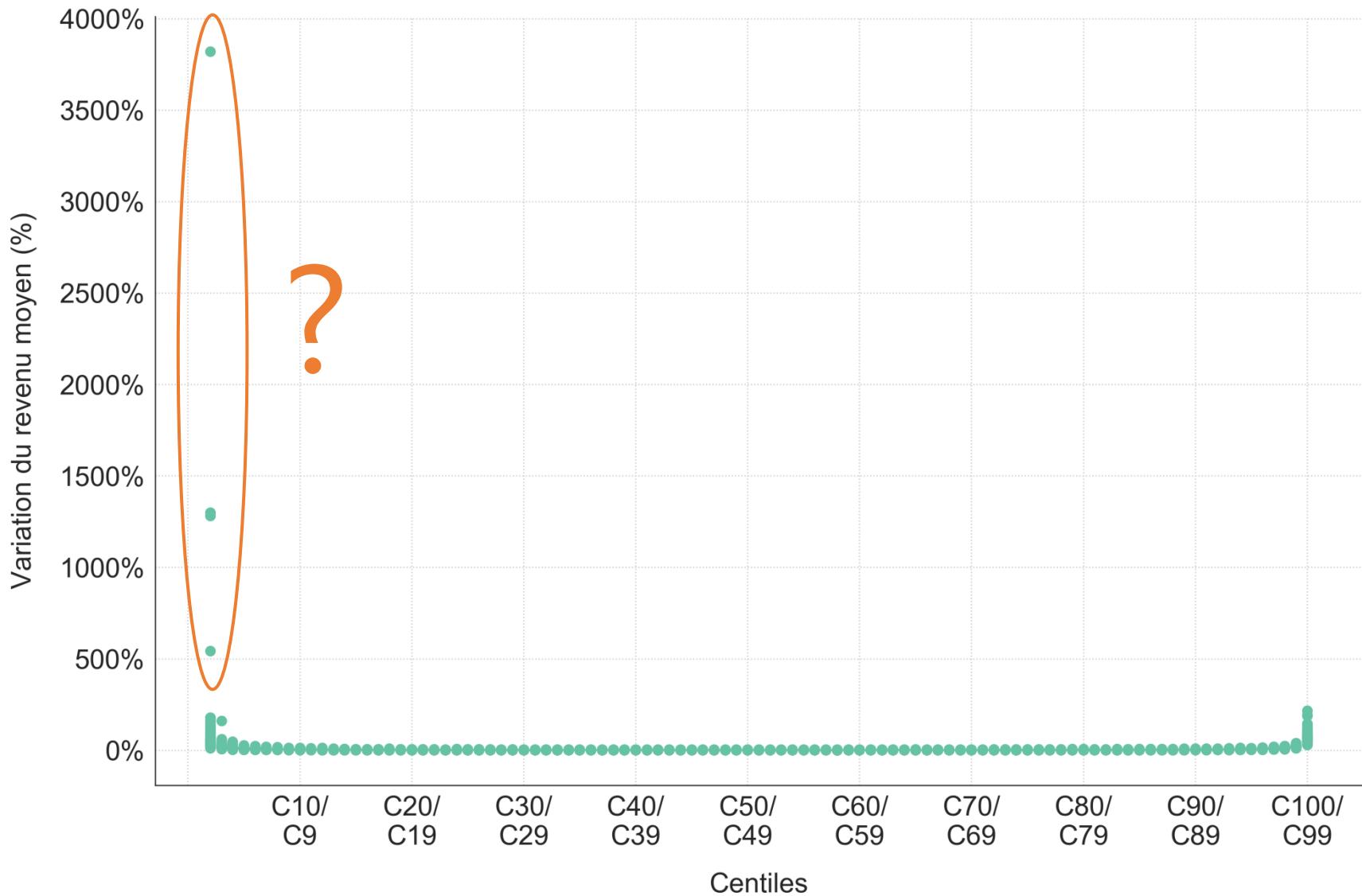
# World Income Distribution

## Distribution des revenus en Lituanie, par centiles



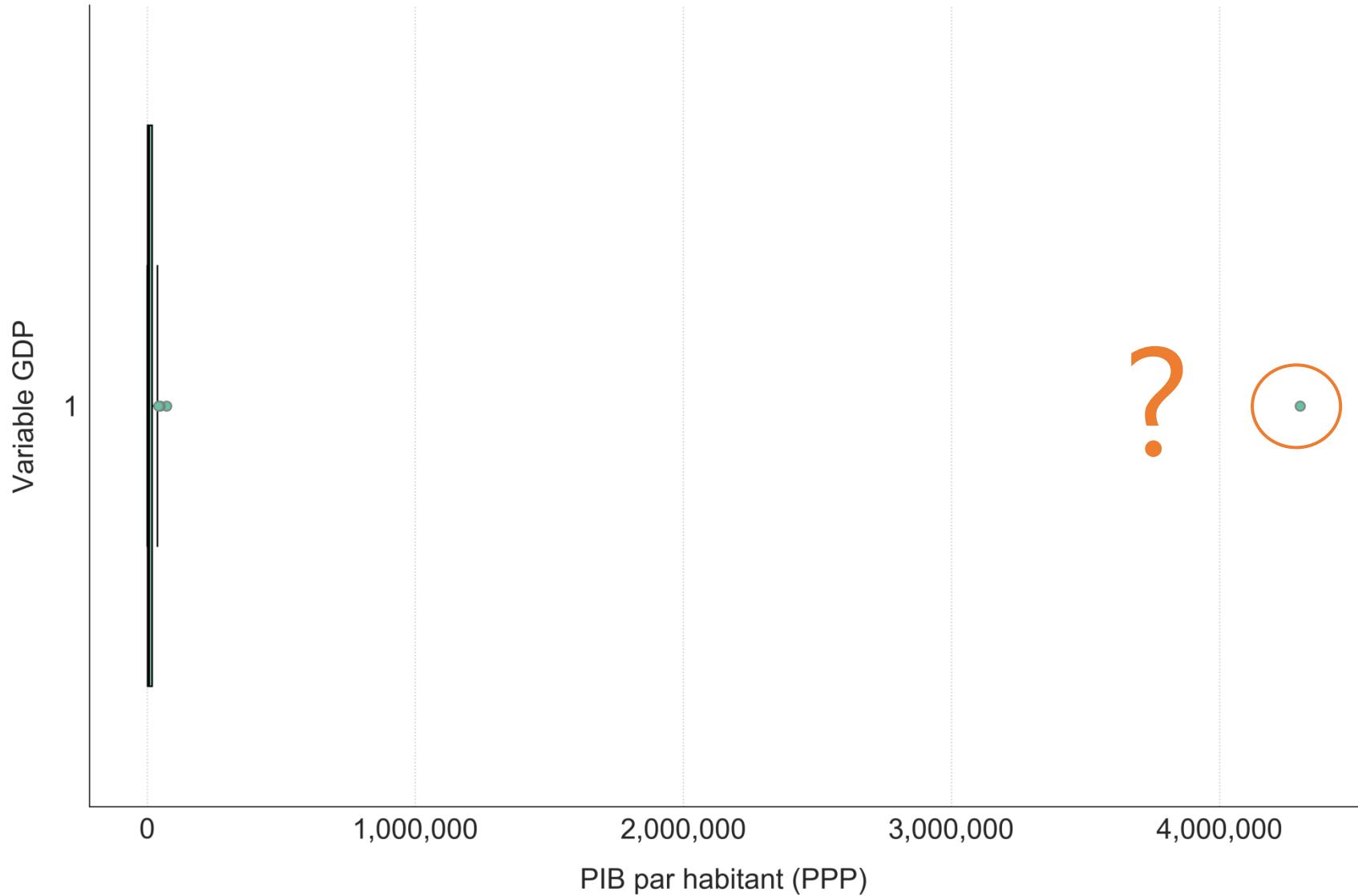
# World Income Distribution

## Dispersion des variations de revenus moyens entre centiles



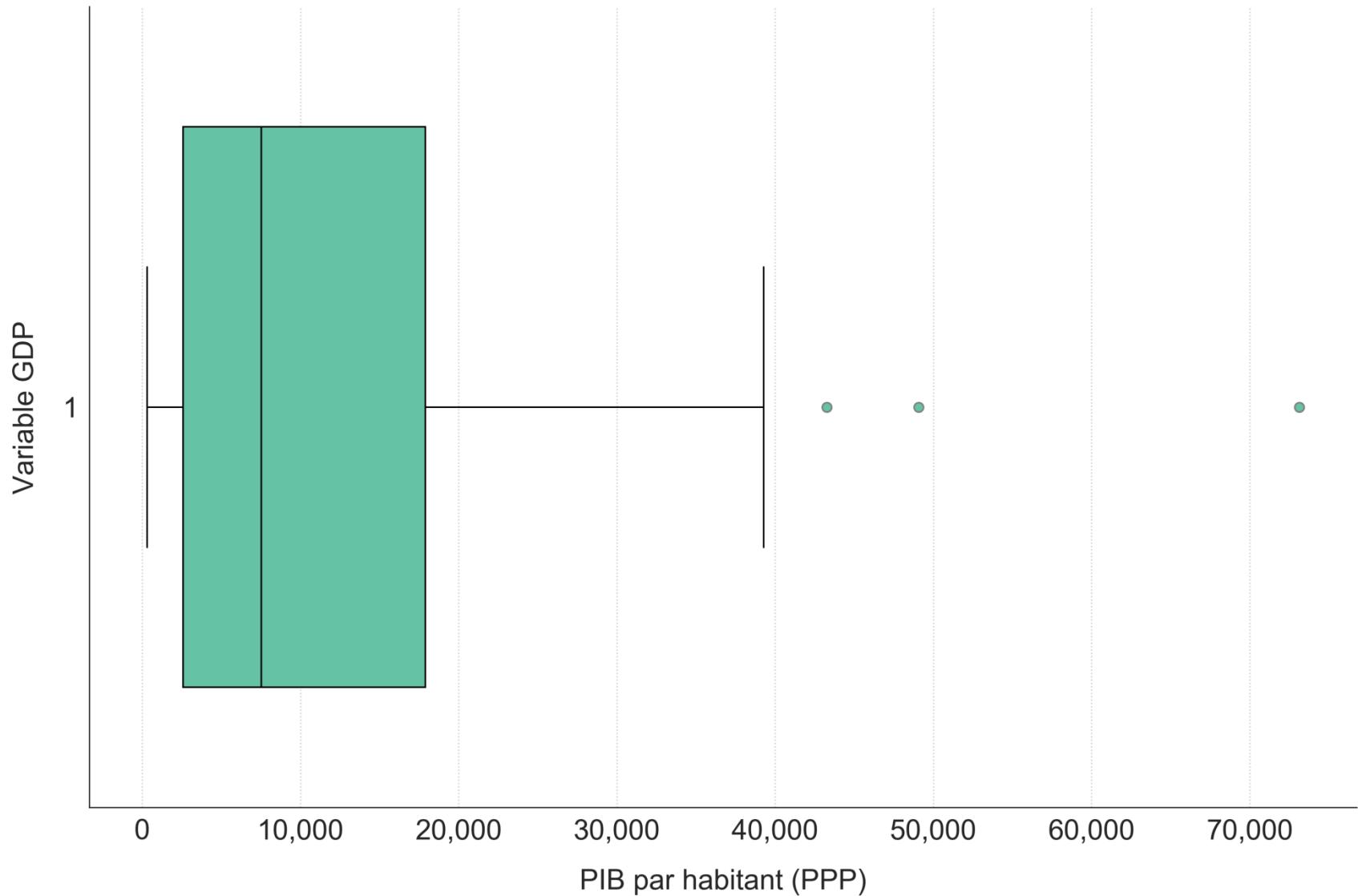
# World Income Distribution

Dispersion de la variable gdp



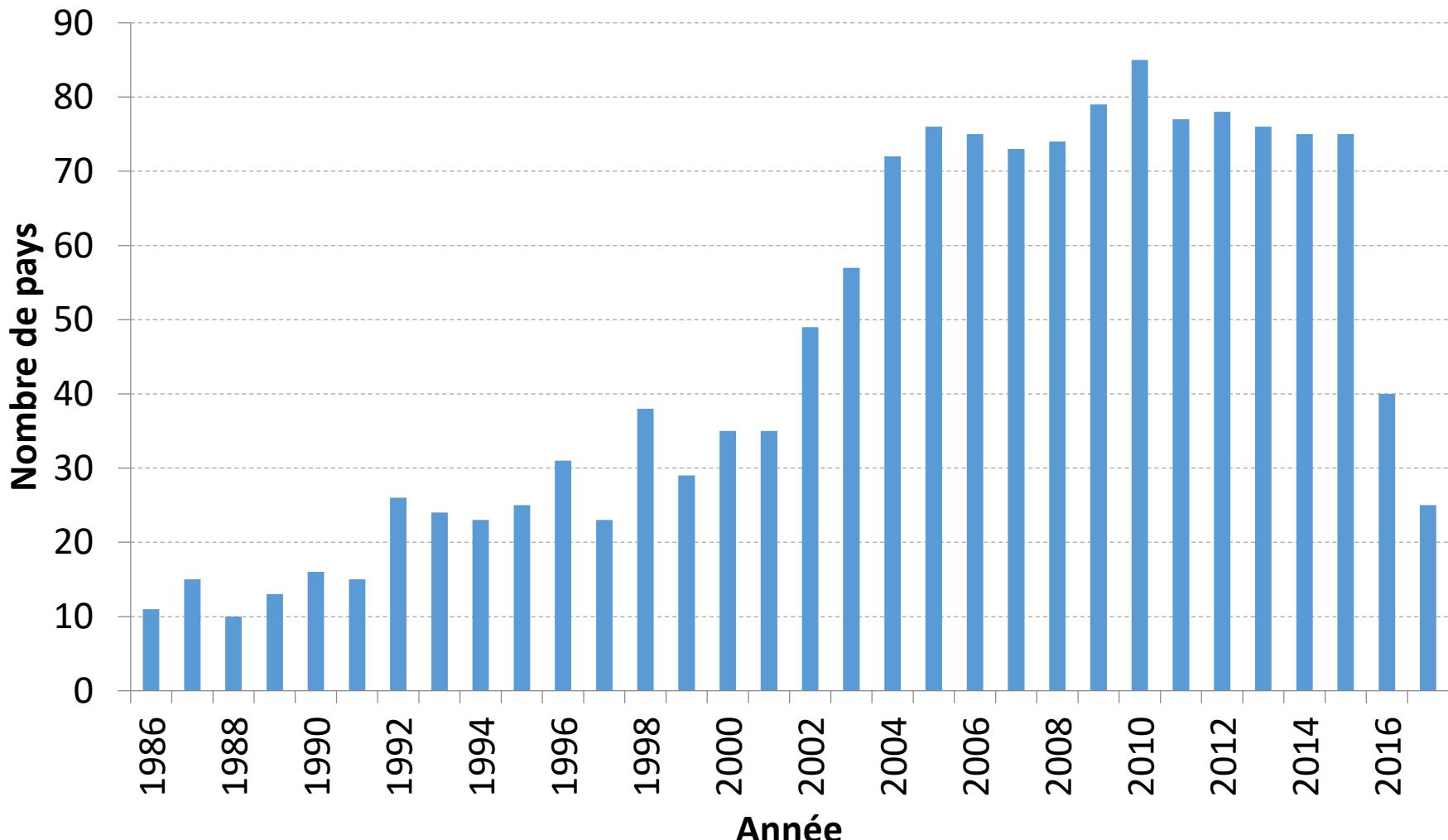
# World Income Distribution

Dispersion de la variable gdp



# Indice de Gini de la Banque Mondiale

Données disponibles par année



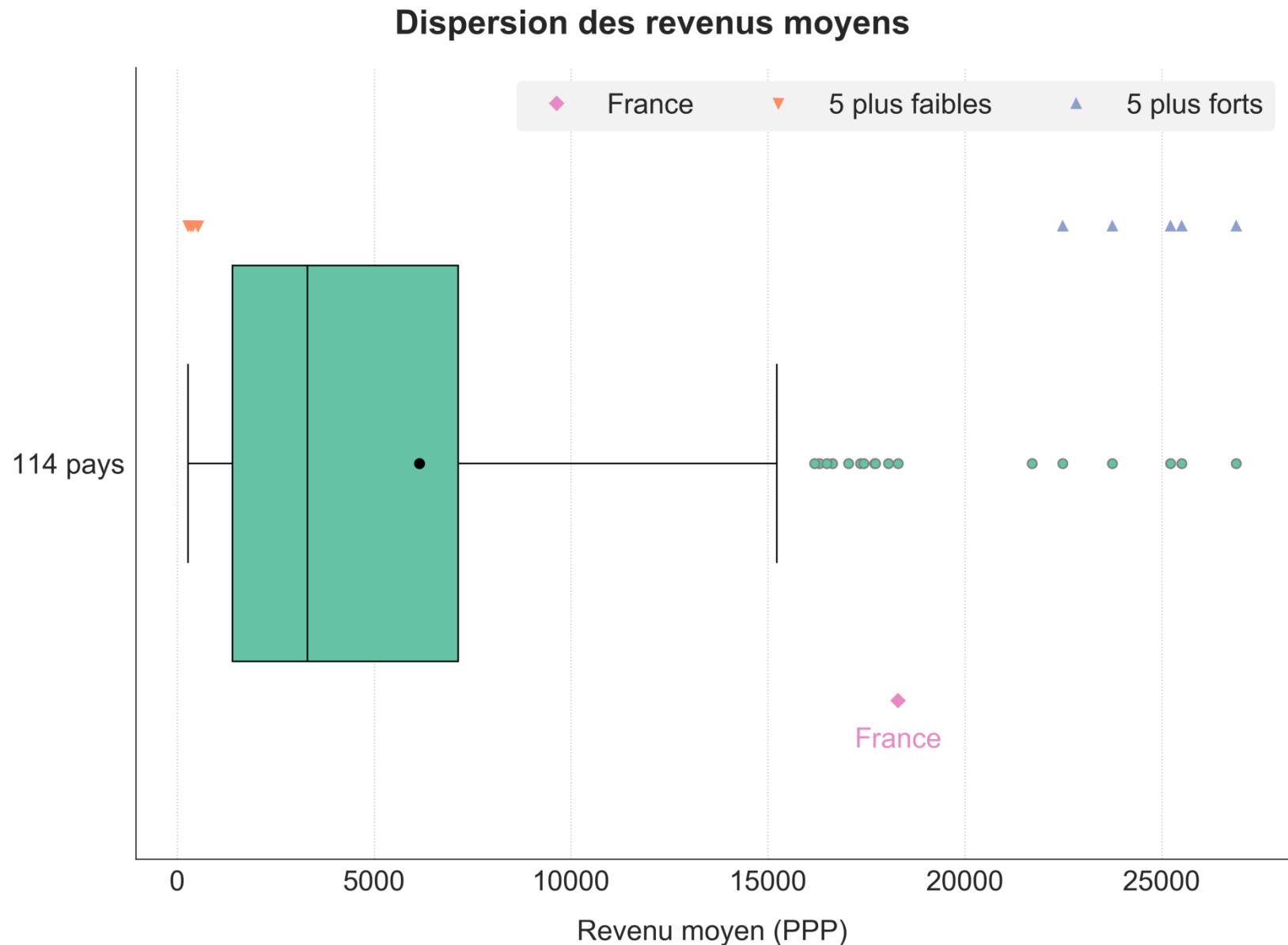
# Création du dataframe income

Colonne	Description	Source
Index : double index country_code et year	Combinaison : <ul style="list-style-type: none"><li>- Code du pays ;</li><li>- Année des données centiles</li></ul>	Centiles
revenu_moyen	Revenu moyen du pays en PPP	Centiles
gini	Indice de Gini estimé par la banque mondiale	Gini
gini_from_centiles	Indice de Gini estimé à partir des centiles des revenus	Centiles
population	Population du pays	Population
nb_annees_gini	Nombre d'années où l'indice de Gini est estimé par la Banque Mondiale entre 2003 et 2015	Gini

# Dataframe income

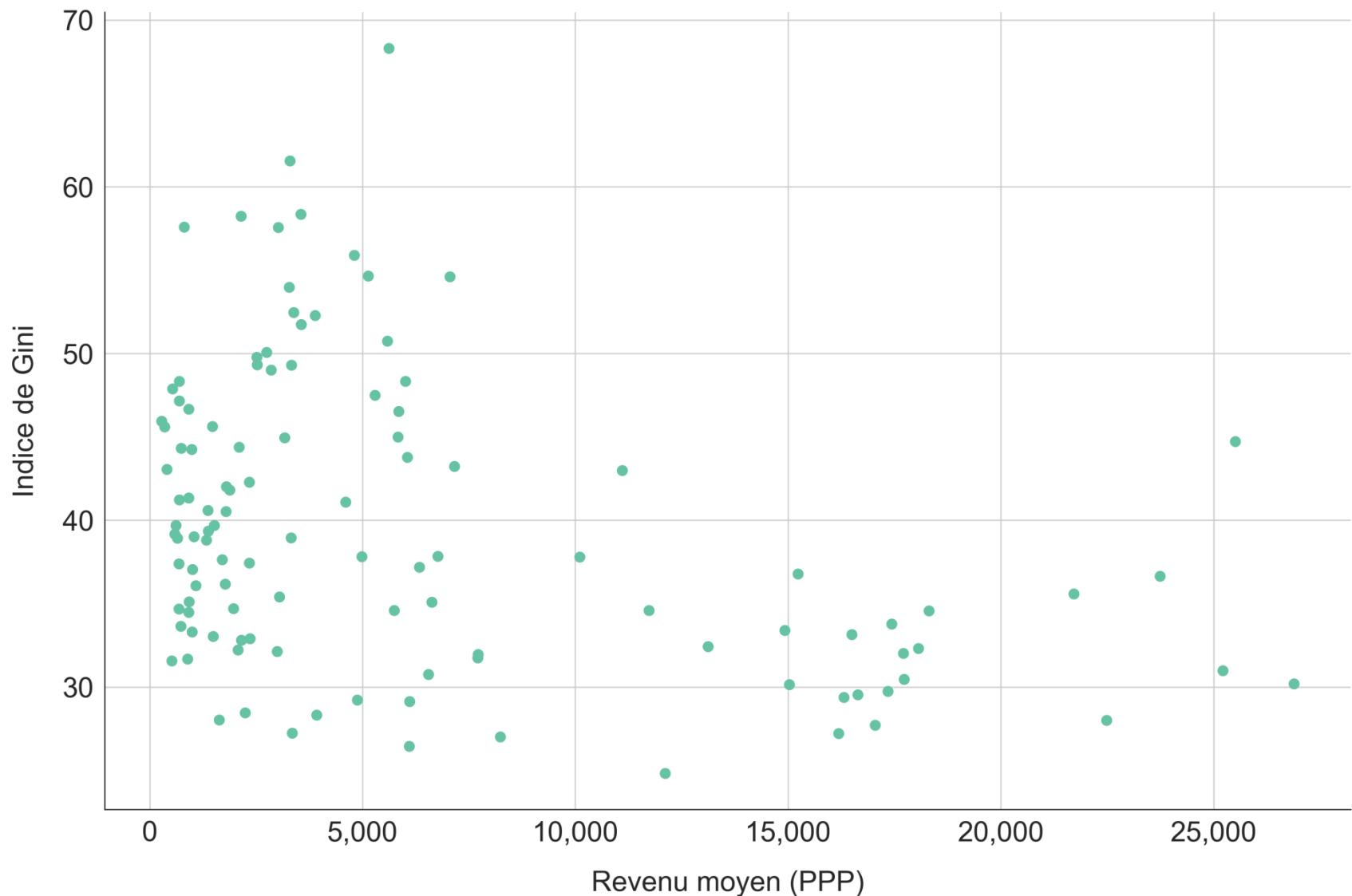


# Dataframe income



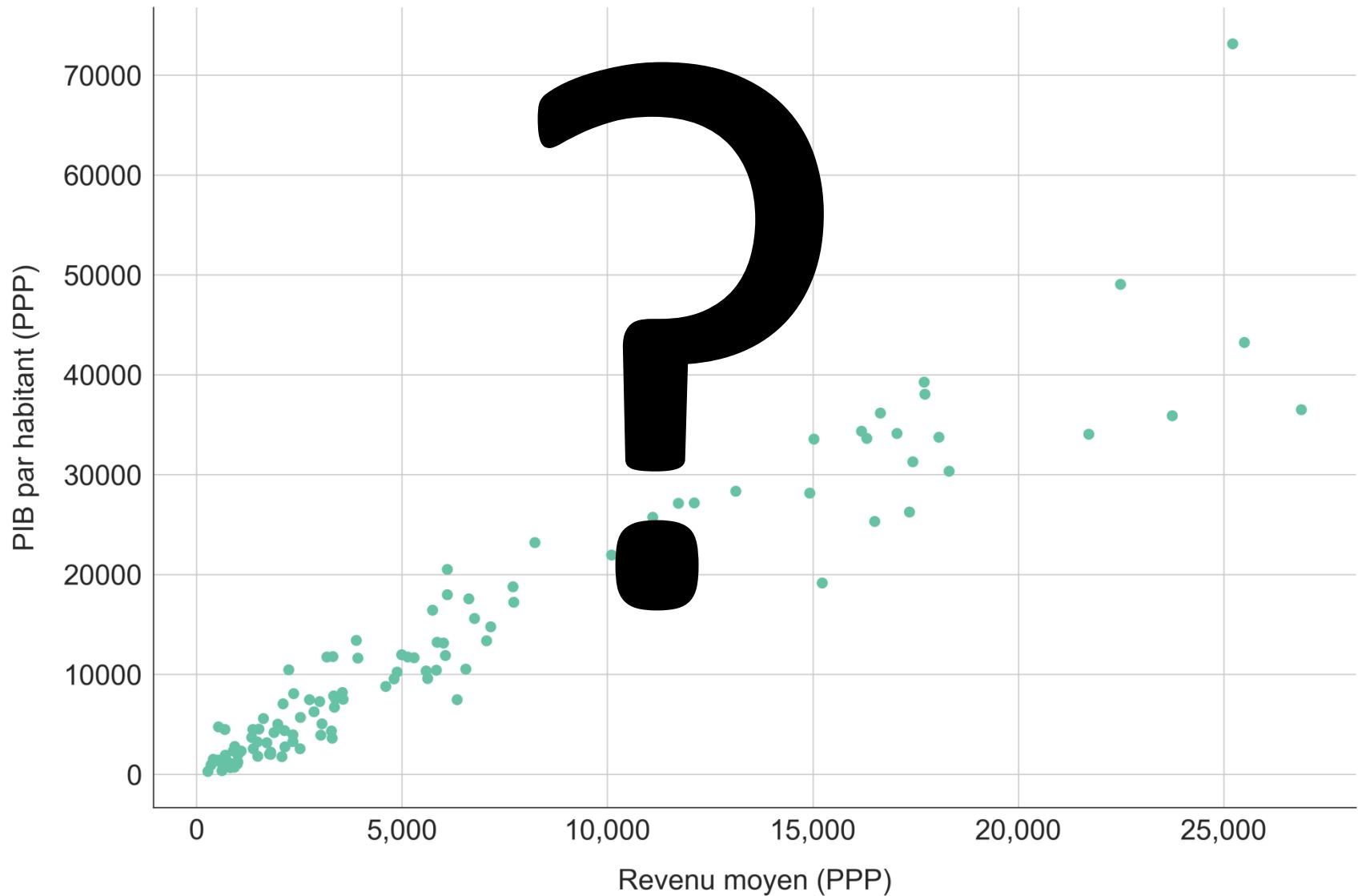
# Dataframe income

Diagramme de dispersion des revenus moyens et des indices de Gini



# Dataframe income

Diagramme de dispersion des revenus moyens et du PIB par habitant



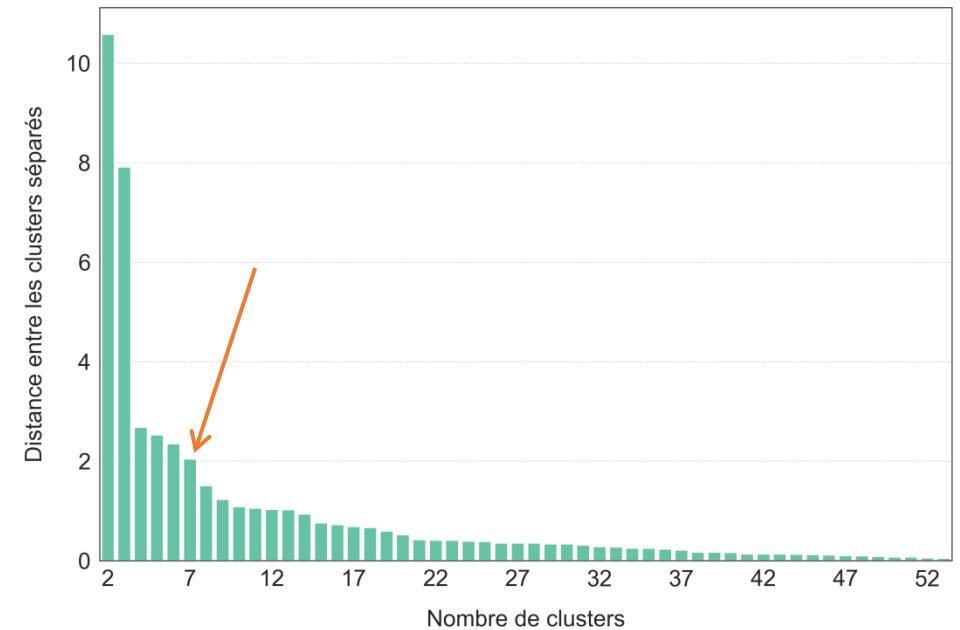
# **CONCENTRATION DES REVENUS**

## **COMPARAISONS ENTRE PAYS**

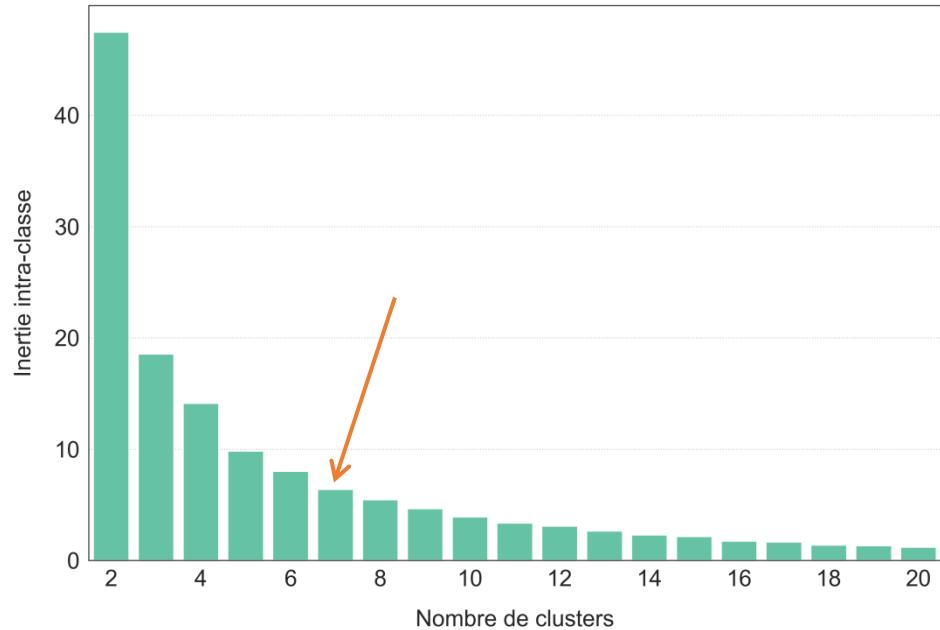
# Choix des pays

## Classification hiérarchique vs K-means

Dendrogramme - Choix du nombre de clusters



KMeans - Choix du nombre de clusters



7 clusters

# Choix des pays

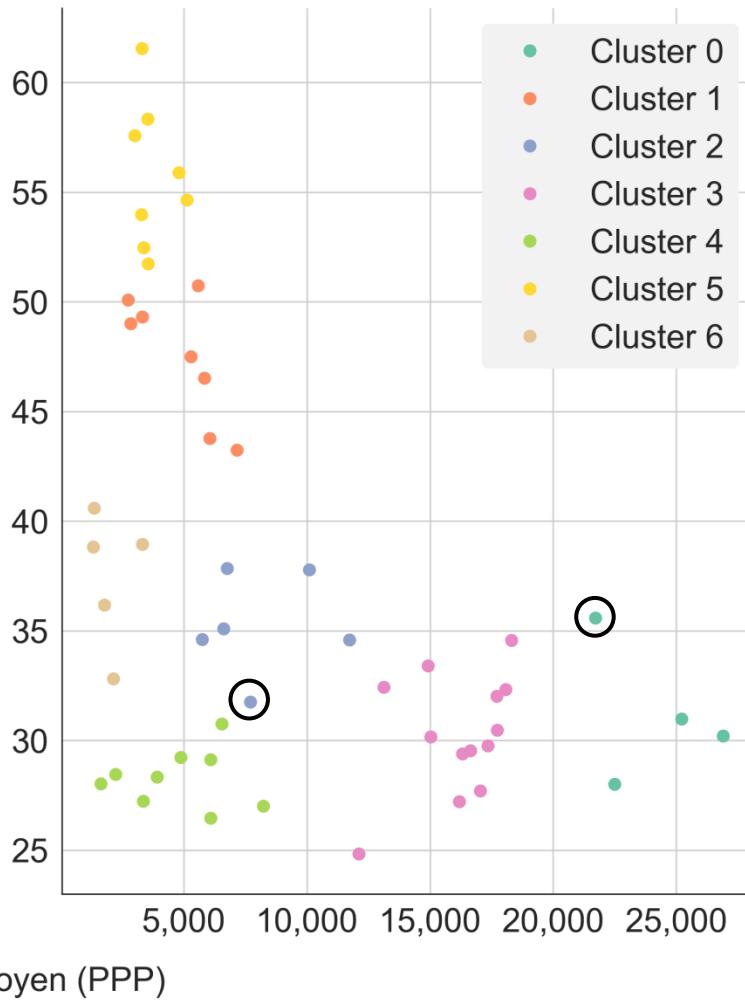
## Classification hiérarchique vs K-means

Comparaison des classifications

Classification hiérarchique



Algorithme kmeans

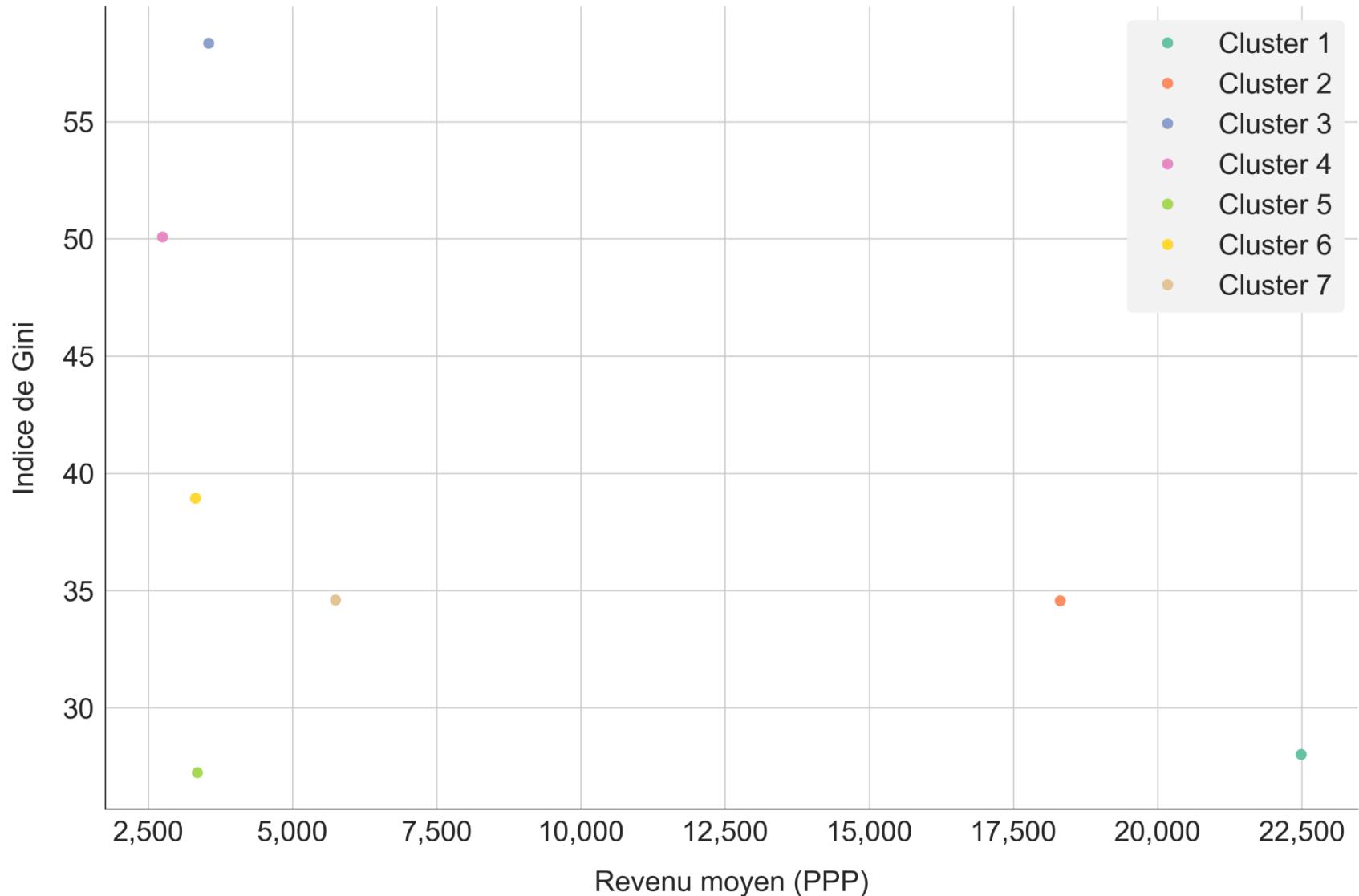


# Choix des pays

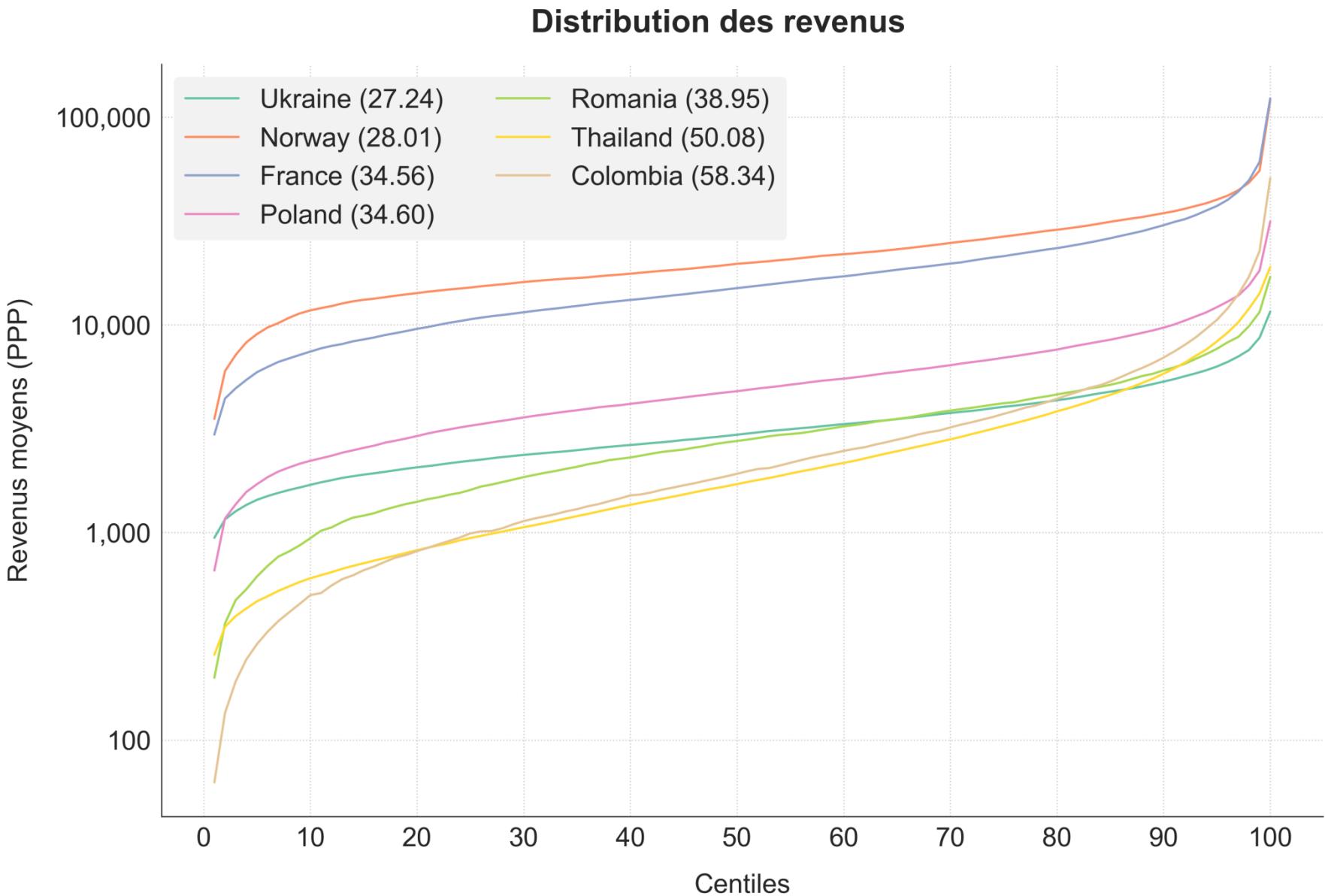


# Choix des pays

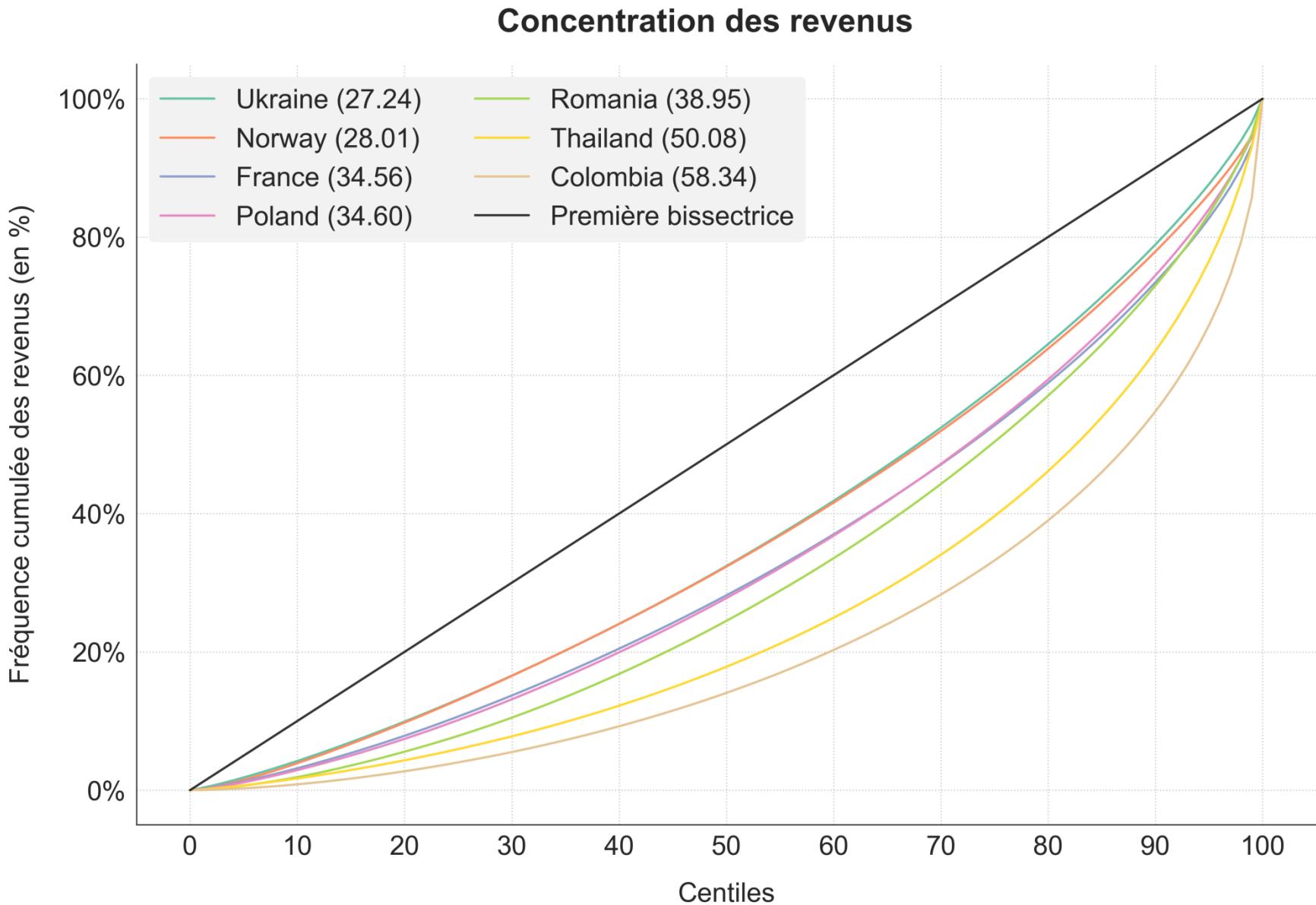
## Comparaison des classifications



# Distribution des revenus

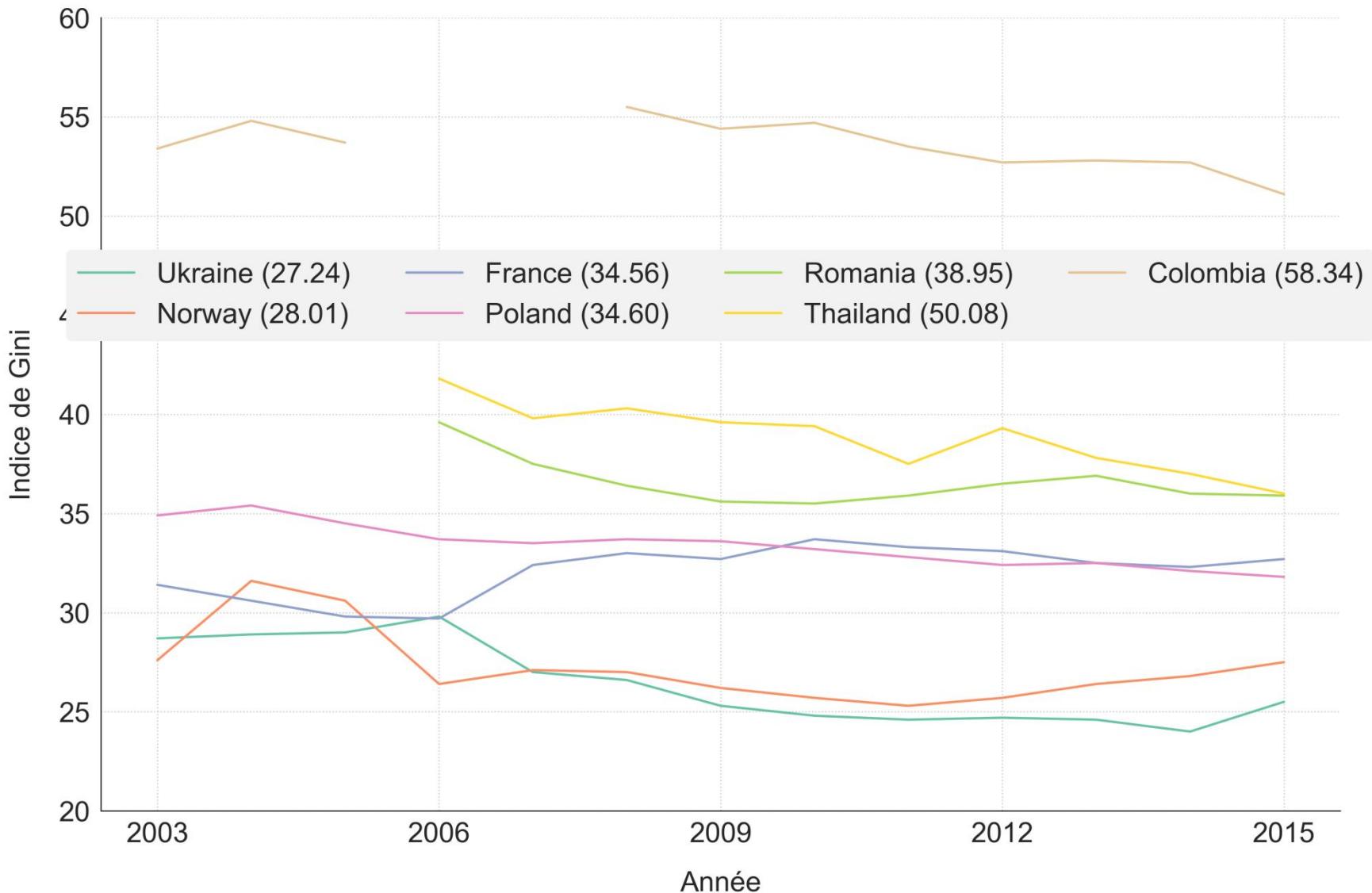


# Concentration des revenus

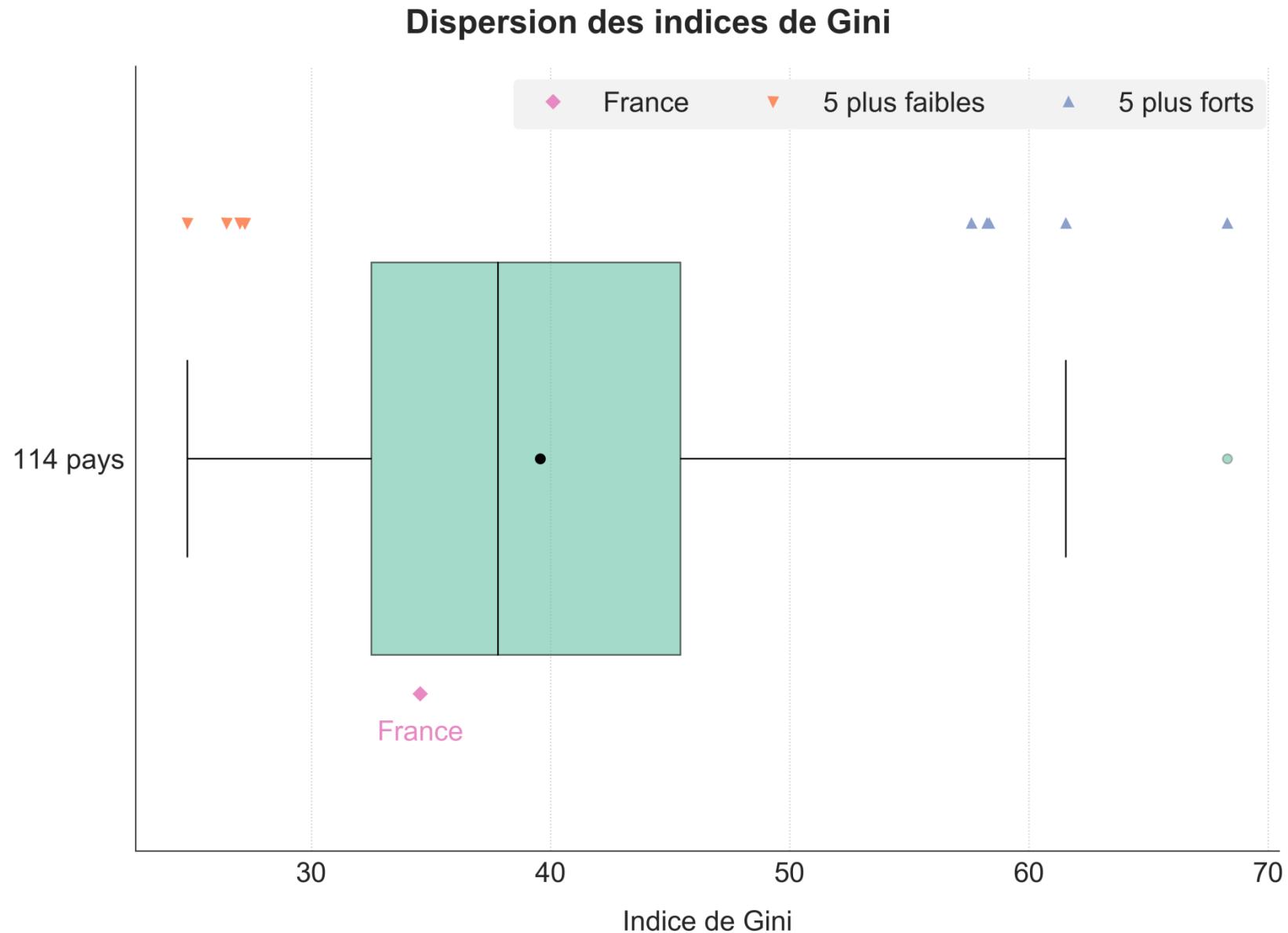


# Evolution de l'indice de Gini

## Evolution de l'indice de Gini dans le temps



# Dispersion des indices de Gini

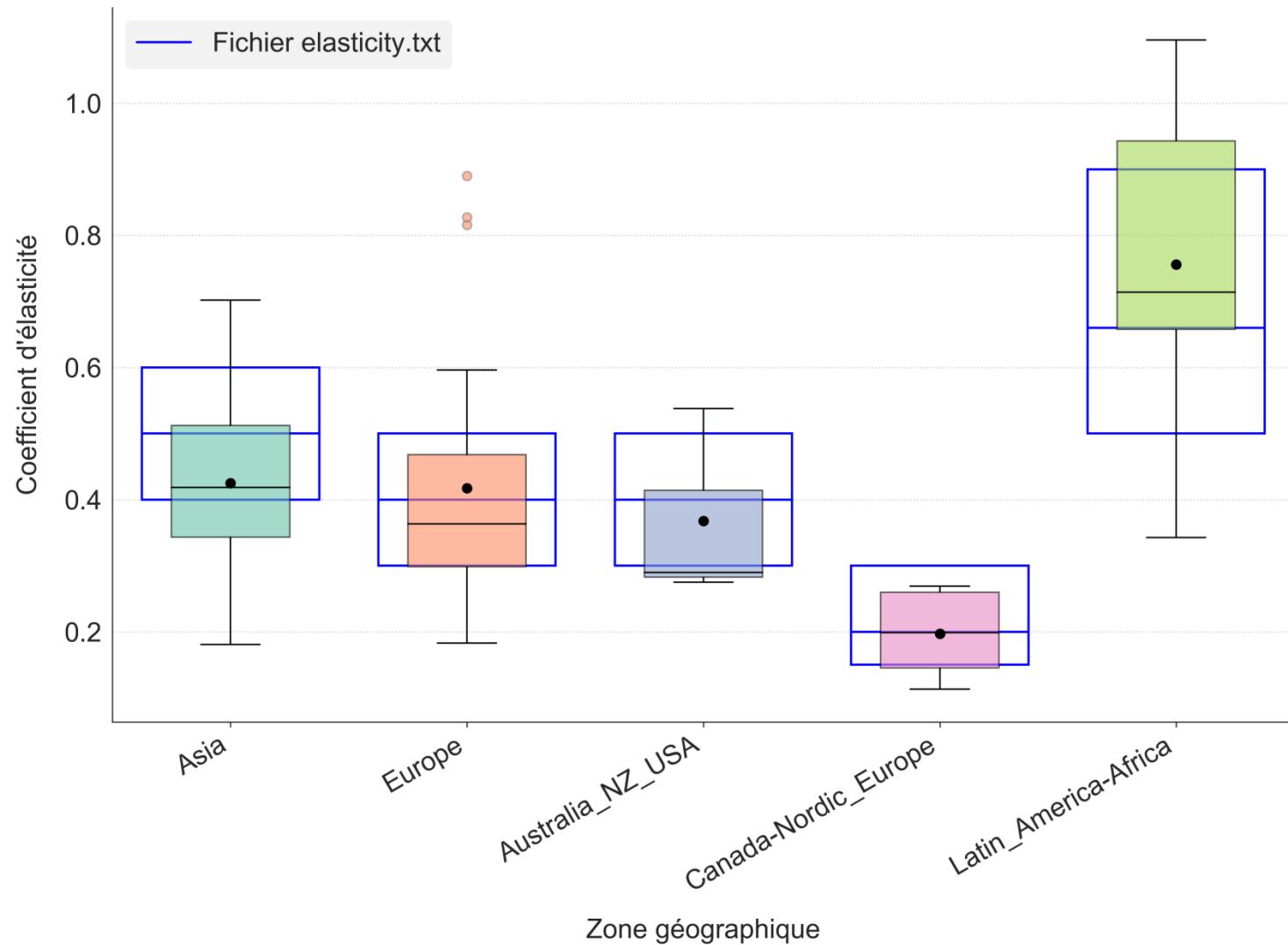


# **ÉCHANTILLONNAGE DES INDIVIDUS**

# Coefficient d'élasticité

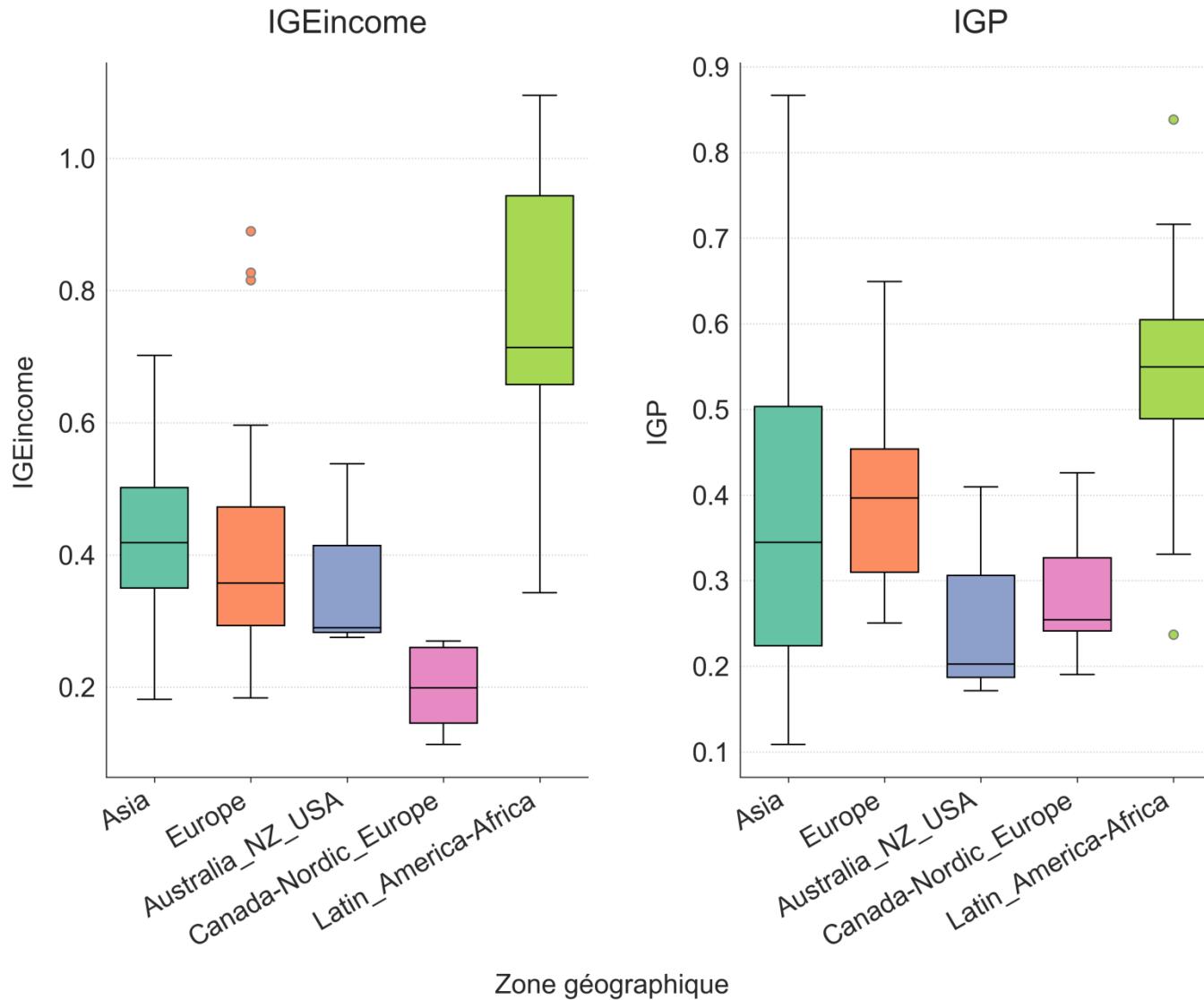
Dispersion du coefficient d'élasticité par zone géographique

Base GDIM vs elasticity.txt



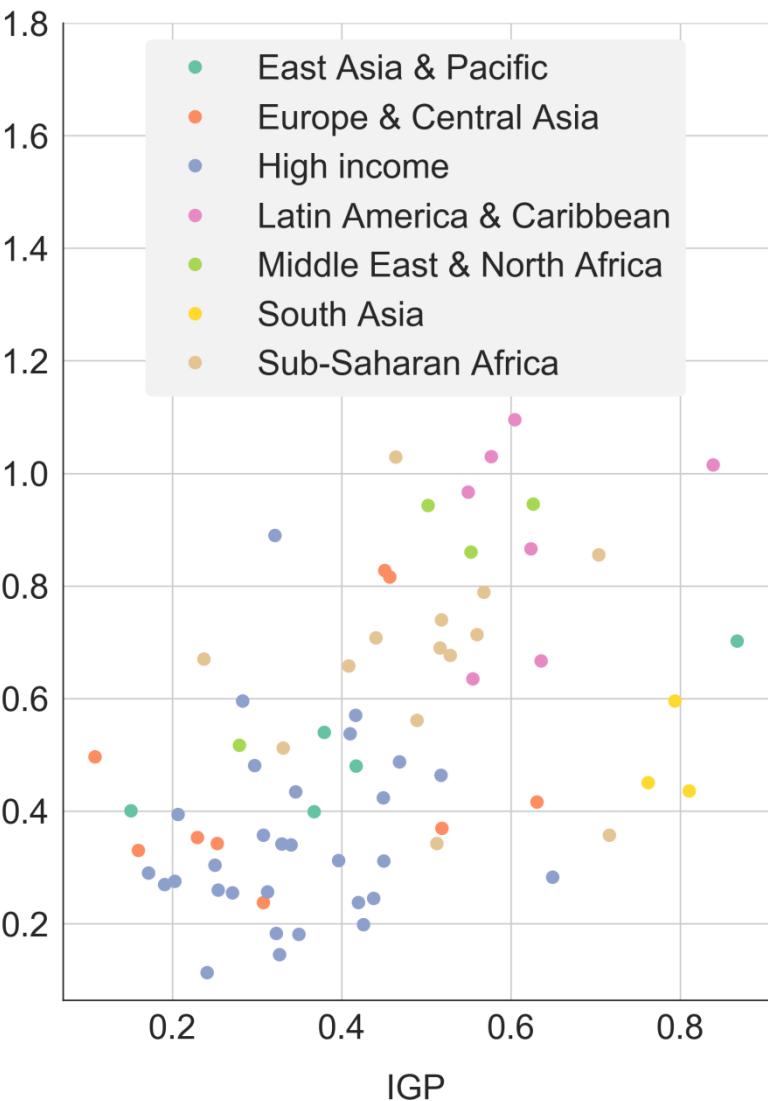
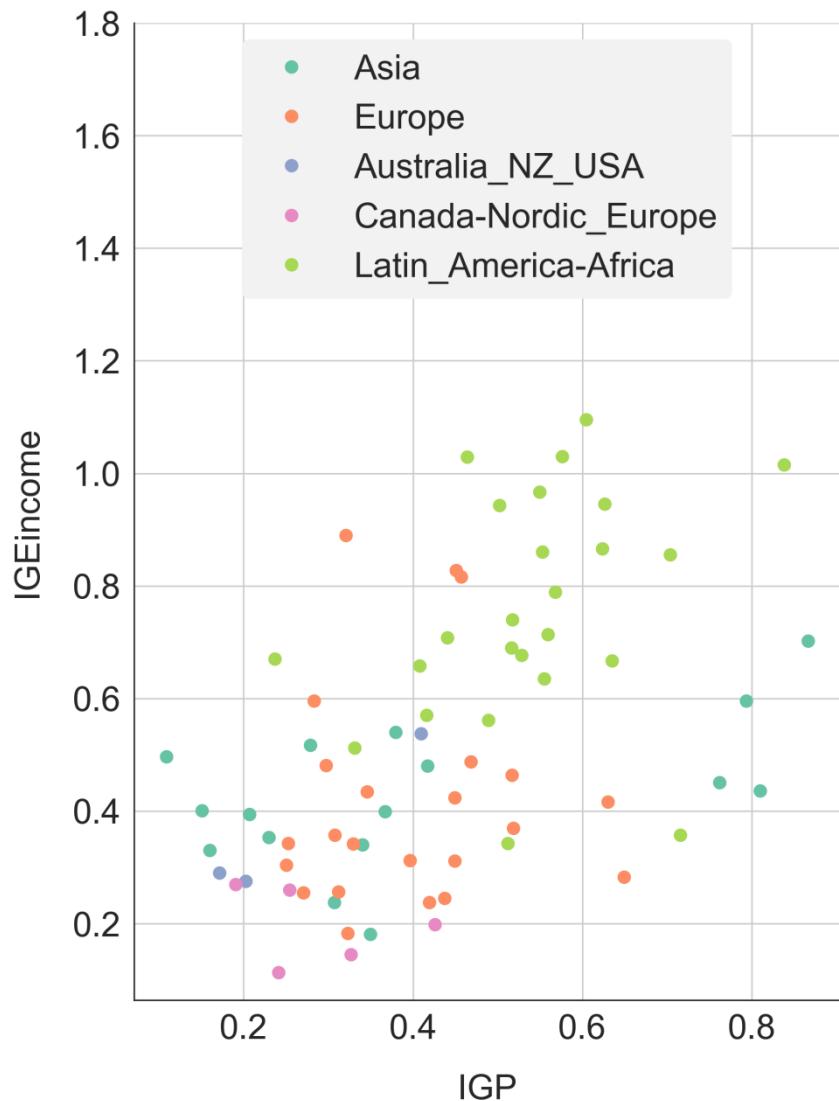
# Coefficient d'élasticité

Dispersion des variables IGEincome et IGP, par zone géographique



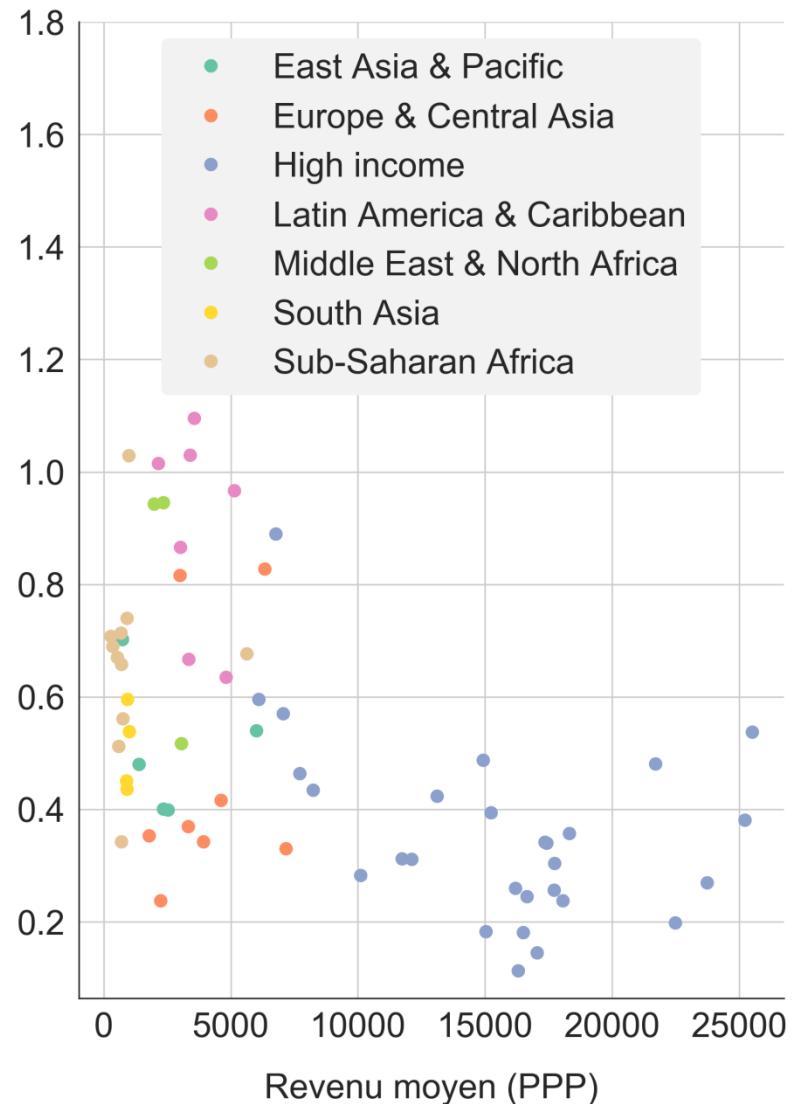
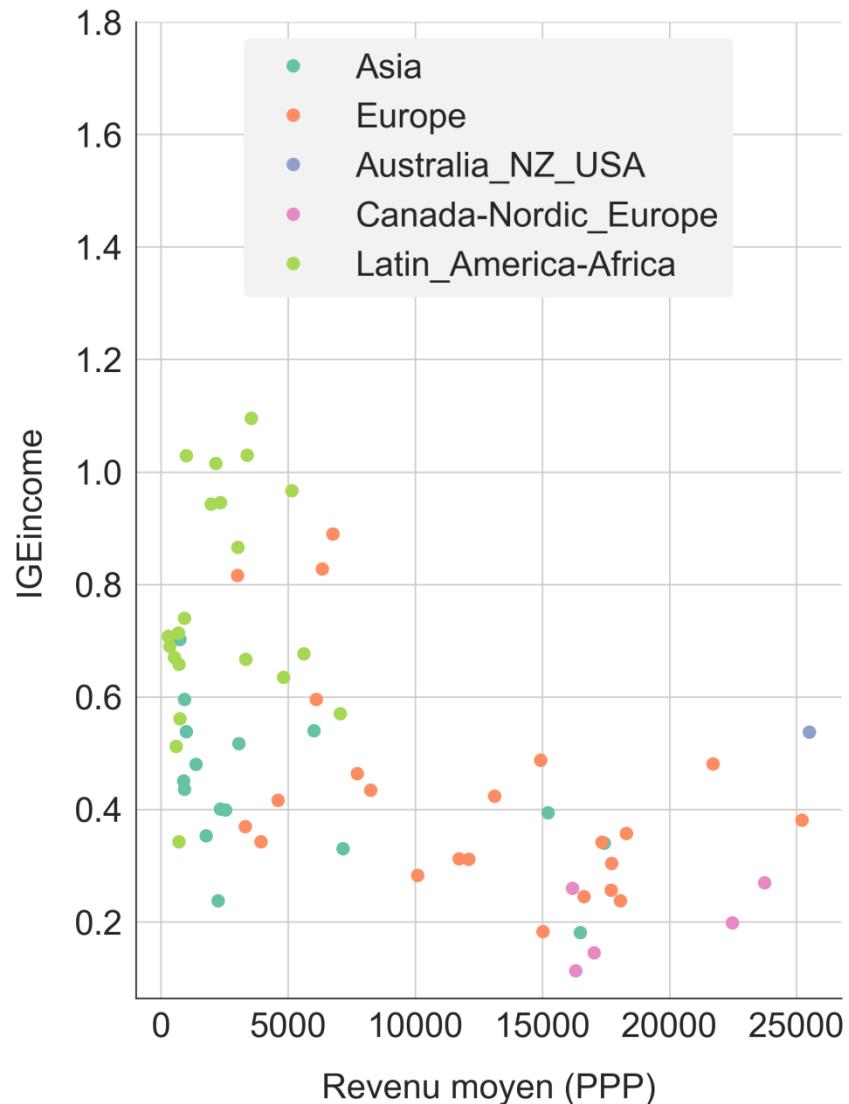
# Coefficient d'élasticité

Diagramme de dispersion d'IGEincome en fonction d'IGP



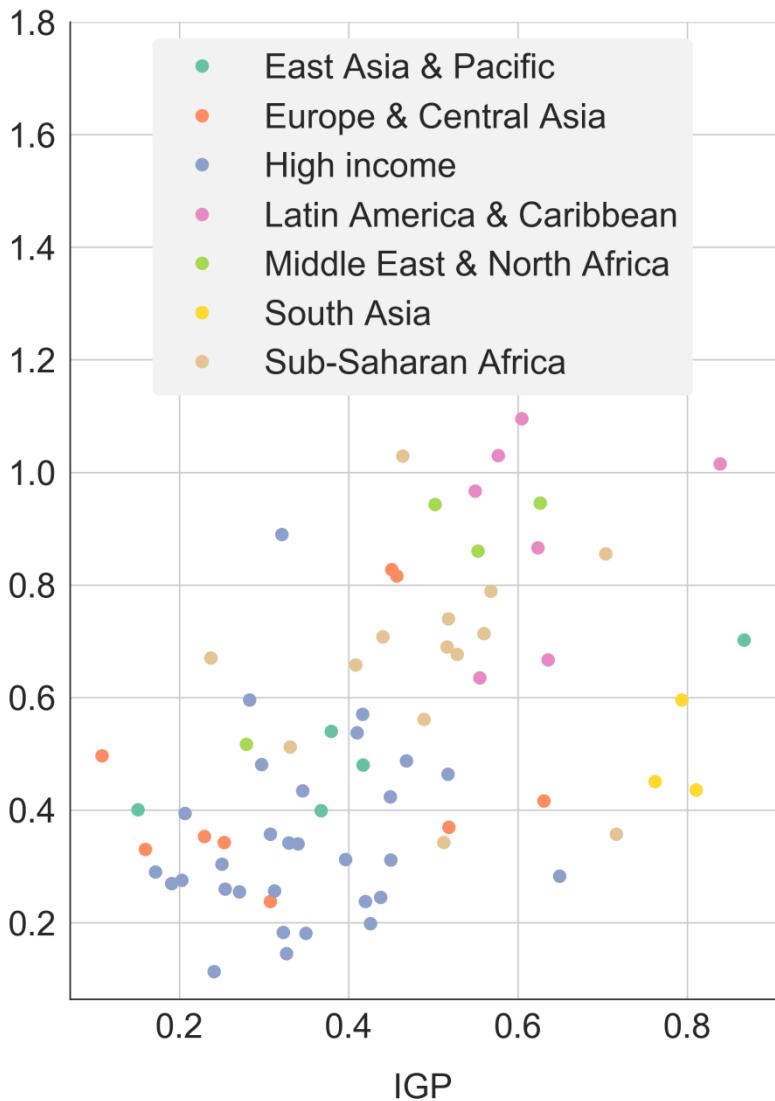
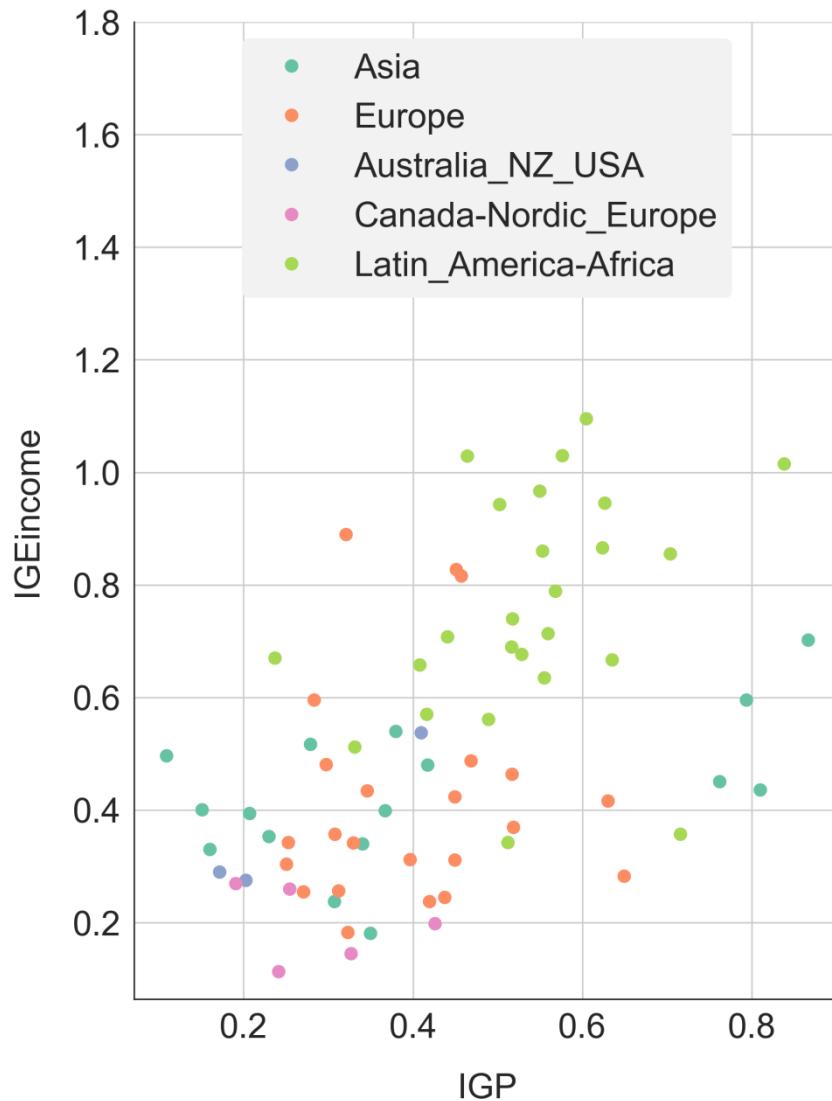
# Coefficient d'élasticité

Diagramme de dispersion d'IGEincome en fonction du revenu moyen



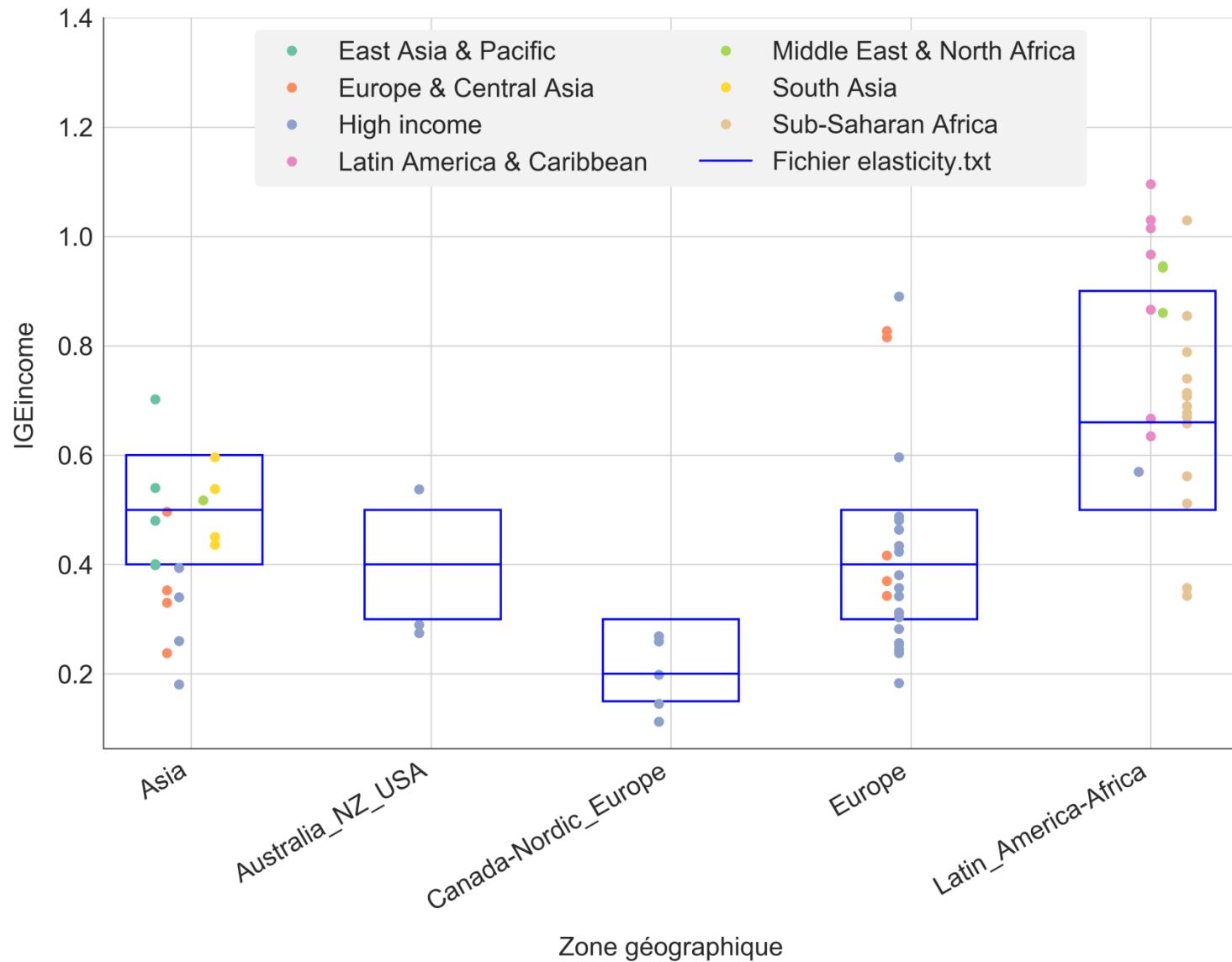
# Coefficient d'élasticité

Diagramme de dispersion d'IGEincome en fonction d'IGP



# Coefficient d'élasticité

Diagramme de dispersion d'IGEincome en fonction de la zone géographique



# Calcul des probabilités conditionnelles

Génération aléatoire

Revenu des parents  
Erreur  $\varepsilon$

Revenu des enfants  
 $\ln(\text{enfants}) = \alpha + p_j \ln(\text{parents}) + \varepsilon$



Calcul des quantiles

Classe du revenu des parents

Classe du revenu de l'enfant



Calcul des probabilités conditionnelles

A = Distribution des paires de quantiles

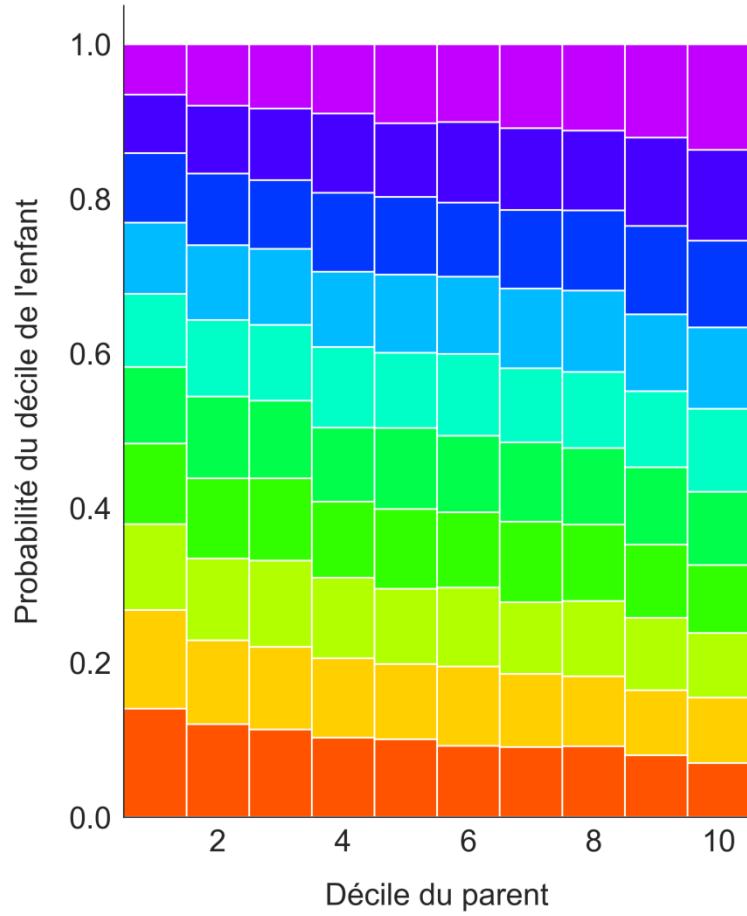
B = Distribution des quantiles du revenu des enfants

Probabilité conditionnelle = A / B

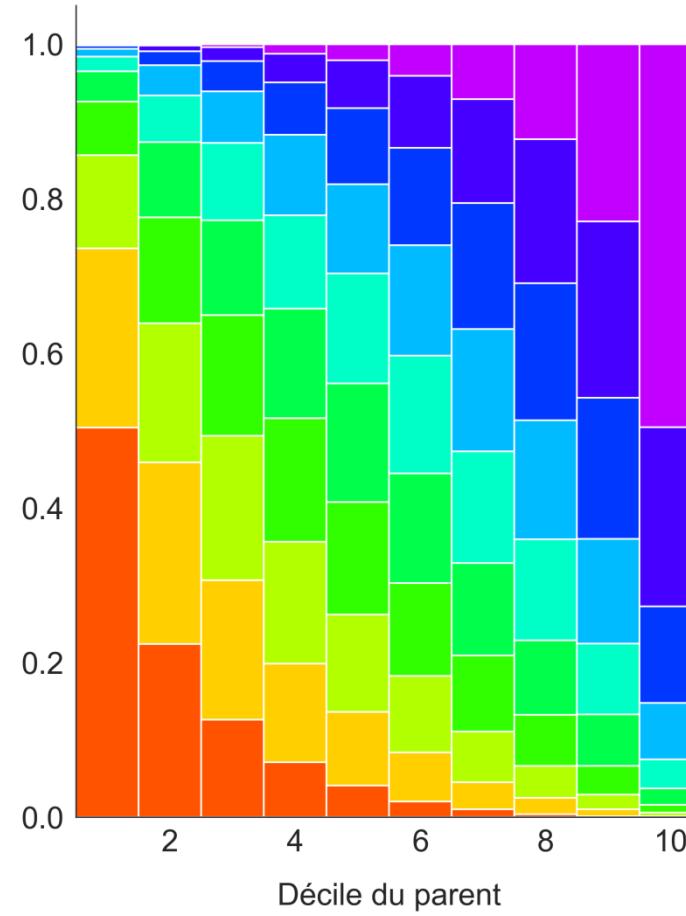
# Calcul des probabilités conditionnelles

## Vérification des probabilités conditionnelles - Valeurs extrêmes de p

Finland -  $p = 0.11$



Colombia -  $p = 1.10$

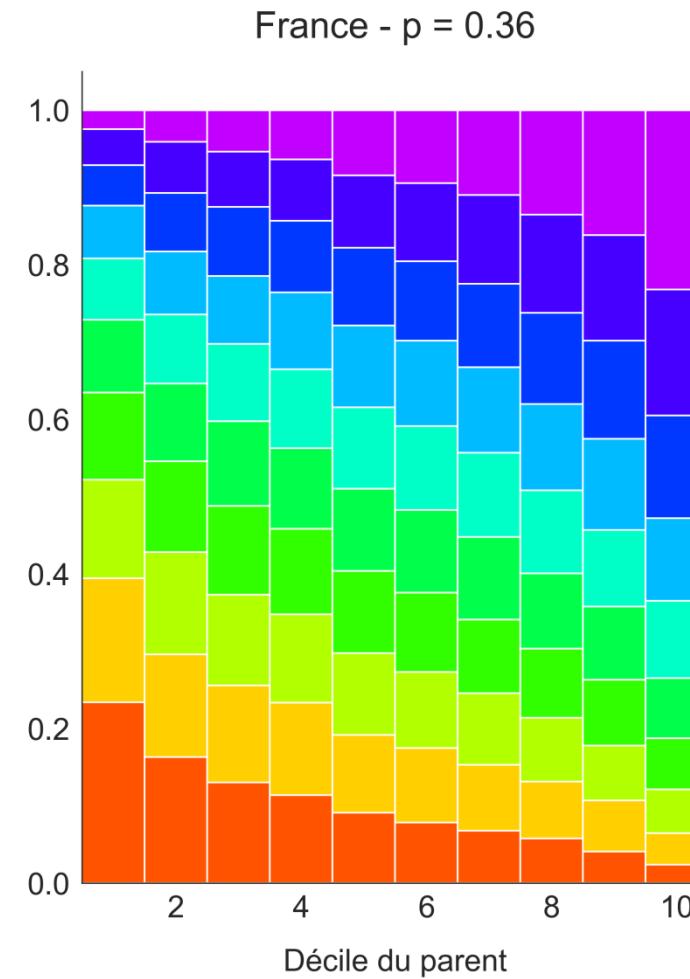
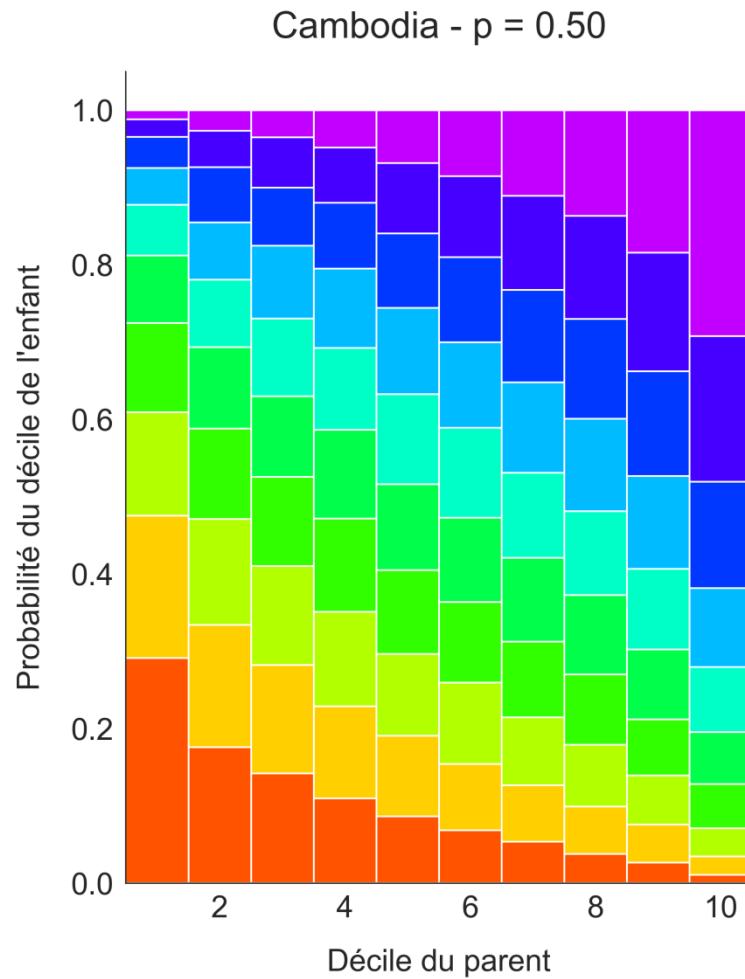


Décile de l'enfant

1 2 3 4 5 6 7 8 9 10

# Calcul des probabilités conditionnelles

Vérification des probabilités conditionnelles - Valeur médiane de  $p$  et France



Décile de l'enfant

1 2 3 4 5 6 7 8 9 10

# Génération de l'échantillon

Centiles de la World Income Distribution

500 enfants pour chaque paire (pays , centile) -> revenu de l'enfant



Calcul du nombre de parents

Probabilité conditionnelle -> centile du revenu pour n parents



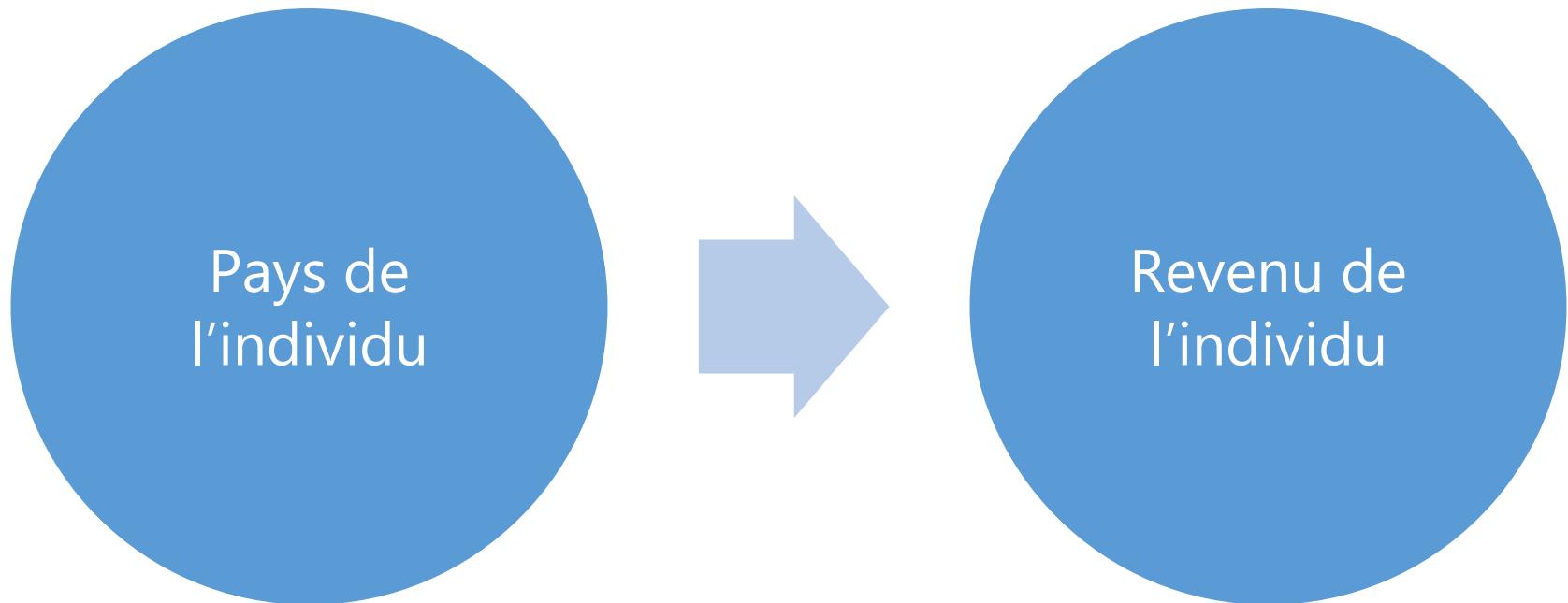
Ajout des autres informations

Revenu moyen du pays

Indice de Gini du pays

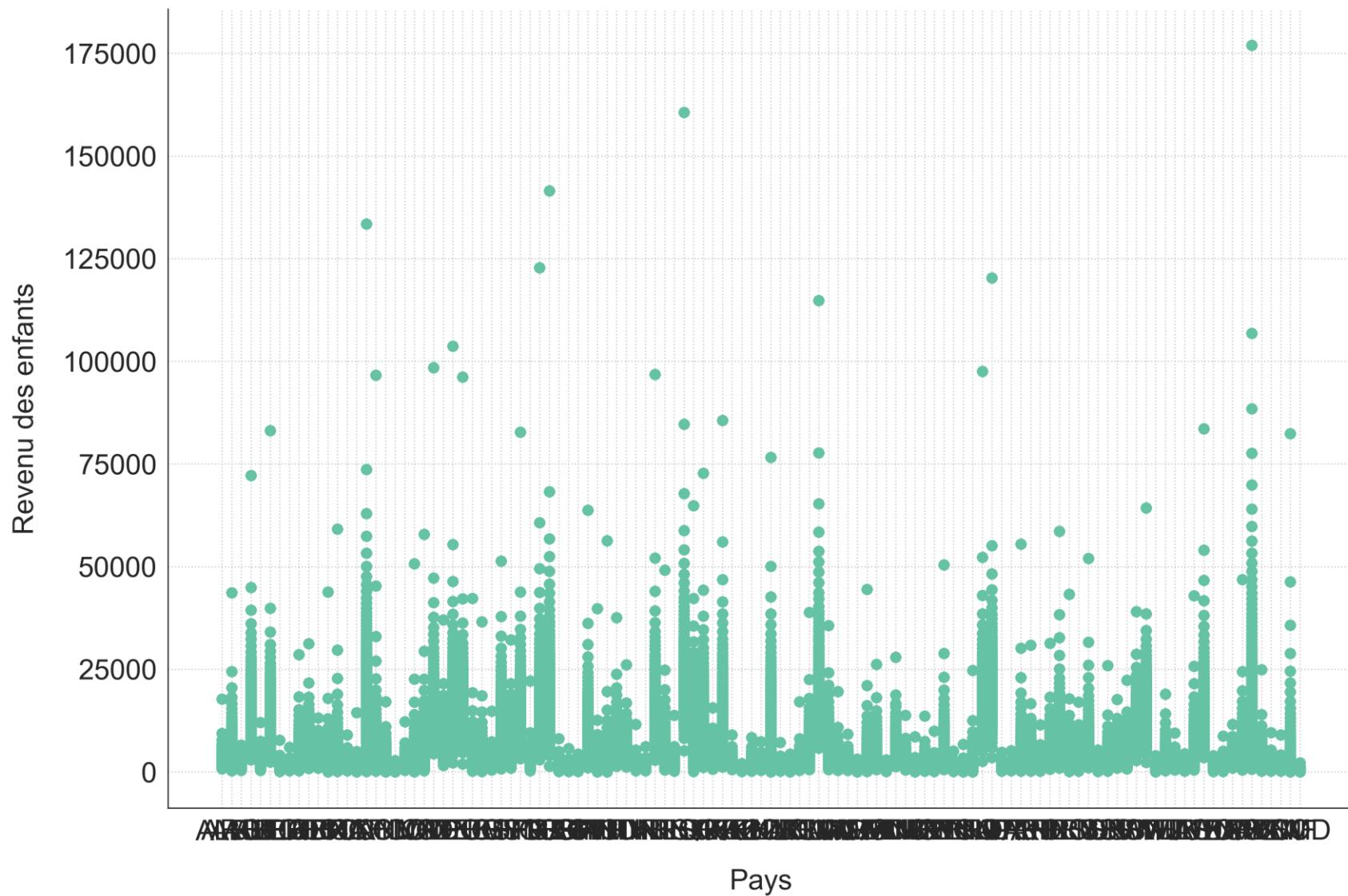
# MODÉLISATION DU REVENU DES INDIVIDUS

# Premier modèle



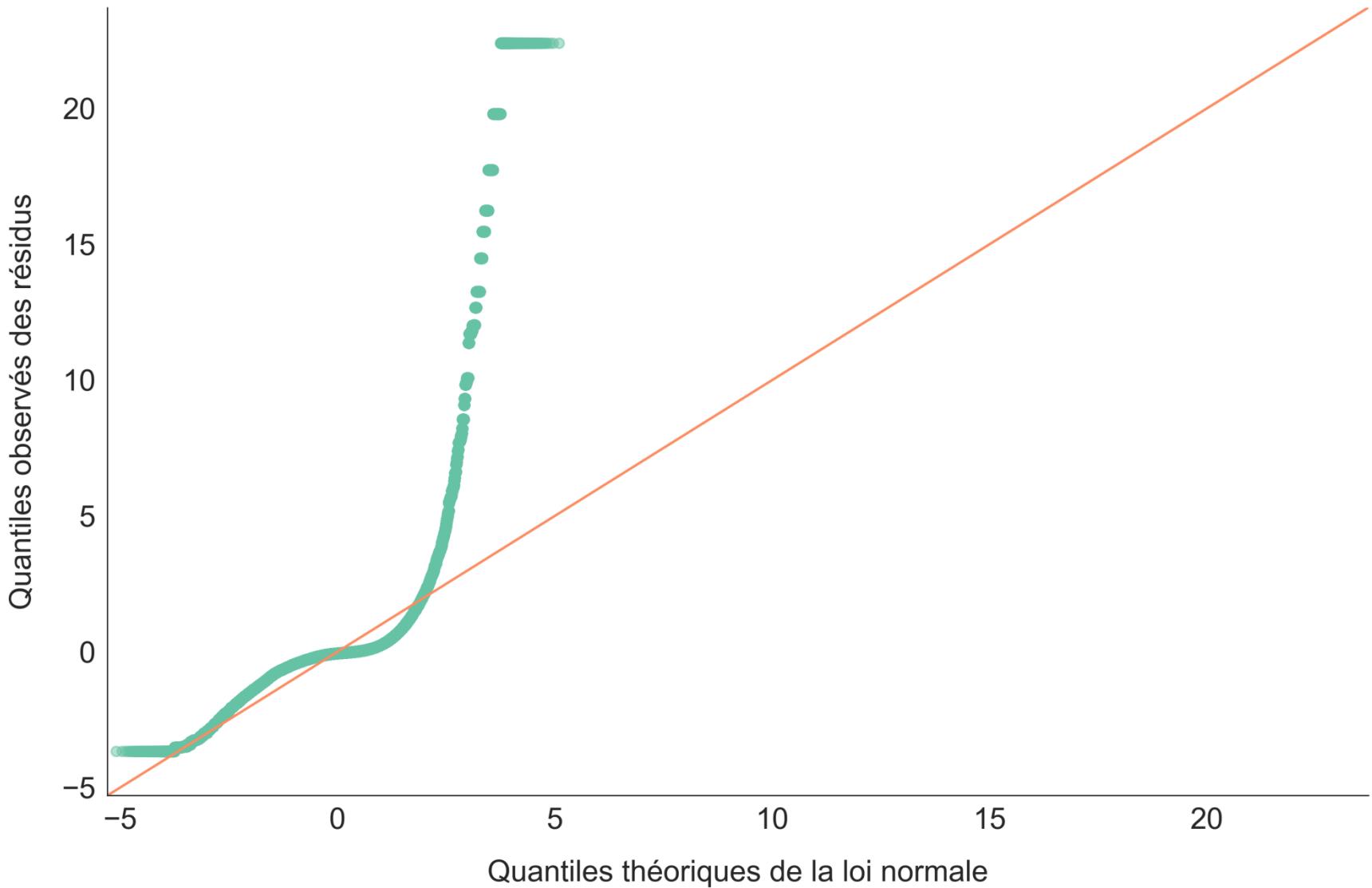
# Premier modèle

Diagramme de dispersion du revenu des enfants en fonction du pays



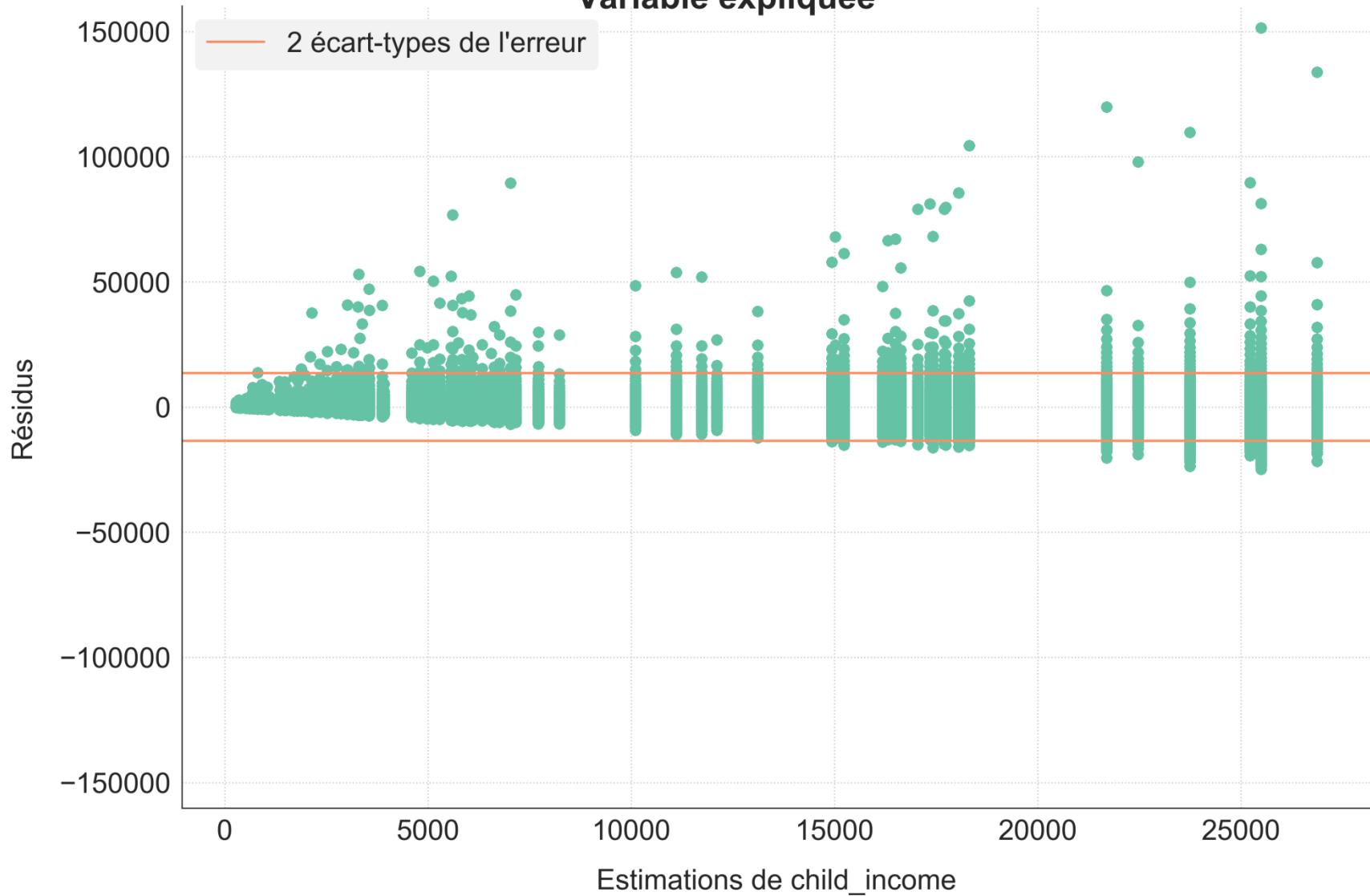
# Premier modèle

Droite de Henry : vérification de la normalité des résidus

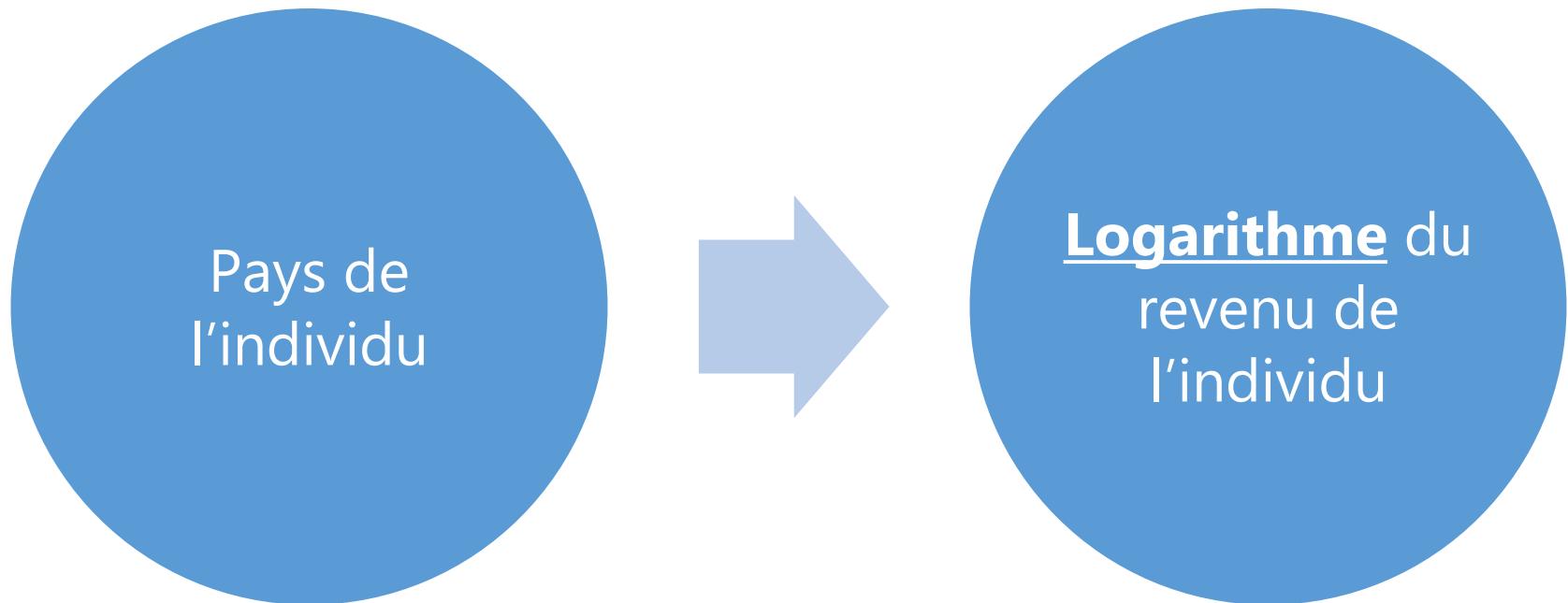


# Premier modèle

Vérification de la linéarité de la relation et de l'homoscédasticité des résidus  
Variable expliquée

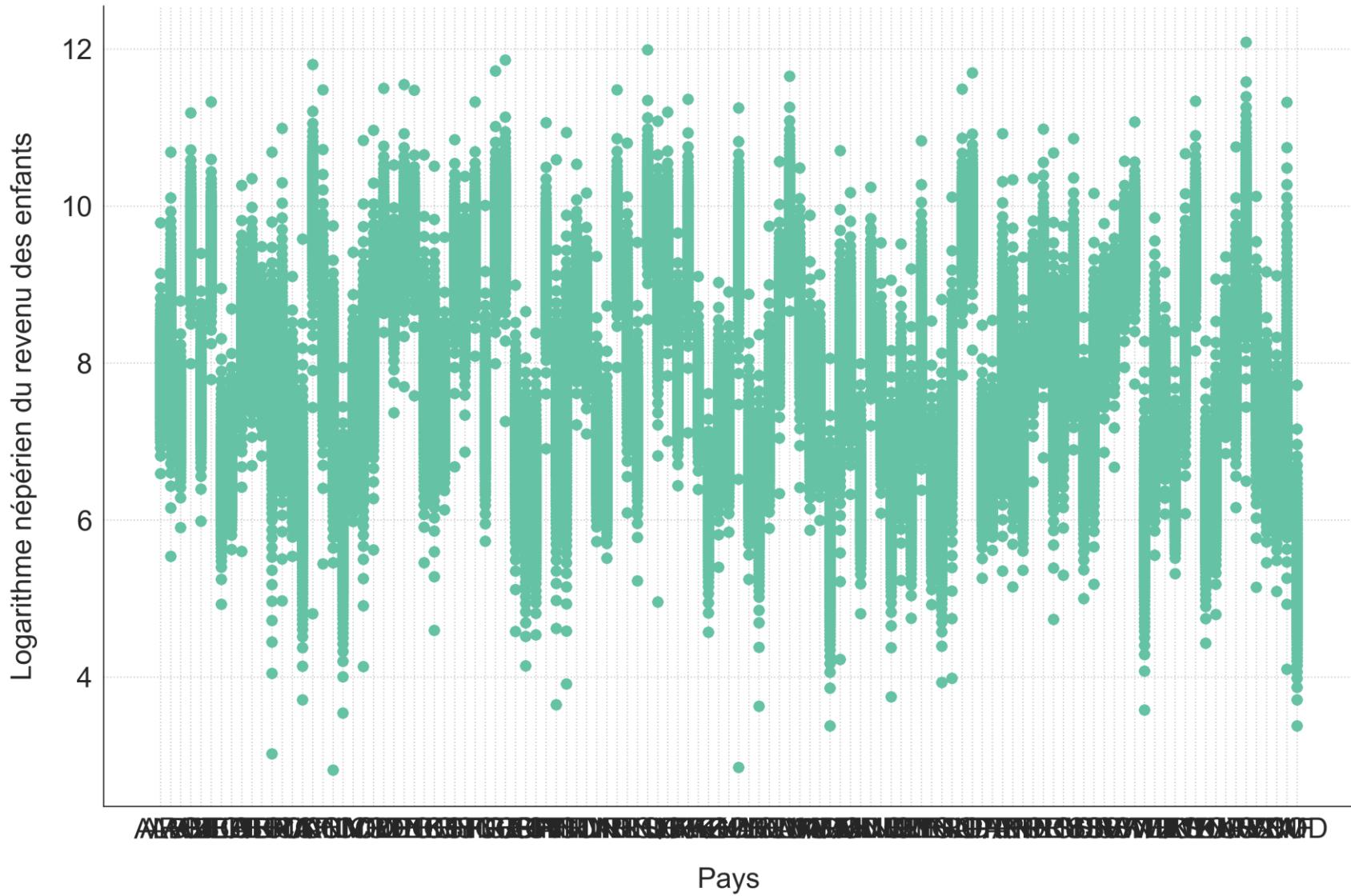


# Deuxième modèle



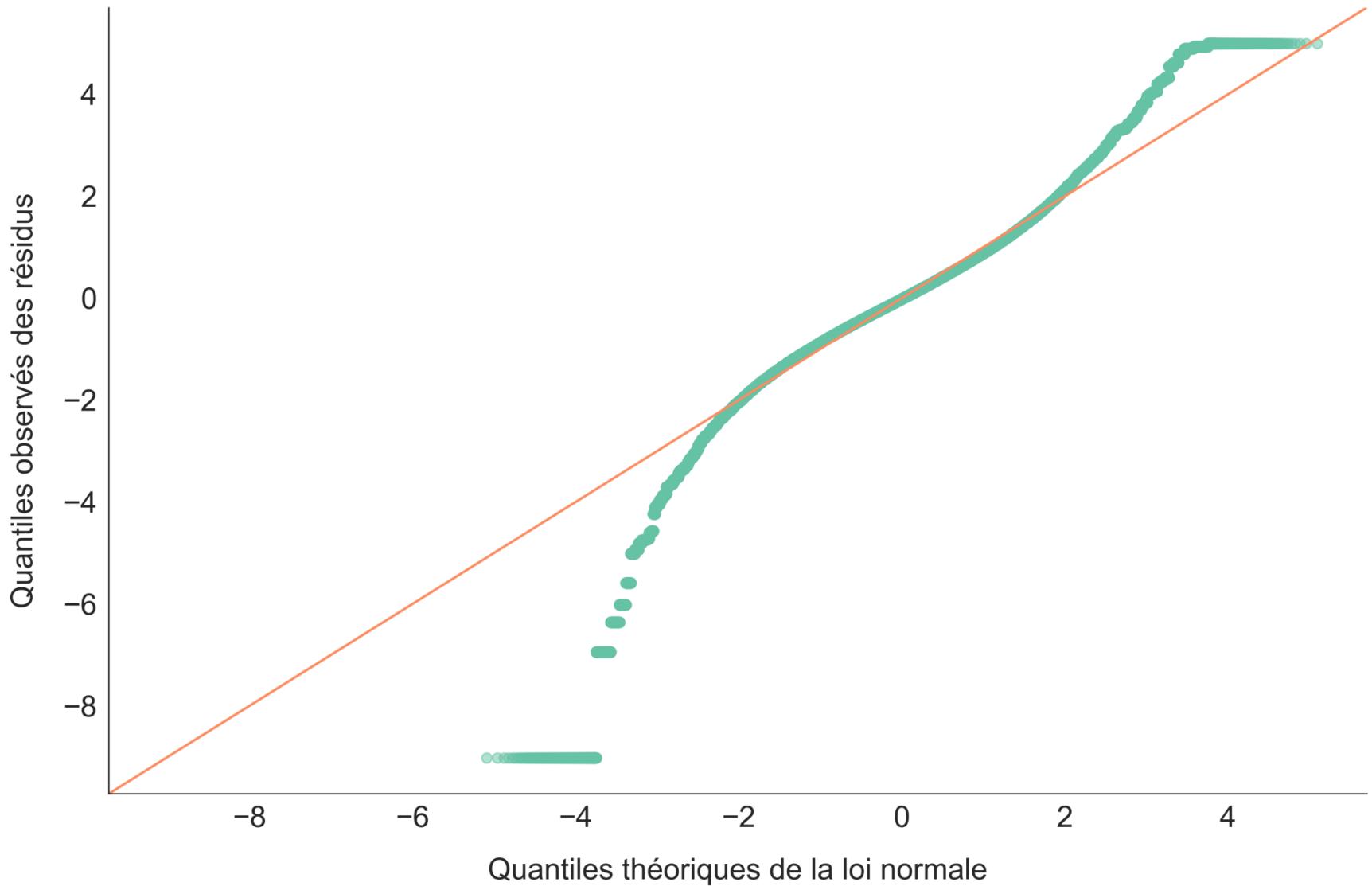
# Deuxième modèle

Diagramme de dispersion du logarithme du revenu des enfants en fonction du pays



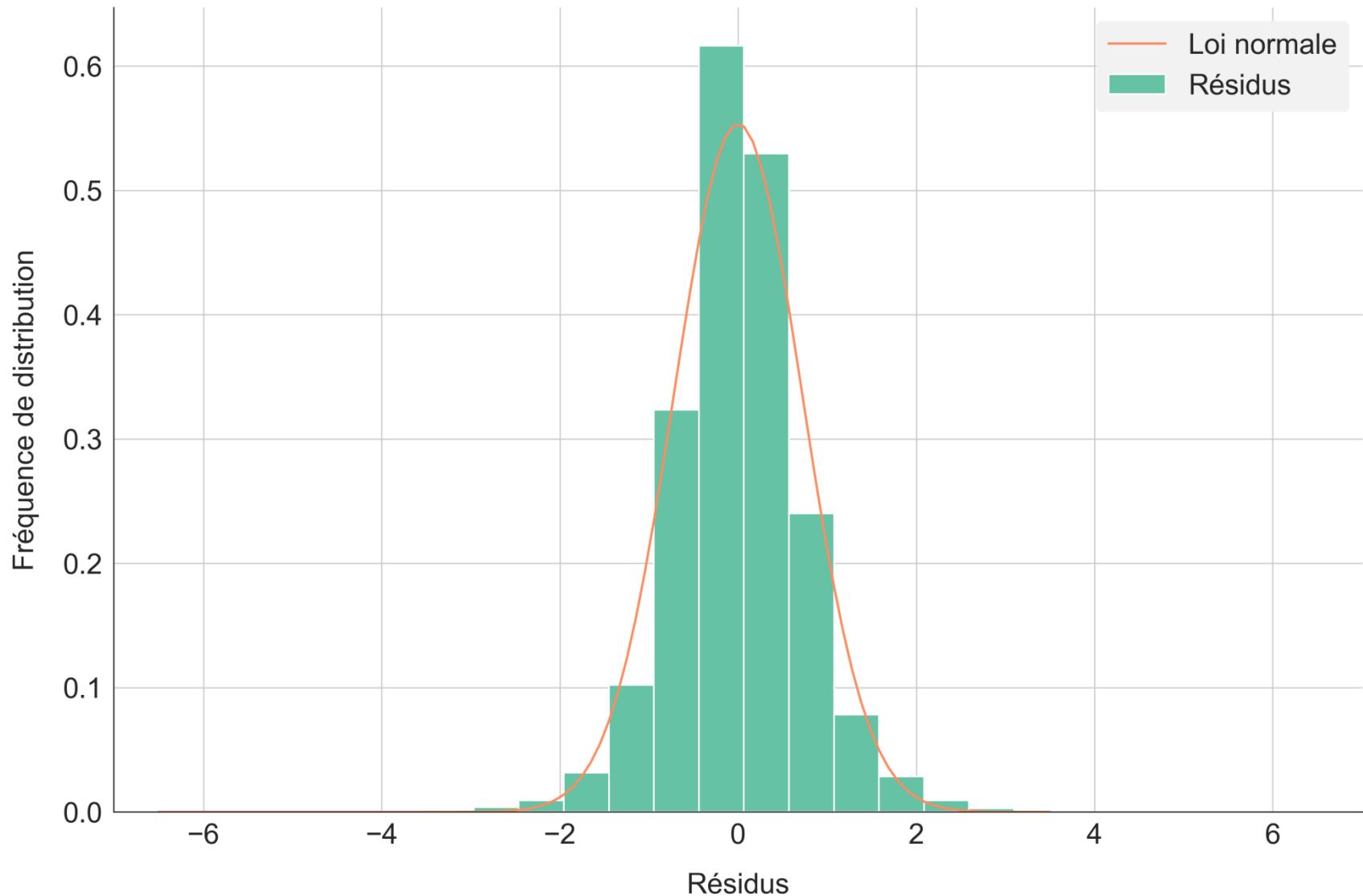
# Deuxième modèle

Droite de Henry : vérification de la normalité des résidus



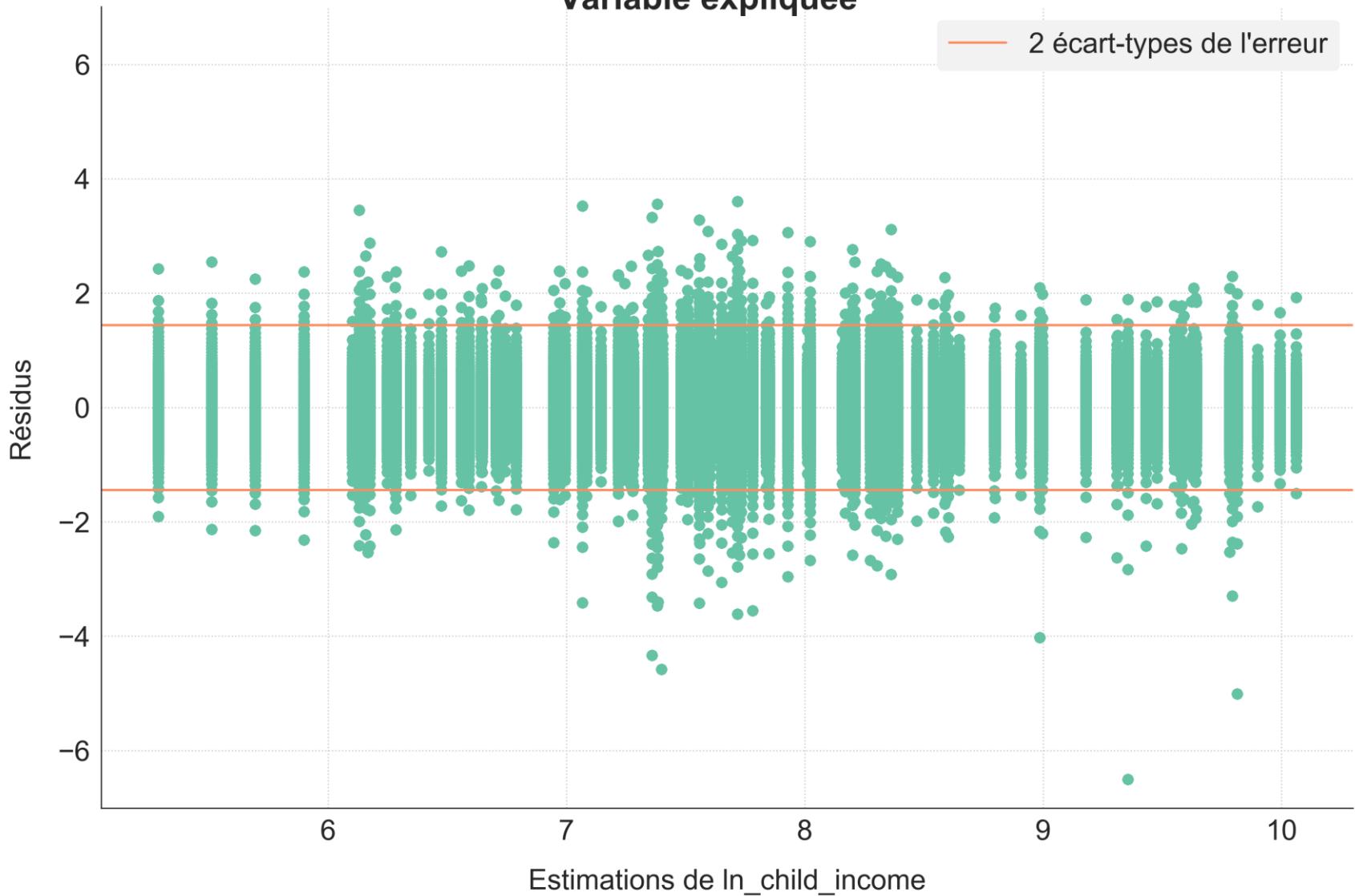
# Deuxième modèle

Histogramme de la distribution des résidus



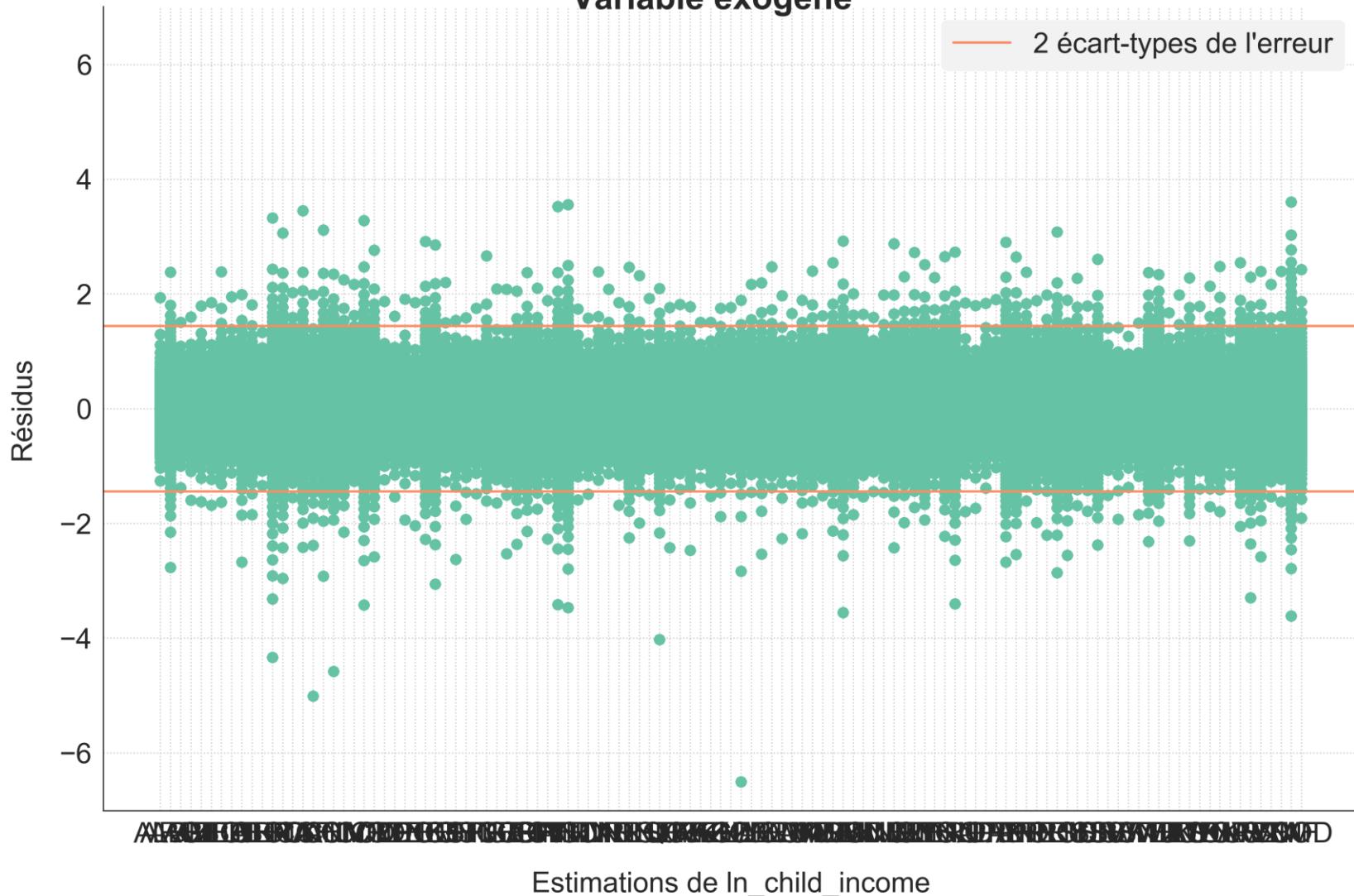
# Deuxième modèle

Vérification de la linéarité de la relation et de l'homoscédasticité des résidus  
Variable expliquée



# Deuxième modèle

Vérification de la linéarité de la relation, de l'homoscédasticité et de l'indépendance des résidus  
Variable exogène



# Deuxième modèle

Statistique	Signification	Interprétation
$R^2 = 0,729$ $R^2 \text{ ajusté} = 0,729$	Qualité explicative du modèle	Le modèle explique 72,9% de la variance totale
Proba(F-stat) = 0,00	Test de signification globale du modèle	Modèle significatif
AIC = 12 310 000 BIC = 12 320 000	Qualité de prédiction du modèle	-

ANOVA		Somme des carrés	Degrés de liberté	Proba(>F)	Omega <sup>2</sup>
Expliquée	pays	7 881 920	112	0,000	0.728999
	Résiduelle	2 929 765	5 642 307		
	Totale	10 811 685	5 642 419		

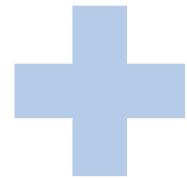
# Deuxième modèle

Estimations par rapport aux valeurs réelles



# Troisième modèle

Logarithme du  
revenu moyen du  
pays

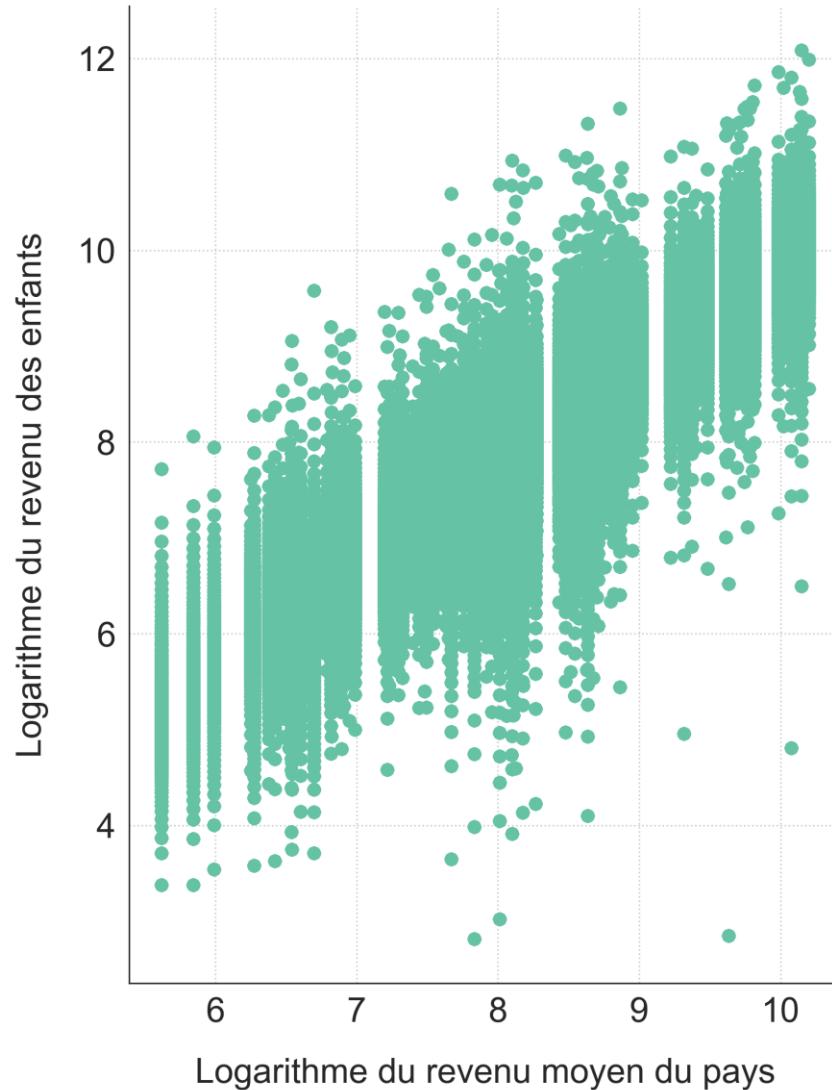


Logarithme du  
revenu de  
l'individu

Indice de Gini du  
pays

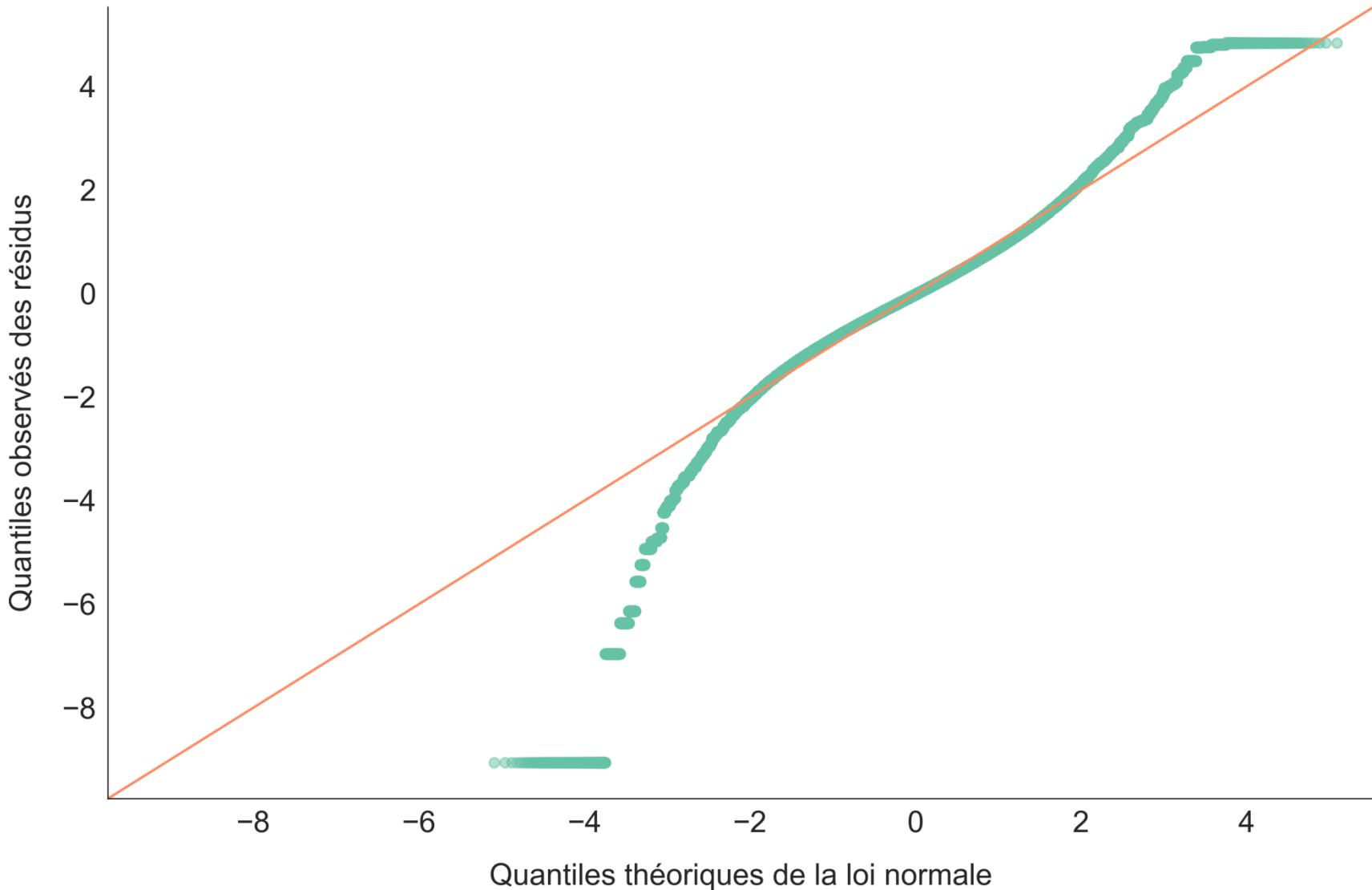
# Troisième modèle

Diagramme de dispersion du revenu des enfants en fonction du pays



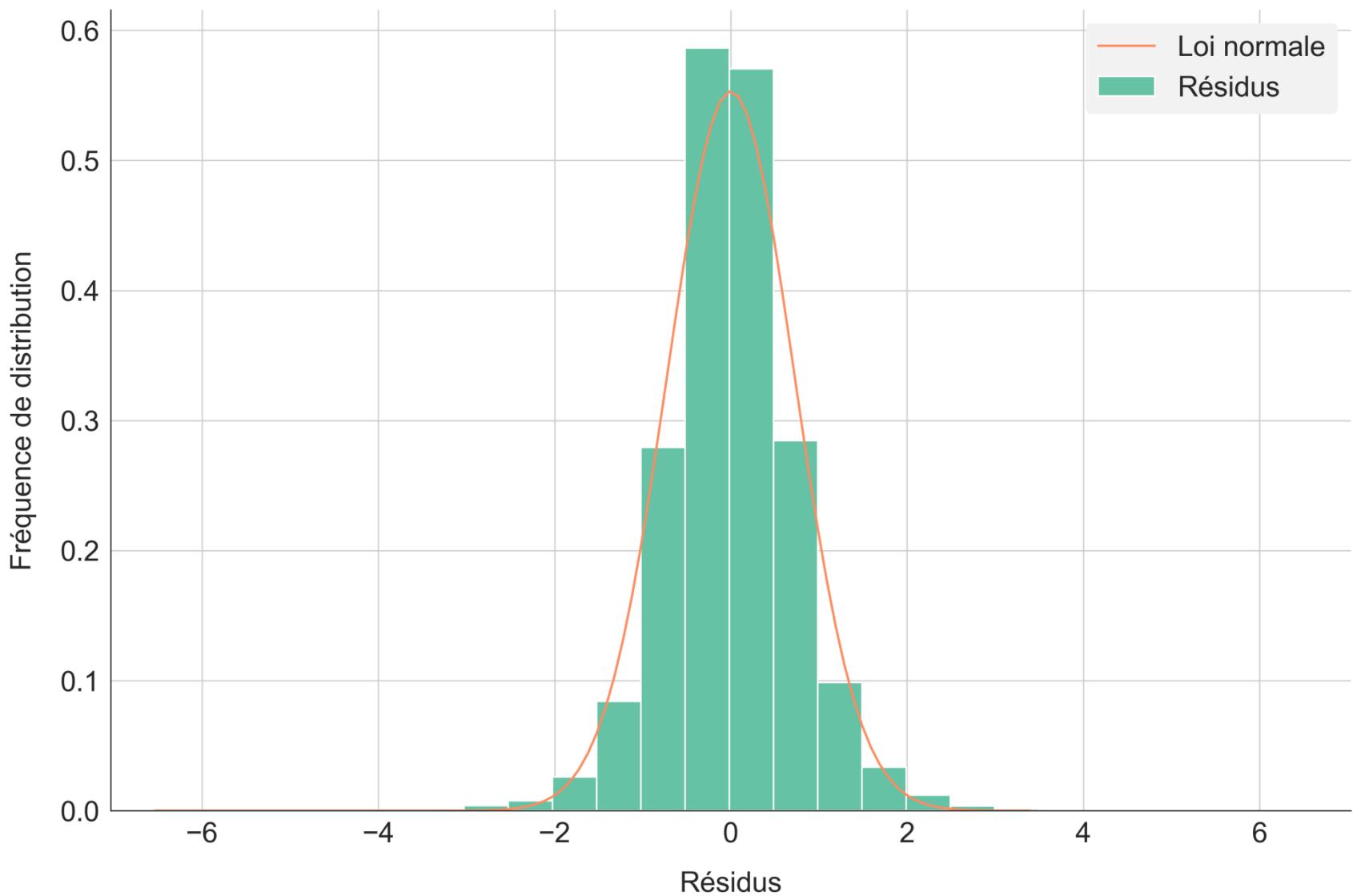
# Troisième modèle

Droite de Henry : vérification de la normalité des résidus



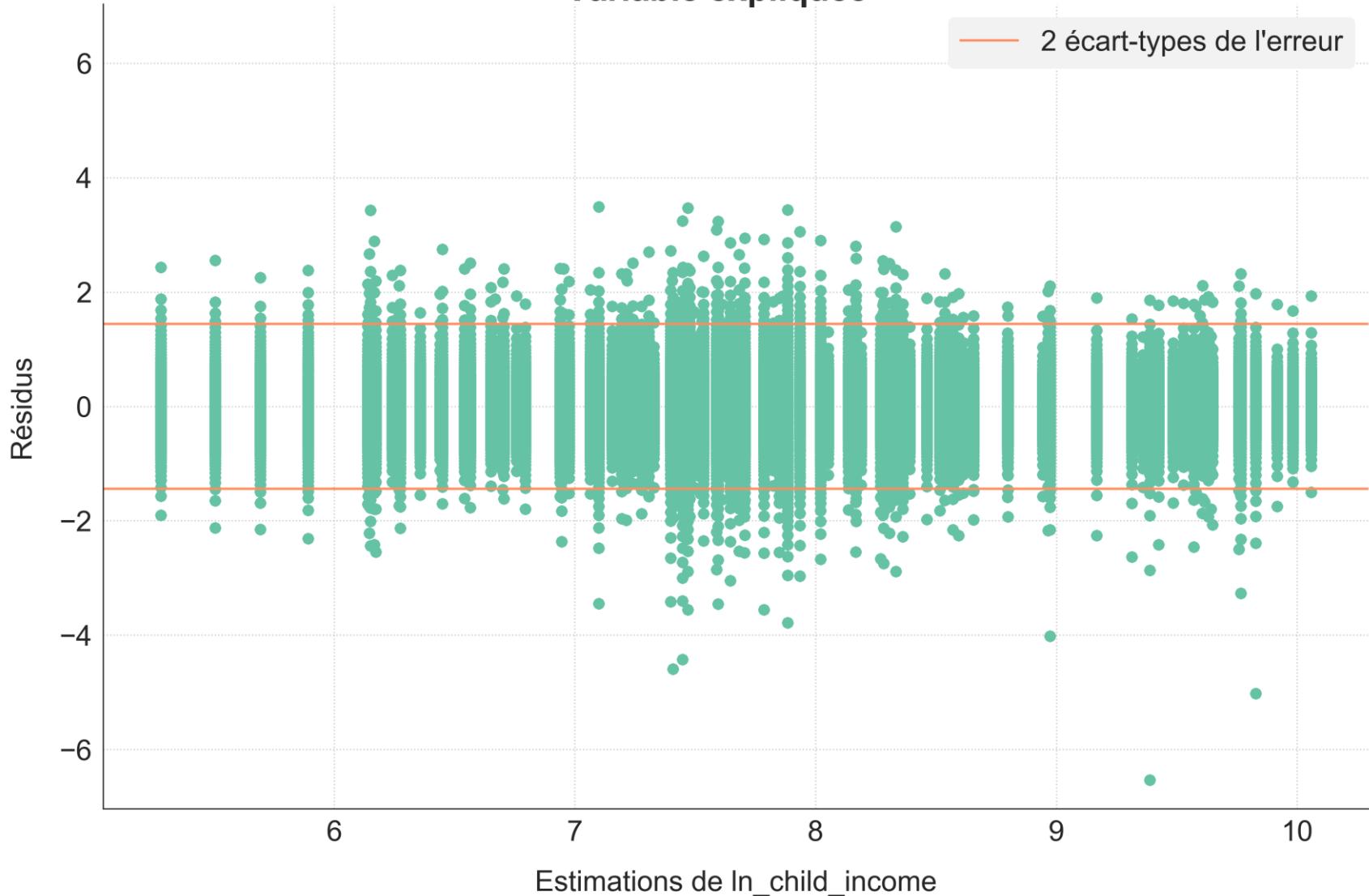
# Troisième modèle

Histogramme de la distribution des résidus



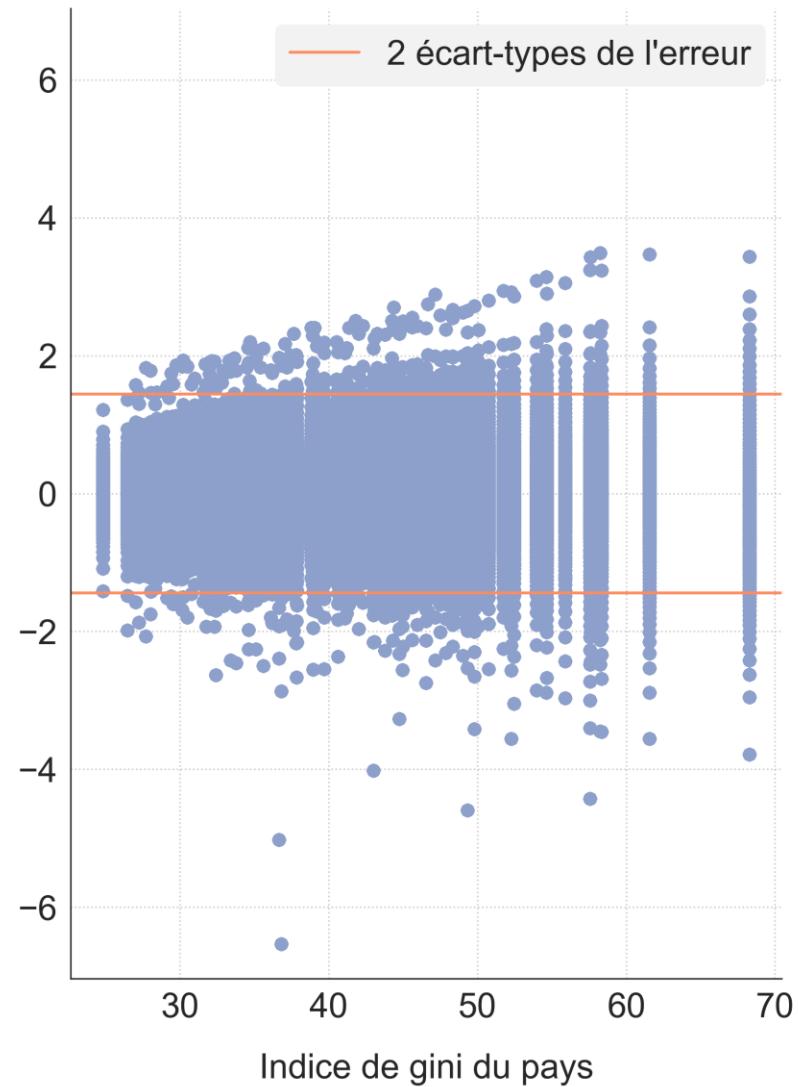
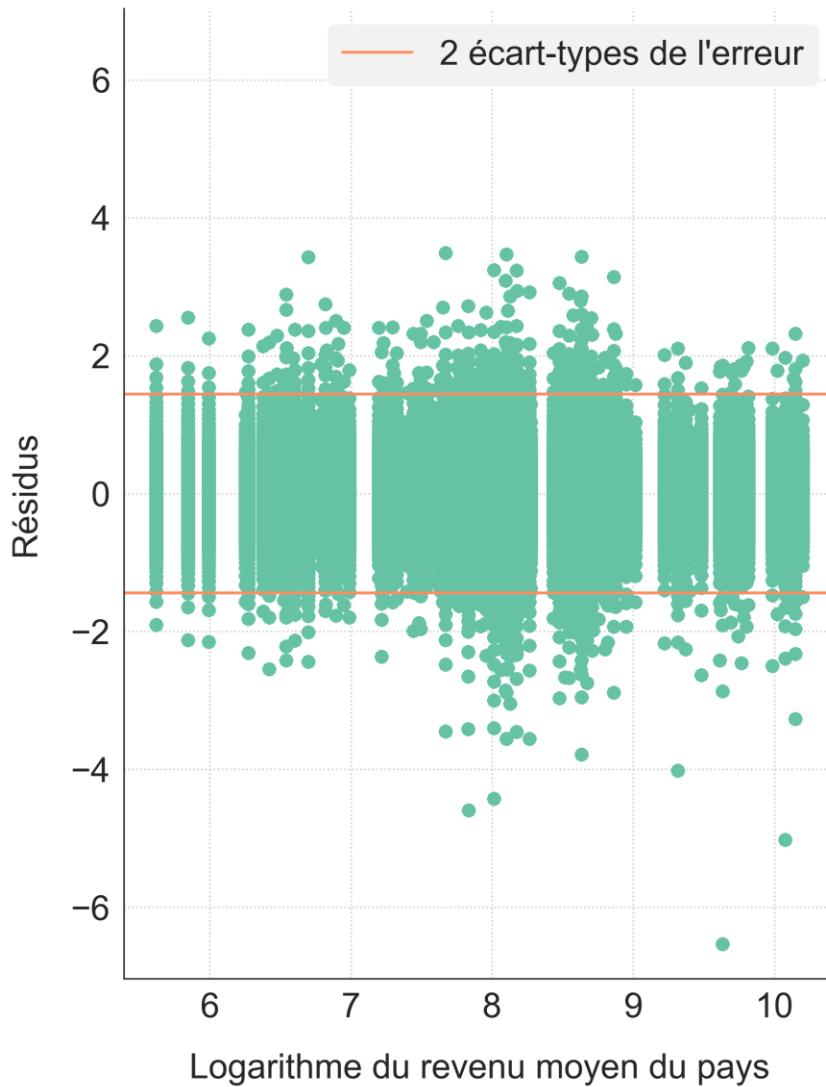
# Troisième modèle

Vérification de la linéarité de la relation et de l'homoscédasticité des résidus  
Variable expliquée



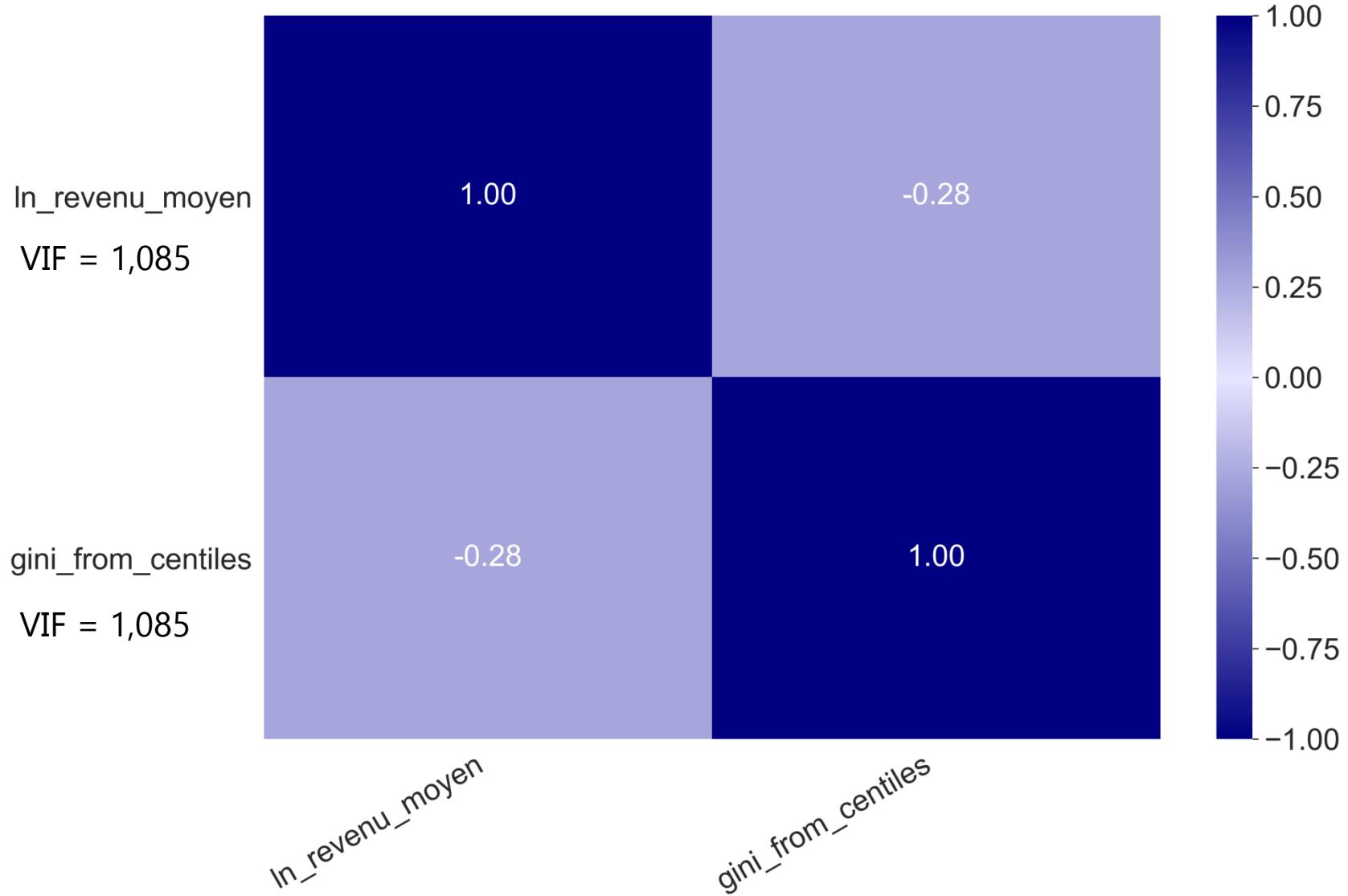
# Troisième modèle

Vérification de la linéarité de la relation, de l'homoscédasticité et de l'indépendance des résidus  
Variables exogènes



# Troisième modèle

Heatmap de la matrice de corrélation des variables explicatives



# Troisième modèle

Statistique	Signification	Interprétation
$R^2 = 0,729$ $R^2 \text{ ajusté} = 0,729$	Qualité explicative du modèle	Le modèle explique 72,9% de la variance totale
Proba(F-stat) = 0,00	Test de signification globale du modèle	Modèle significatif
AIC = 12 320 000 BIC = 12 320 000	Qualité de prédiction du modèle	Similaire au deuxième modèle

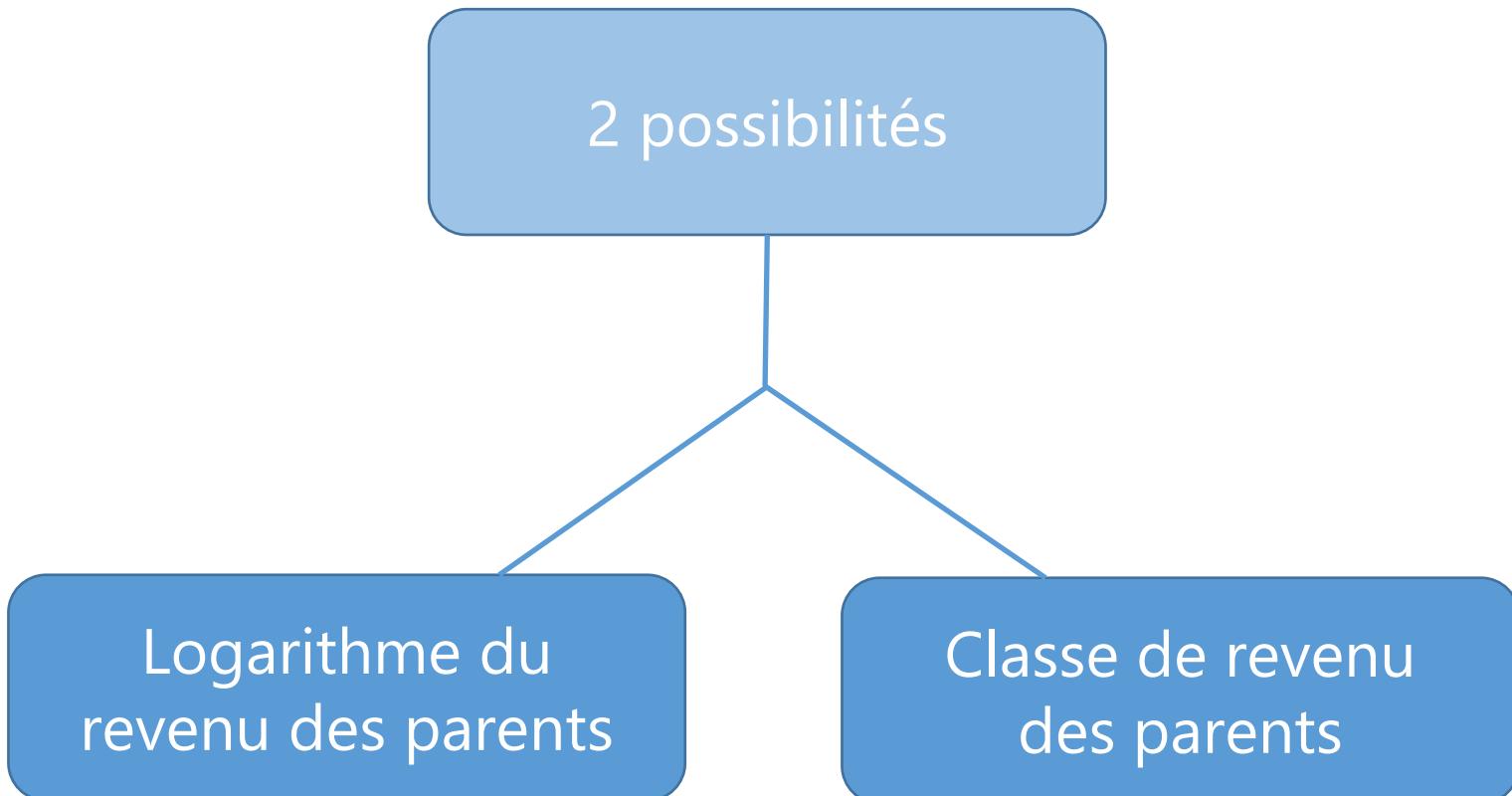
ANOVA		Somme des carrés	Degrés de liberté	Proba(>F)	Omega <sup>2</sup>
Expliquée	ln(revenu moyen)	6 662 550	1	0,000	0,686
	Indice de Gini	113 513	1	0,000	0,011
Résiduelle		2 934 565	5 642 417		
Totale		9 710 628	5 642 419		

# Troisième modèle

Estimations par rapport aux valeurs réelles

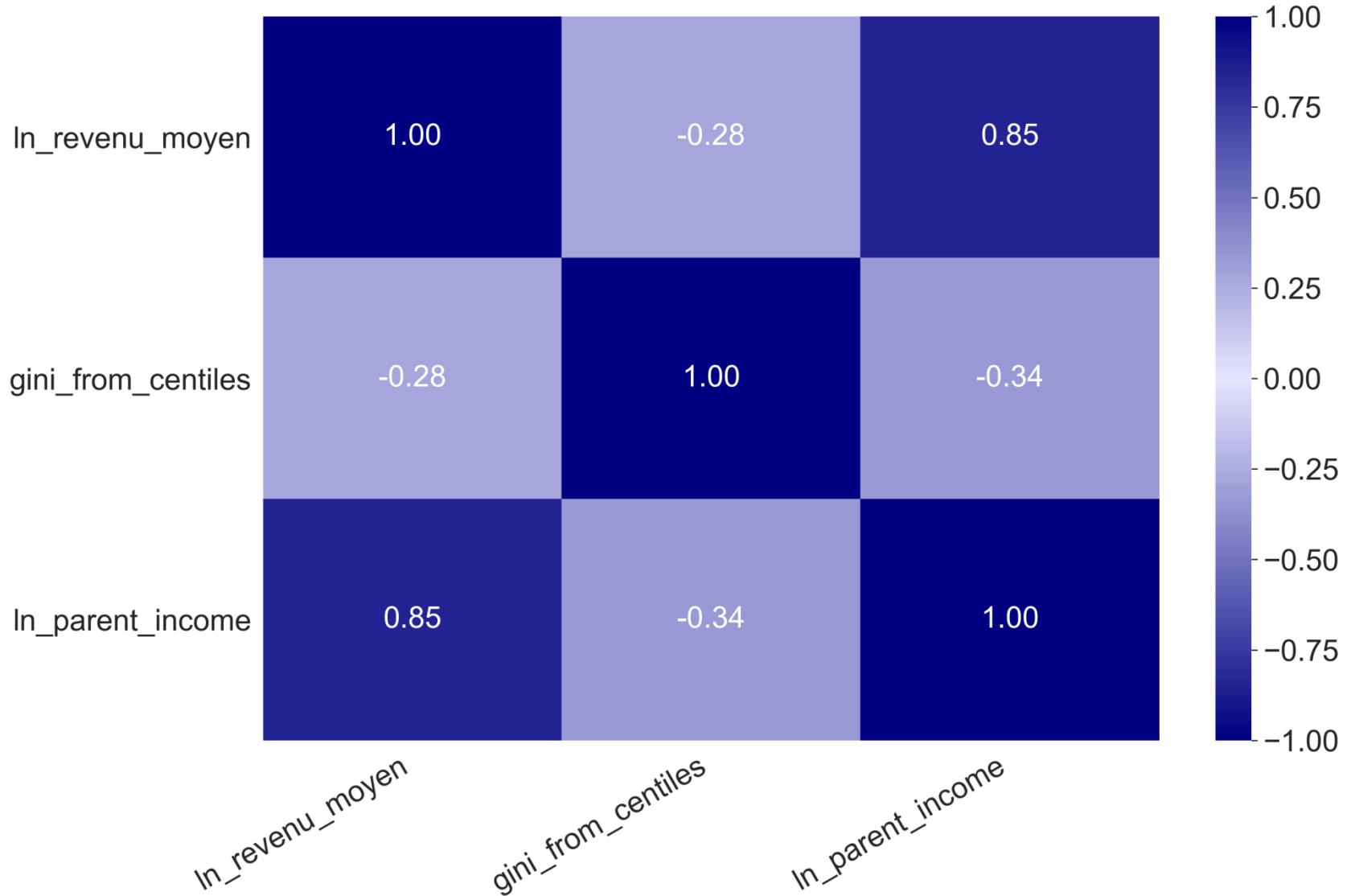


# Ajout du revenu des parents

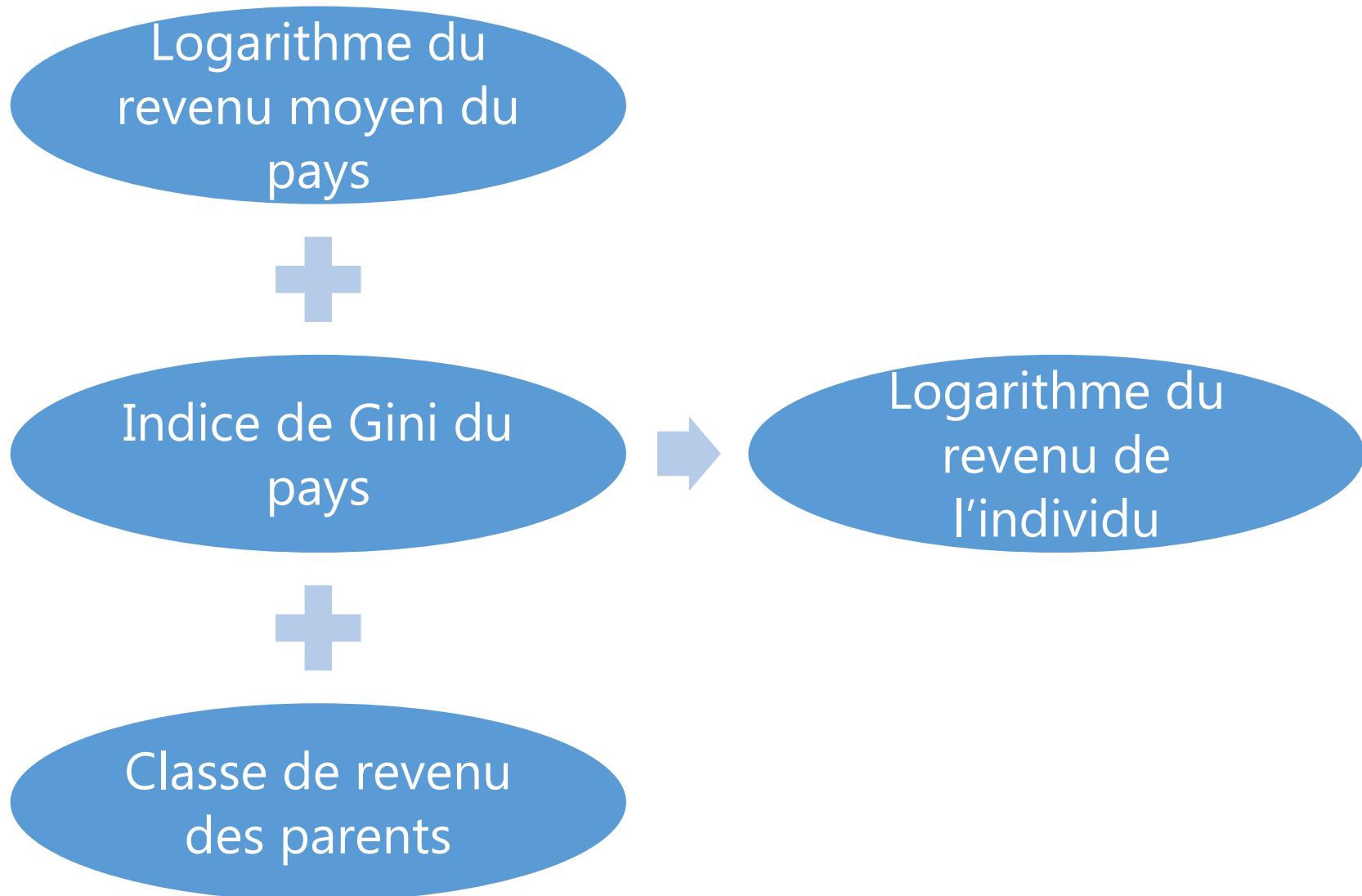


# Ajout du revenu des parents

Heatmap de la matrice de corrélation des variables explicatives

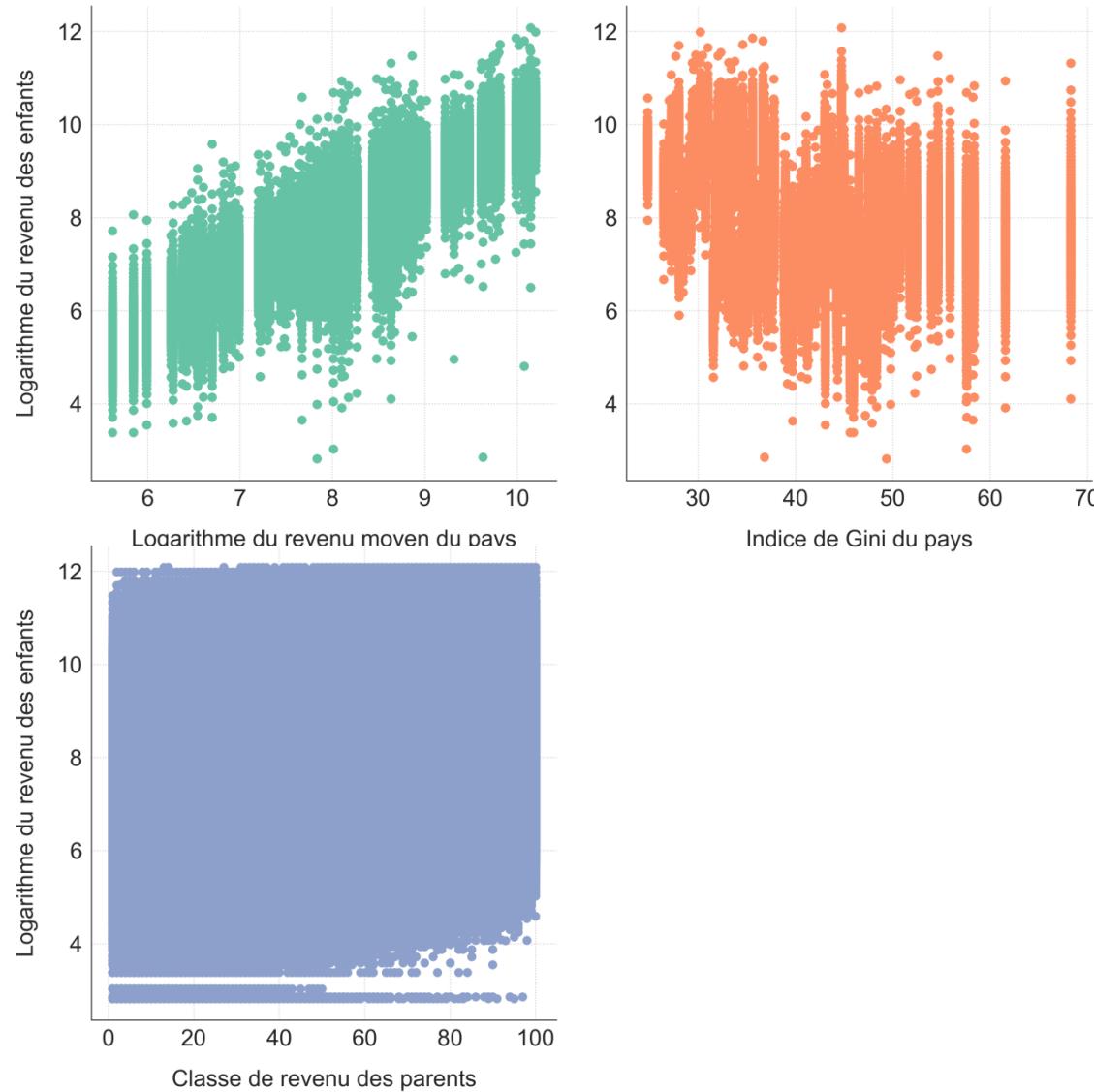


# Modèle final



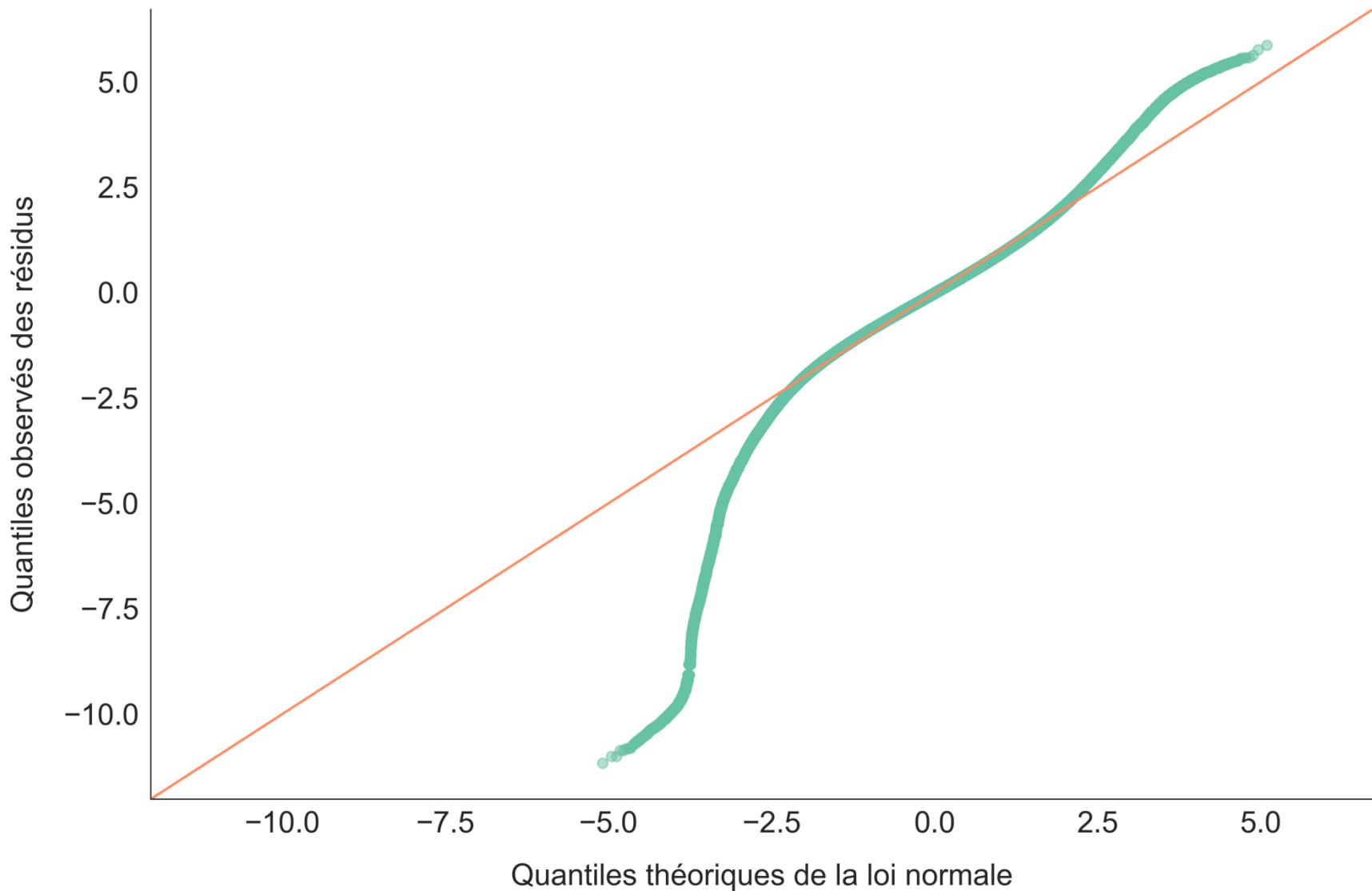
# Modèle final

Diagramme de dispersion du revenu des enfants en fonction du pays et du revenu des parents



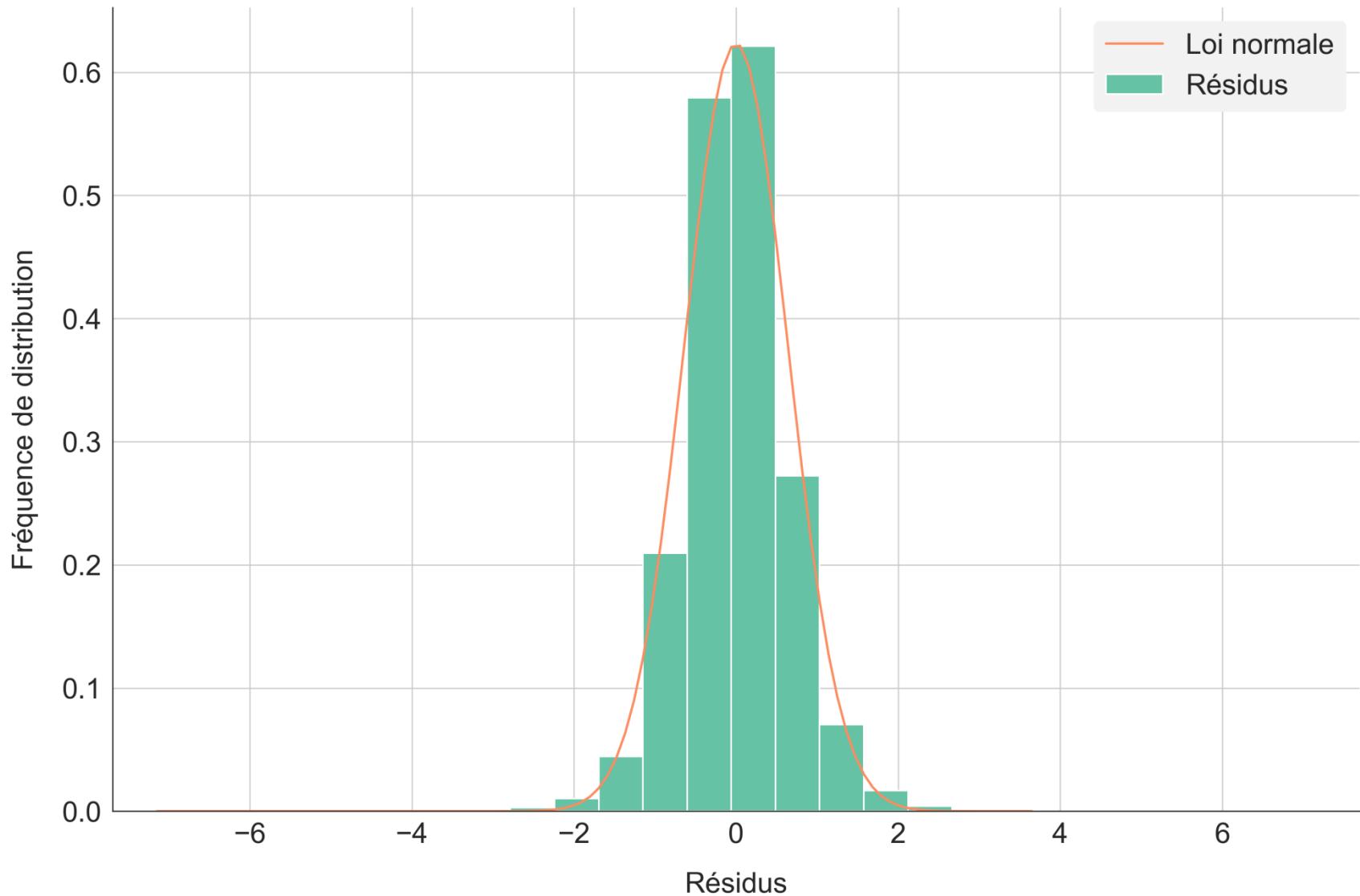
# Modèle final

Droite de Henry : vérification de la normalité des résidus



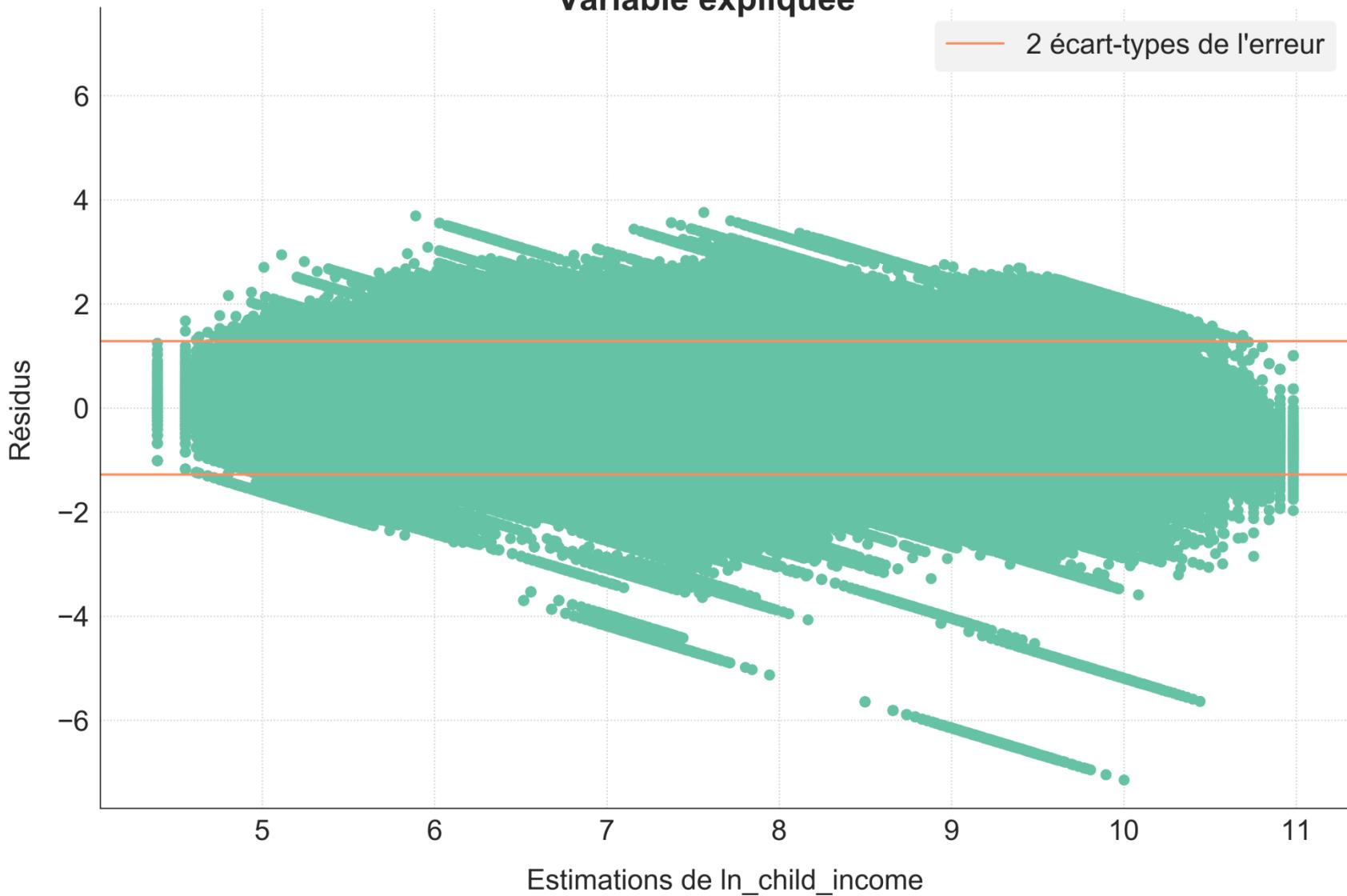
# Modèle final

Histogramme de la distribution des résidus



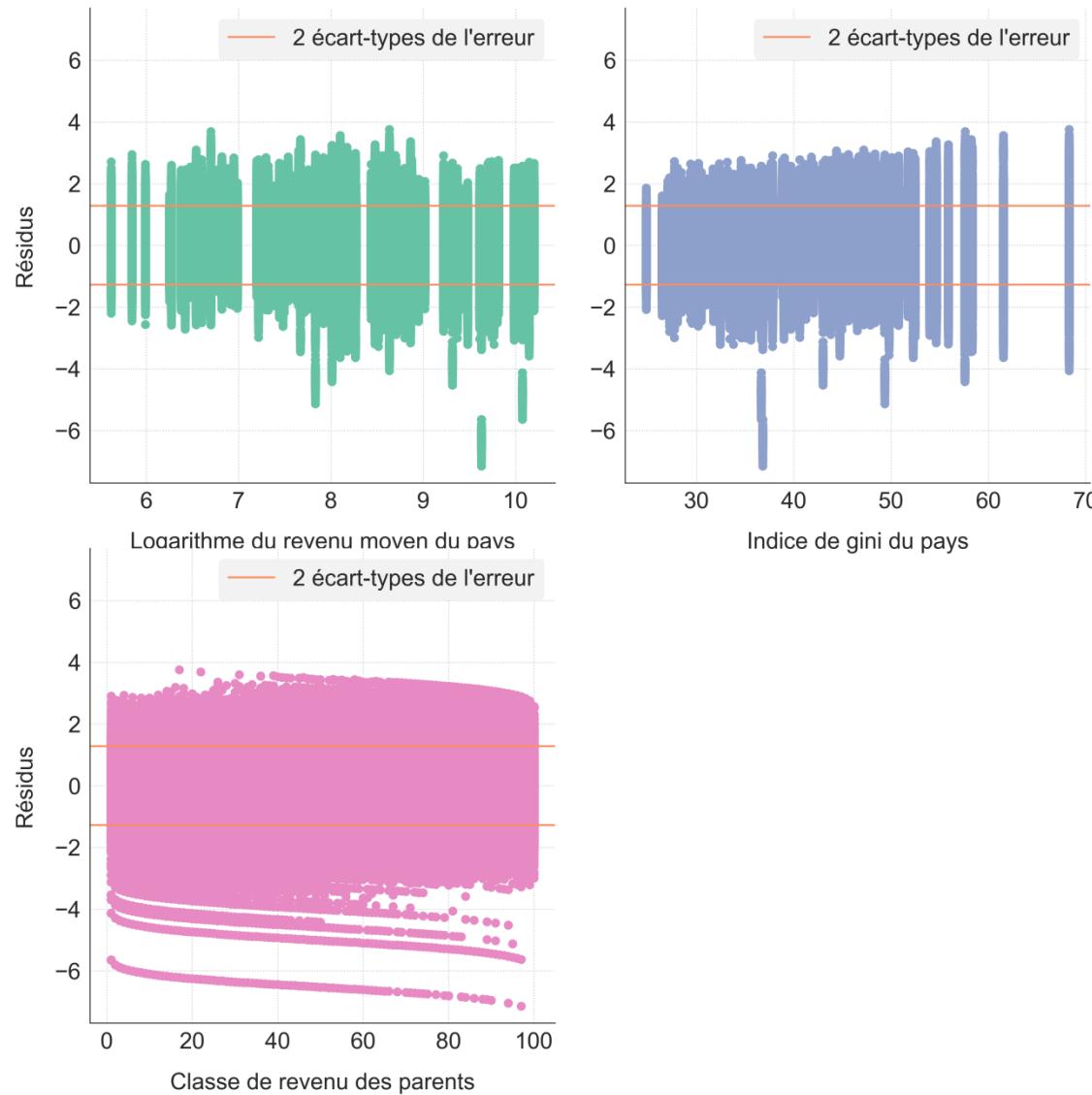
# Modèle final

Vérification de la linéarité de la relation et de l'homoscédasticité des résidus  
Variable expliquée



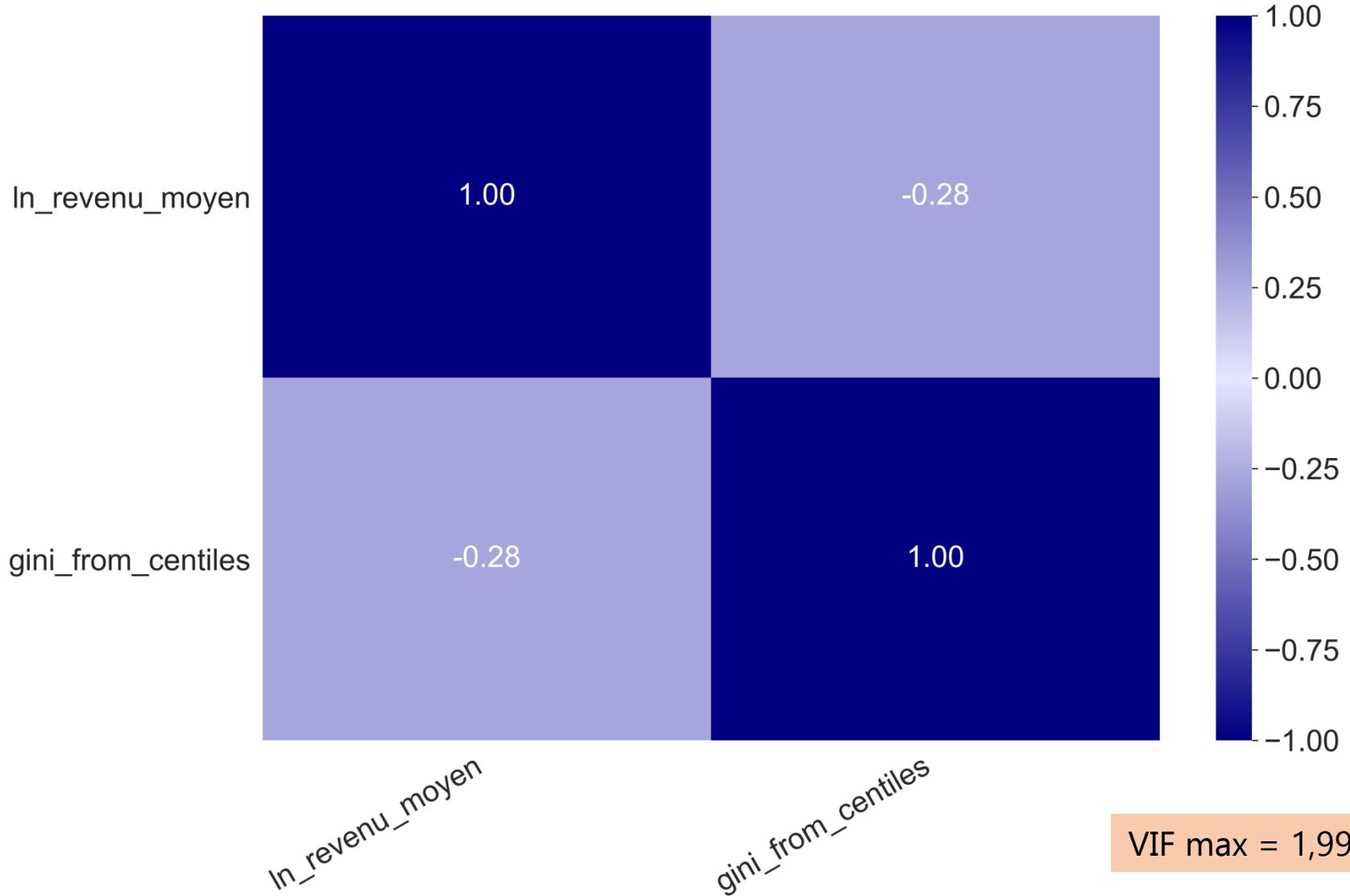
# Modèle final

Vérification de la linéarité de la relation, de l'homoscédasticité et de l'indépendance des résidus  
Variables exogènes



# Modèle final

Heatmap de la matrice de corrélation des variables explicatives



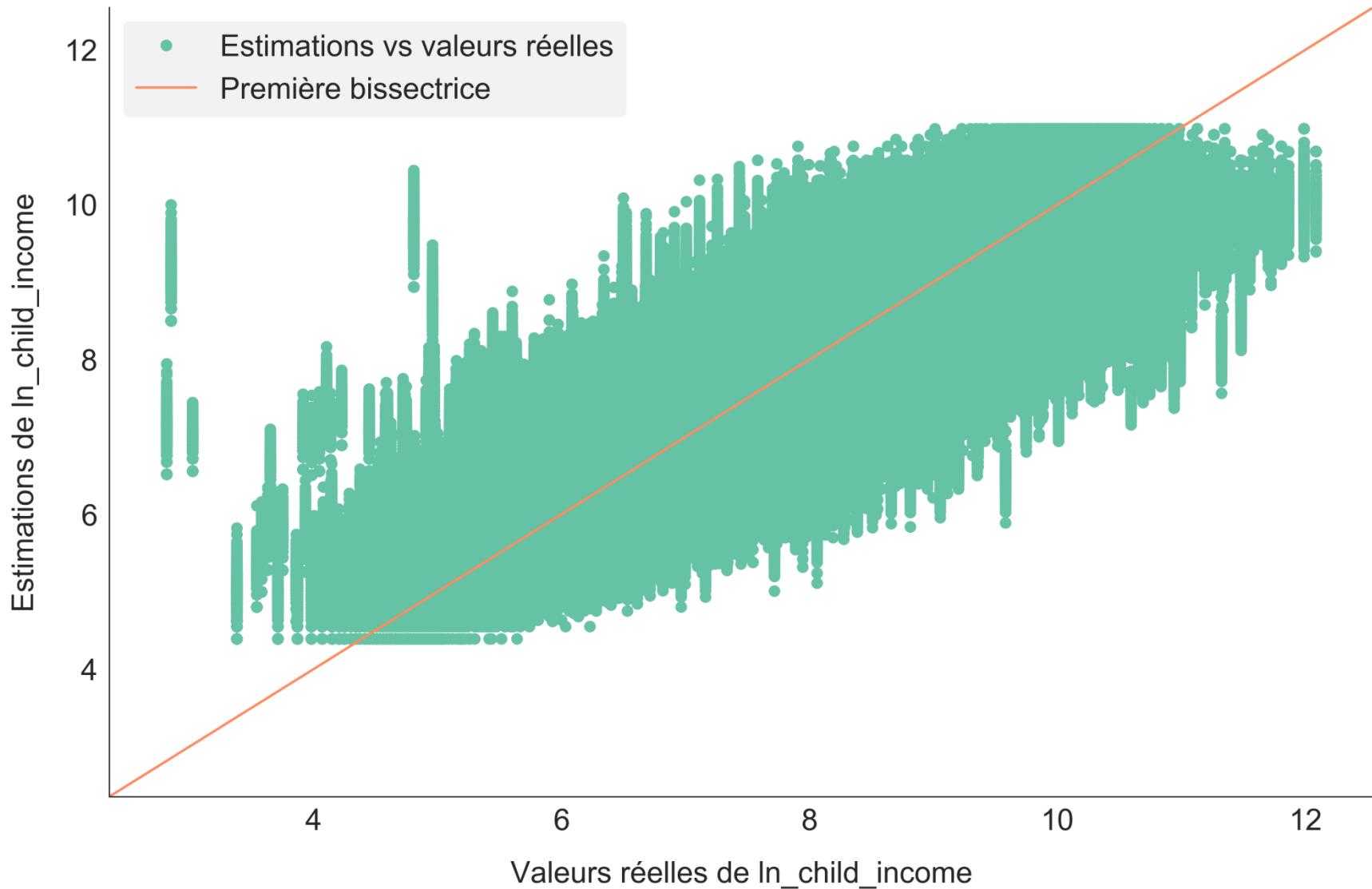
# Modèle final

Statistique	Signification	Interprétation
$R^2 = 0,786$ $R^2$ ajusté = 0,786	Qualité explicative du modèle	Le modèle explique 78,6% de la variance totale
Proba(F-stat) = 0,00	Test de signification globale du modèle	Modèle significatif
AIC = 10 980 000 BIC = 10 980 000	Qualité de prédiction du modèle	Meilleur modèle obtenu

ANOVA		Somme des carrés	Degrés de liberté	Proba(>F)	Omega <sup>2</sup>
Expliquée	In(revenu moyen)	6 662 860	1	0,000	0,686
	Indice de Gini	113 493	1	0,000	0,012
	Classe de revenu	622 871	99	0,000	0,064
Résiduelle		2 311 694	5 642 318		
Totale		9 710 918	5 642 419		

# Modèle final

Estimations par rapport aux valeurs réelles



# CONCLUSION

# QUESTIONS - RÉPONSES

