Projet n°7

Effectuez une prédiction de revenus

Introduction

Contexte :

Notre banque souhaite cibler de **nouveaux clients potentiels** et en particulier les jeunes en âge d'ouvrir leur premier compte bancaire. Notre cible seront les enfants de nos clients actuels.

Enjeu :

Cibler les **prospects** susceptibles d'avoir des **hauts revenus**.

Mission :

Créer un modèle permettant de déterminer le revenu potentiel d'une personne.

Sommaire

1. Les ressources disponibles
2. La distribution des revenus
3. La mobilité intergénérationnelle
4. Modélisation des données :
    1. ANOVA
    2. Régressions multiples

Les ressources disponibles

Le jeu de données

- Aperçu du jeu de données :

	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

- Pour chaque pays, nous disposons du revenu moyen divisé en 100 quantiles.

Le jeu de données

- Année(s) des données utilisées : **2004 et 2006 à 2011**
- Nombre de pays présents : **116**
- Population couverte par l'analyse ≈ **92 %**

- *De quel type de quantiles s'agit-il ?*

  Les classes de revenu sont découpés en centiles

- *Échantillonner une population en utilisant des quantiles est-il une bonne méthode ? Pourquoi ?*

  Le découpage en quantile est une bonne méthode car elle est robuste aux valeurs atypiques (i.e les très riches).

  Exemple d'un dataset avec un outlier :

  X: 1,2,2,3,1000

  Médiane = 2

  Moyenne = 201,6

Qu’est-ce que le PPP ?

- Purchasing Power Parity *(Parité de pouvoir d’achat)*
- Combien coûte un panier de biens dans un certains pays ?
- Cela permet de comparer des pays en retirant certains biais comme la politique monétaire.

Pays	USA	Japon
PIB	$16	$18
Prix d’un big mac	$2	$6
Panier de biens	8	3
GDP (PPP)	$16	$6

Marche à suivre

- Nous disposons des revenus des parents et nous souhaitons cibler les enfants de nos clients actuels.
- Notre objectif est d'établir une relation entre les revenus d'un parent et celui de ses enfants.
- Pour cela, nous allons simuler les revenus des parents en utilisant la formule du coefficient d' élasticité

$$\ln(Y_{child}) = p_j \ln(Y_{parent}) + \varepsilon$$

- A partir de cette formule nous allons pouvoir créer des probabilités conditionnelles de ce type :
- « Si cette individu vient du pays x et qu'il appartient à la classe de revenu de centile y, alors ses enfants ont une probabilité z_1 d'appartenir à la classe 1, z_2 d'appartenir à la classe 2 … z_{100} d'appartenir à la classe 100. »
- Nous disposerons alors de toutes les informations nécessaires afin de procéder à :
  - ANOVA (Analyse de la variance) entre le revenu d'un individu et son pays d'origine
  - Régressions multiples entre le revenu d'un individu et différentes variables explicatives

La distribution des revenus

Définition : Indice de Gini

- **Indice de Gini** : c'est une mesure statistique qui permet de rendre compte de le répartition d'une variable dans une population

- Dans notre cas, nous allons l'utiliser pour mesurer le niveau d'inégalité de la variable « *revenu* » au sein de la population d'un pays

- Note : l'indice est compris en 0 et 1, plus il est proche de 0, plus le pays est égalitaire. Si IG=0, tout le monde a le même salaire

Top des 5 indices les plus haut/bas

Top 5 des pays les égalitaires

Pays	Gini	Revenu moyen
Slovénie	0,23	11986€
Slovaquie	0,25	6036€
République Tchèque	0,25	8154€
Suède	0,25	16024€
Ukraine	0,26	3316€

Top 5 des pays les inégalitaires

Pays	Gini	Revenu moyen
République Centrafricaine	0,56	803€
Guatemala	0,57	2121€
Colombie	0,57	3512€
Honduras	0,60	3264€
Afrique du sud	0,67	5562€

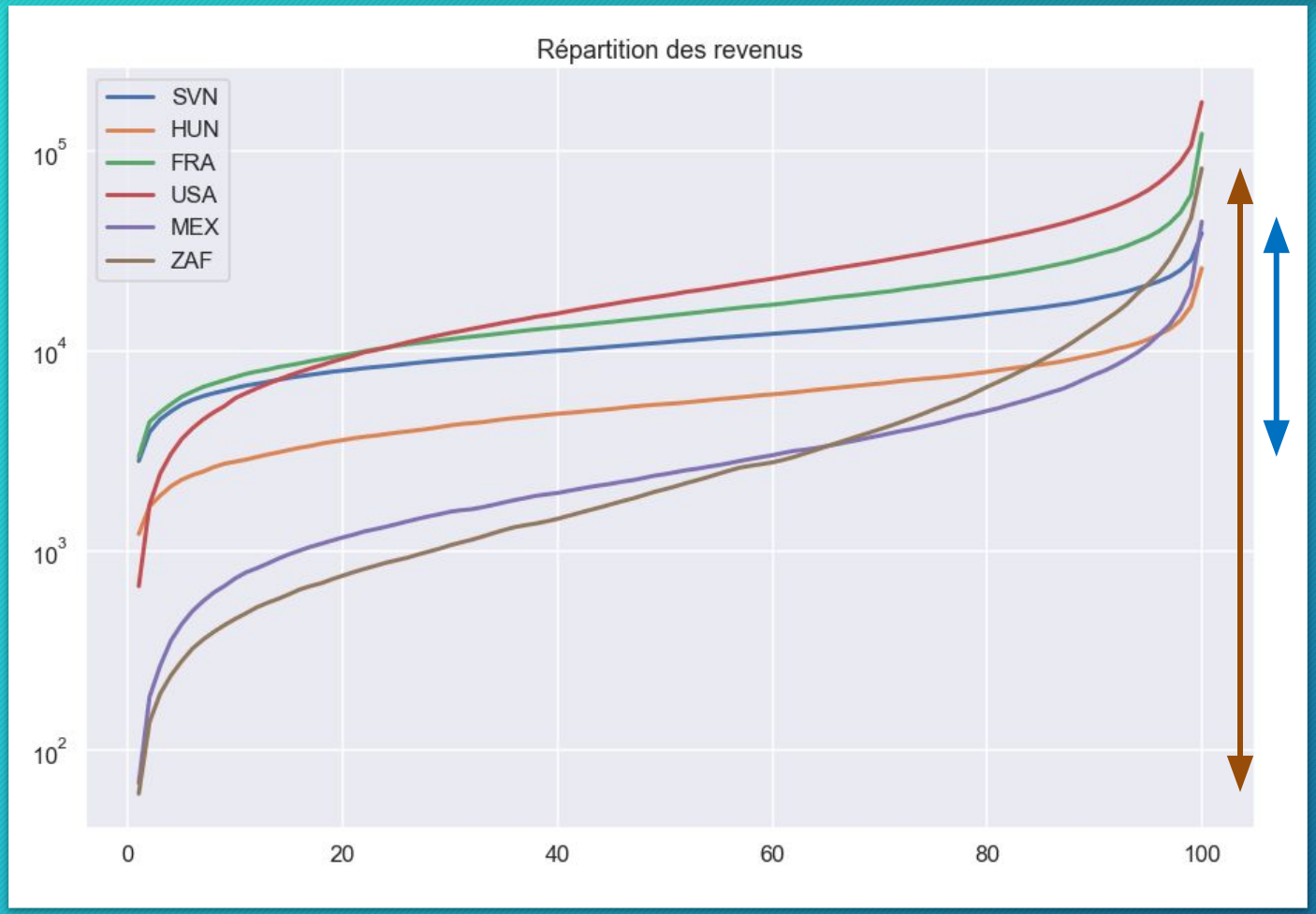
Classement de la France : 39e IG = 0,33

Sélection de pays

- Pour illustrer les différences de répartition de revenu entre les pays, j'ai sélectionné (en me basant sur l'indice de Gini):

Pays	Gini
Slovénie	0,23
Hongrie	0,27
France	0,33
Etats-Unis	0,43
Mexique	0,51
Afrique du sud	0,67

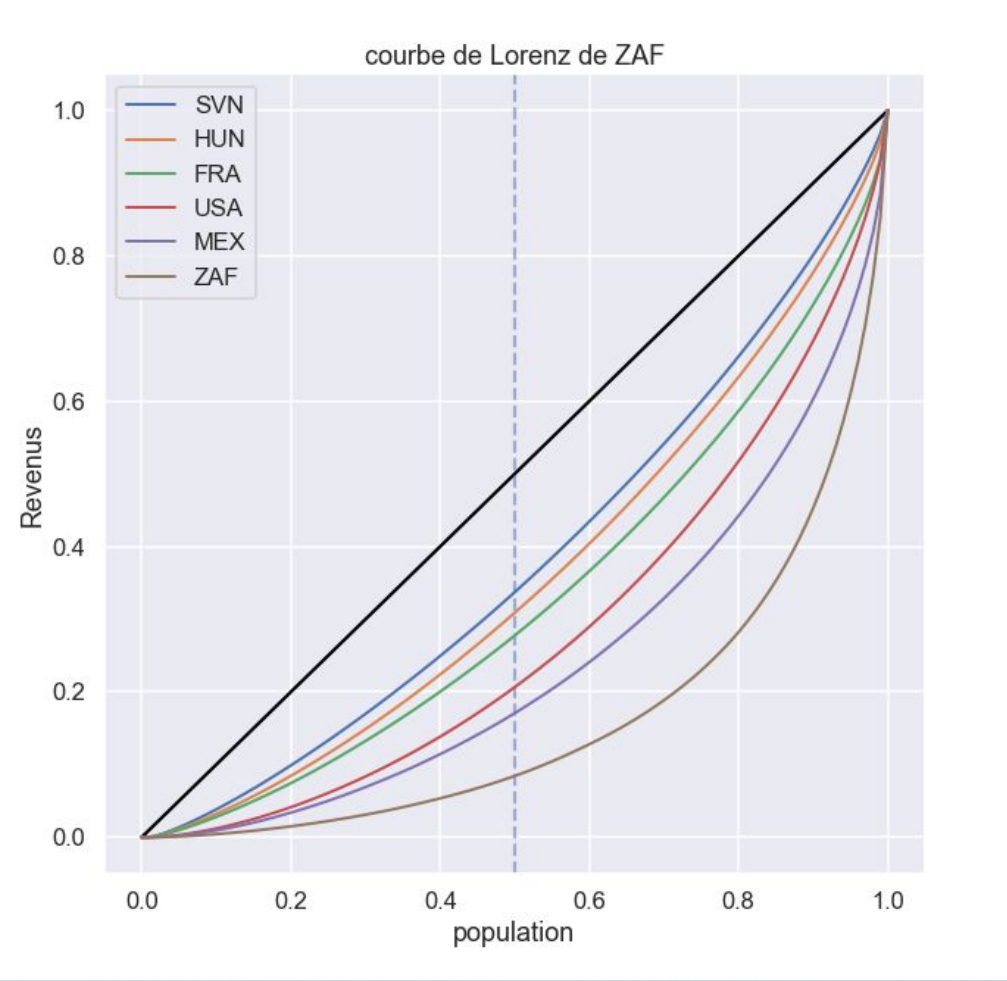
Diversité des pays en terme de distribution des revenus



Origines (supposées) :

- ZAF : apartheid
- USA : esclavage
- MEX : narco-trafiquants, corruption

Courbes de Lorenz



Les 50% les plus pauvres possèdent …

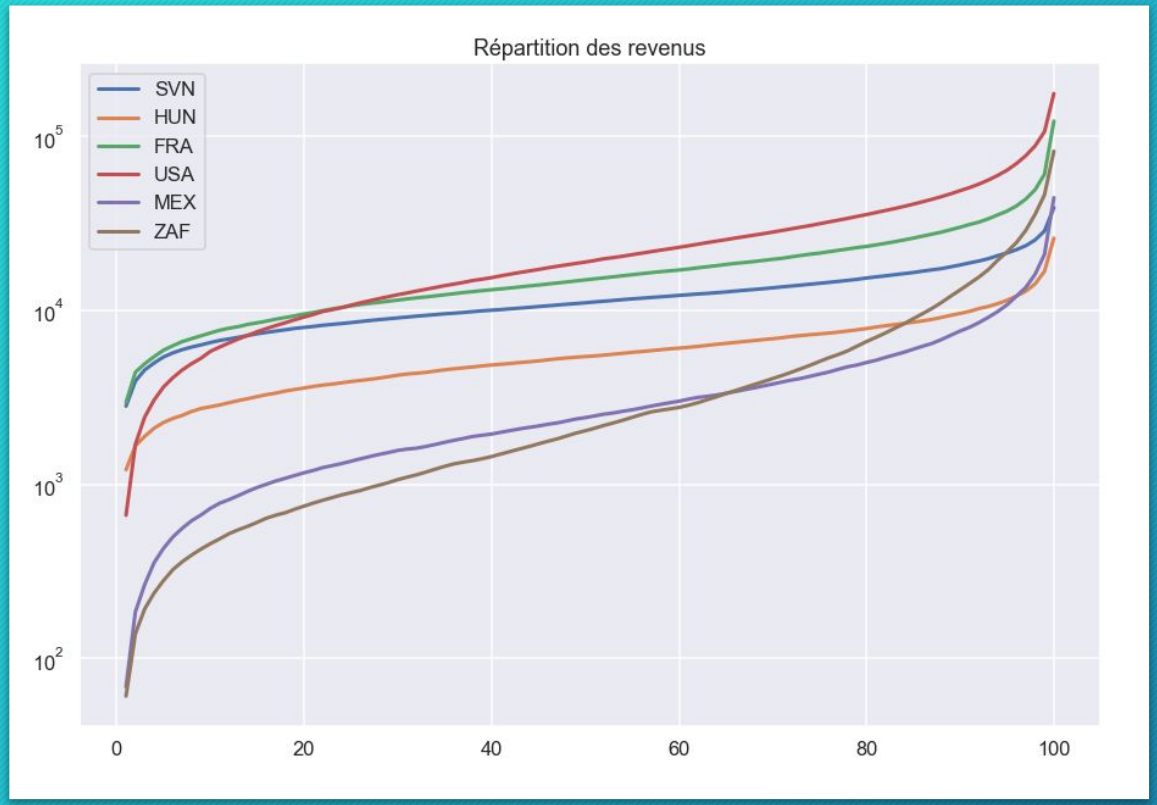
SVN ≈ 35%

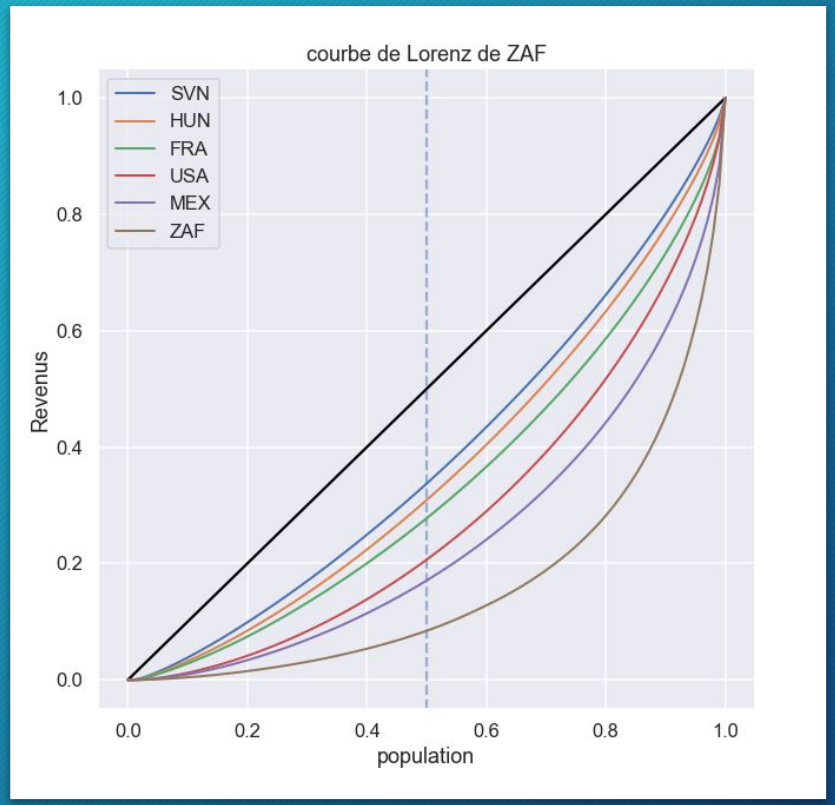
USA ≈ 20%

ZAF ≈ 10 %

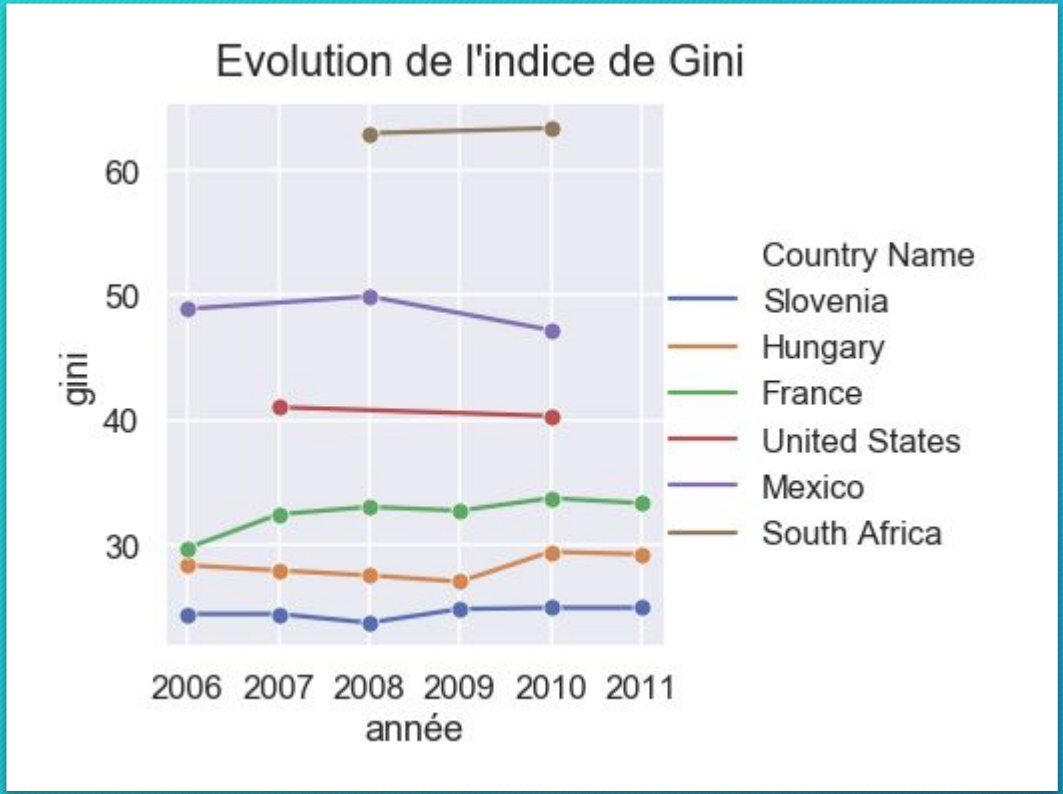
… des revenus du pays.

Comparaison : richesse ≠ inégalité





L’indice de Gini est-il stable dans le temps ?



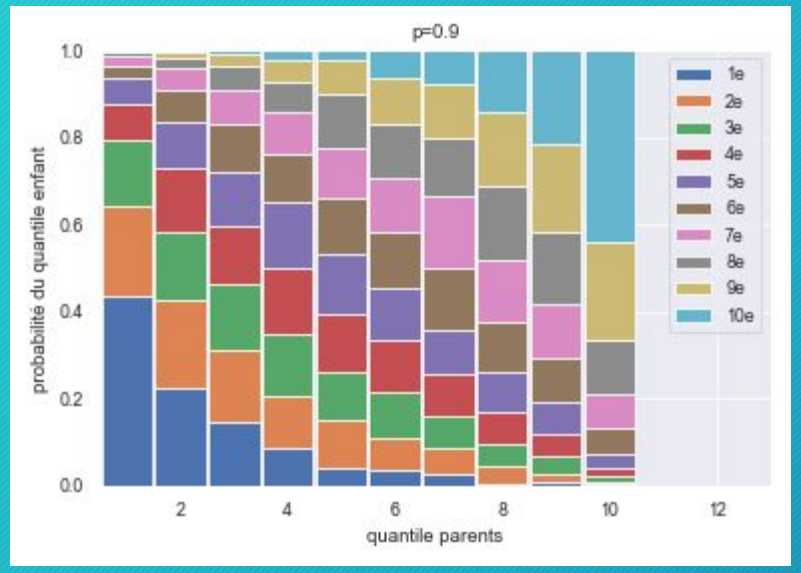
- Afin d’utiliser l’IG en temps que facteur explicatif pour nos régressions, il faut s’assurer que cet indice est stable dans le temps
- Le graphique nous laisse penser qu’aucune importante fluctuation impacte cette indice

La mobilité intergénérationnelle du revenu

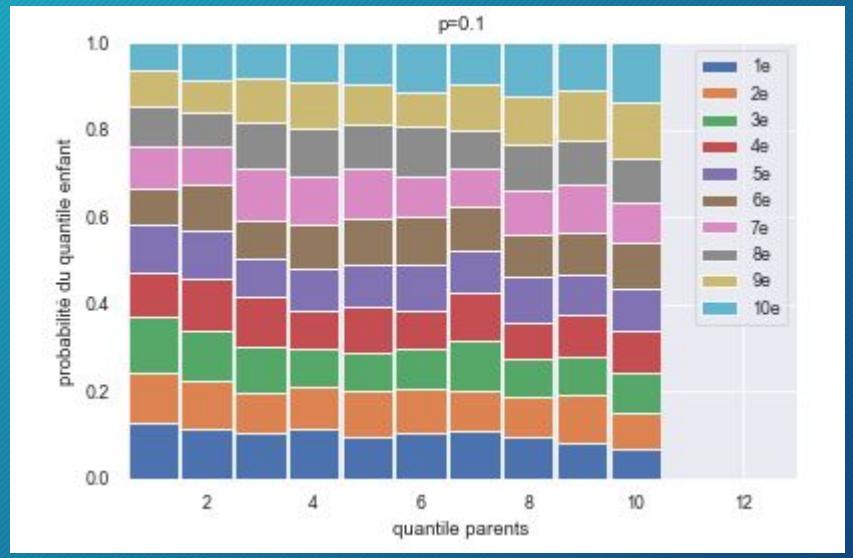
Définition : Coefficient d’élasticité

Hypothèse de cet exemple: les classes de revenu sont divisées en 10. ‘p’ est le coefficient d’élasticité

Faible mobilité



Forte mobilité



Simulation des classes de revenu

- Nous allons simuler la **classe de revenu des parents** en utilisant la formule du coefficient d'élasticité

$$\ln(Ychild) = p_j \ln(Yparent) + \varepsilon$$

Définitions des variables :

- « Ychild » : revenu des enfants, cette information est calculée
- « p_j » est le coefficient d'élasticité du pays
- « Yparent » est le revenu des parents, il sera généré selon une loi normale de paramètres (0,1)
- « ε » c'est le bruit des données : « les revenus de l'enfant que l'on ne peut pas expliquer avec les revenus des parents ». « ε » sera également généré selon une loi normale de paramètres (0,1)

Pourquoi avoir simulé Yparent et Epsilon ?

- Nous pouvons généré « Yparent » et « ε » selon une loi normale centré réduite dans la mesure où nous ne sommes pas intéressé par les revenus mais par la classe de revenu.

- L'objectif étant de générer des probabilités conditionnelles. Nous ne sommes pas contraint d'utiliser les données du dataset. En effet, nos données ne sont pas une représentation parfaite de la réalité et nous aurons un meilleur résultat en simulant ces données selon une loi normale.

Probabilités conditionnelles

La formule utilisée précédemment nous donne des probabilités conditionnelles pour chaque pays.

En dupliquant chaque ligne par 1000, on peut appliquer les probabilités conditionnelles pour chaque individu

Exemple : Individus issus de la classe de revenu de quantile 5 dont les parents étaient de classe revenu de quantile 70

country	gini	elasticity	revenu_moyen	gdpppp	quantiles_enf	revenu_classe	classe_parents
FIN	0.276857	0.1	16144.884419	33626.0	70	17835.383	5
FIN	0.276857	0.1	16144.884419	33626.0	70	17835.383	5
FIN	0.276857	0.1	16144.884419	33626.0	70	17835.383	5
FIN	0.276857	0.1	16144.884419	33626.0	70	17835.383	5
FIN	0.276857	0.1	16144.884419	33626.0	70	17835.383	5

Modélisation

22

ANOVA

Les modélisations

- ANOVA
  - Le pays d'origine d'un individu a-t-il un impact sur ses revenus ?
  - Variable dépendante (quantitative) :
    - Revenu de l'individu
  - Variable explicative (qualitative/catégorielle):
    - Pays d'origine

ANOVA : résultats

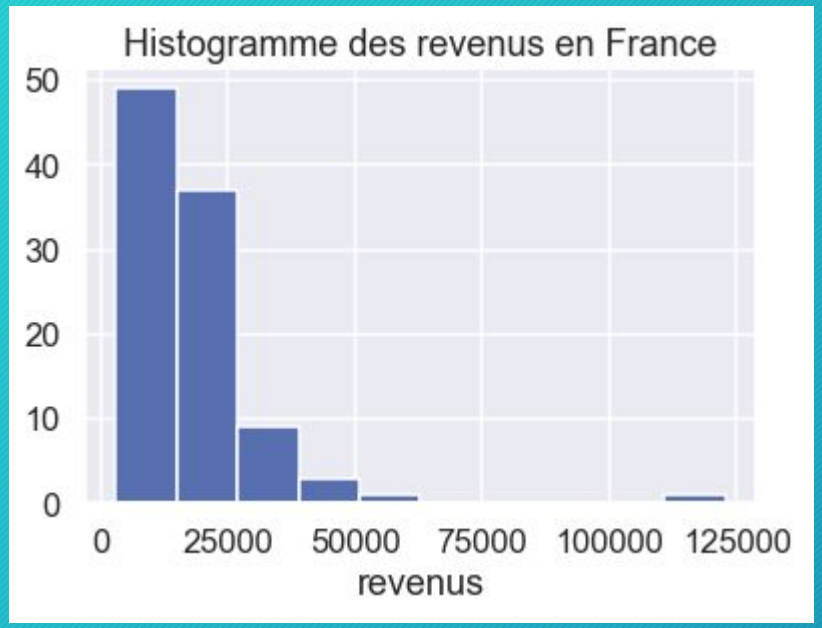
- H_0 = il n'y a pas de différence entre les moyennes des groupes
- H_1 = les moyennes des groupes sont différentes

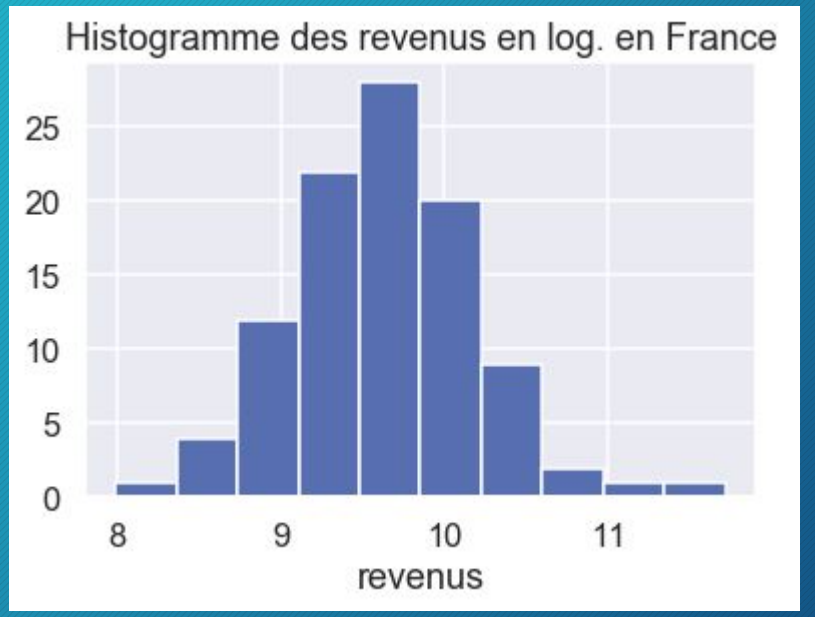
- R^2 (ajusté)= 0,491
- F-stat = 98,43
- P-value ≈ 0
- 64,66% ont une p-value < 0,05

- Seuil alpha retenu = 0,05
- Nous pouvons rejeter l'hypothèse nulle avec une forte présomption.

- Un nombre trop important de pays à une p-value qui dépasse le seuil alpha

Distribution des revenus vs revenus en log





ANOVA en log : résultats

- H_0 = il n'y a pas de différence entre les moyennes des groupes
- H_1 = les moyennes des groupes sont différentes

- R^2 (ajusté)= 0,729 (Première ANOVA : 0,491)
- F-stat = 269 (Première ANOVA : 98,43)
- P-value ≈ 0 (Première ANOVA : 0)
- 91,38% des pays ont une p-value < 0,05 (Première ANOVA : 64,66%)

- Seuil alpha retenu = 0,05
- Nous pouvons rejeter l'hypothèse nulle avec une forte présomption.

- Désormais, on peut affirmer que notre modèle fonctionne pour la plupart des pays

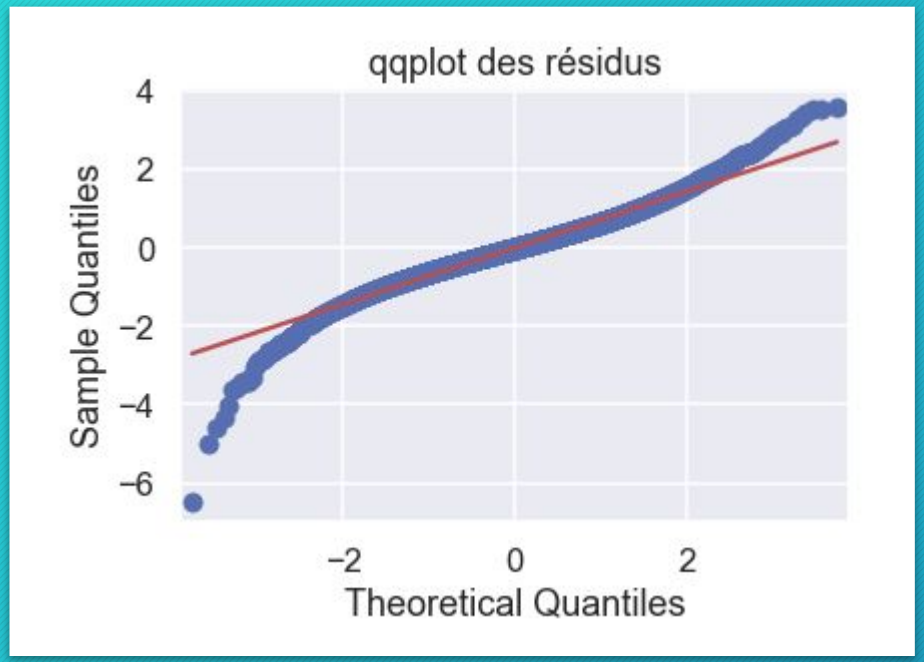
ANOVA : Vérifications des hypothèses

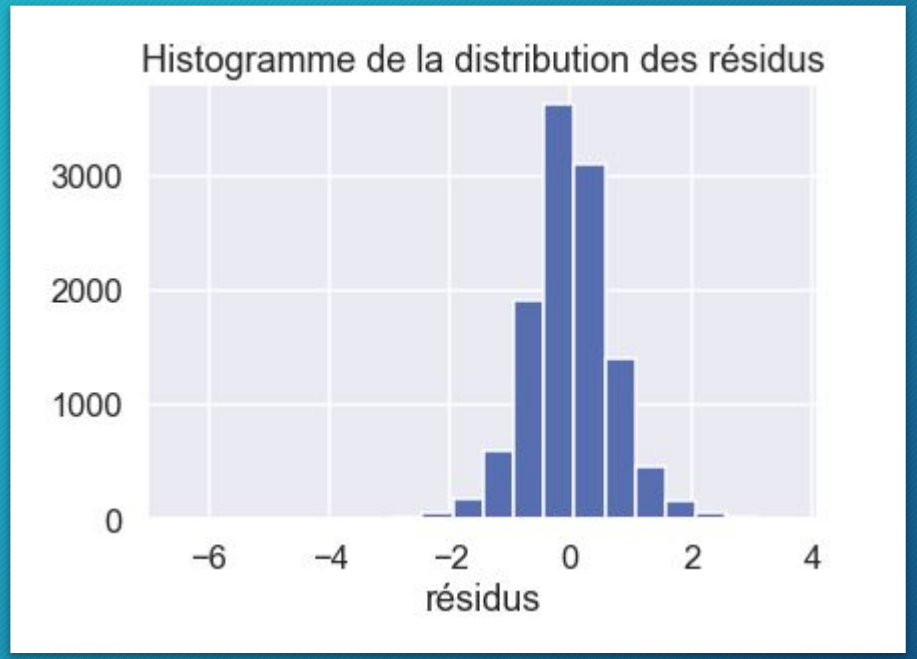
- 3 hypothèses à vérifier la légitimité d'une ANOVA paramétrique :
  - Homoscédasticité
  - Distribution normale des résidus
  - Les individus sont i.i.d (indépendantes et identiquement distribué)

Les individus sont i.i.d

- Les individus sont répartis en classe de 1 à 100. On peut donc affirmer que l'échantillon est représentatif de la population

Les résidus des populations étudiées suivent une distribution normale



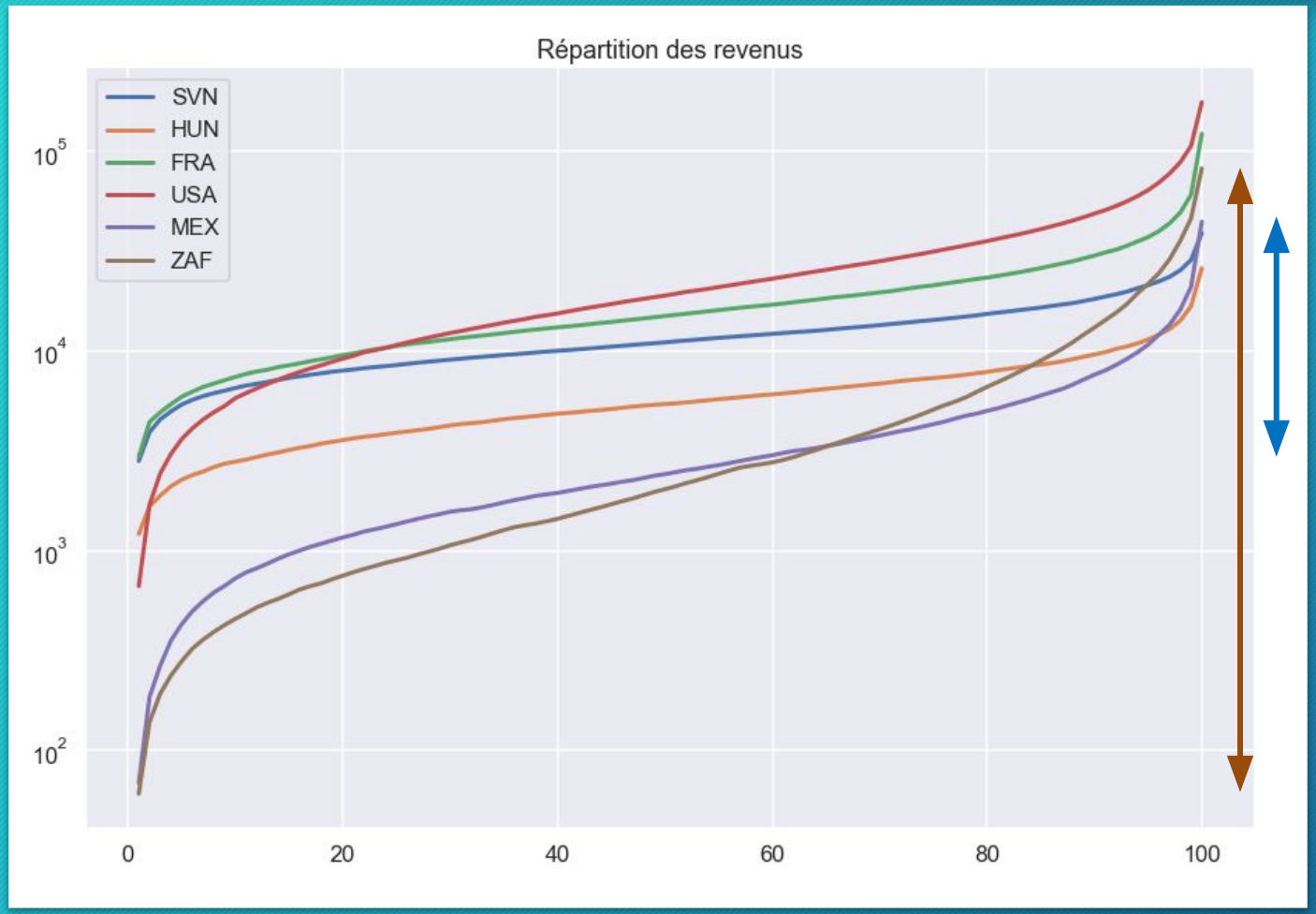


- **Hypothèse** : la distribution est normale si les points sont alignés sur la droite
- D'après la lecture graphique, nous supposons que la distribution est normale

Homoscédasticité

- Homoscédasticité : les variances des différents groupes sont égales
- Test de bartlett :
  - H_0 = les variances sont égales
  - H_1 = les variances sont différentes
  - Stat : 1514,6
  - P-value ≈ 0
- Nous pouvons rejeter l'hypothèse nulle avec une forte présomption.
- L'homoscédasticité n'est pas vérifié. Cependant comme mes autres critères sont vérifiés et que la p-value de l'ANOVA est proche de 0, j'estime que l'ANOVA est robuste à ce critère

Intuition sur la non-homoscédasticité



- En fonction de l'indice de Gini, l' écart entre les plus riches et les plus pauvres varient.
- Cela conduit à une différence entre les variances des pays

Les pays dont la p-value n'est pas satisfaisante

	country	gini	elasticity	revenu_moyen	gdpppp
12	BRA	0.544494	0.6	4759.891127	9559.0000
27	DOM	0.502524	0.7	3523.175376	7505.0000
52	JOR	0.337725	0.5	3018.449854	5082.9316
68	MEX	0.507989	0.7	3847.361739	13434.0000
84	PAN	0.531917	1.0	5084.301678	11767.0000
85	PER	0.478034	0.7	3297.562504	7858.0000
91	ROU	0.373422	0.4	3285.486062	11782.0000
108	UKR	0.255140	0.4	3316.228572	6721.0000
111	VEN	0.434021	0.7	3135.794123	11756.0000
115	ZAF	0.669779	0.7	5562.288691	9602.0000

Pour la majorité, ils ont un indice de Gini élevé (sauf Ukraine)

Point positif : Ce sont des pays où le revenu moyen est assez faible

Modélisation

Régressions multiples

Régressions multiples

1. **Revenu d'un individu** en fonction du **coefficient de gini** et du **revenu moyen du pays**
2. **Revenu d'un individu (log)** en fonction du **coefficient de gini** et du **revenu moyen du pays(log)**
3. **Revenu d'un individu** en fonction du **coefficient de gini**, du **revenu moyen du pays** et de la **classe de revenus des parents**
4. **Revenu d'un individu (log)** en fonction du **coefficient de gini**, du **revenu moyen du pays (log)** et de la **classe de revenus des parents**

: variable dépendante
: variable(s) explicative(s)

Pourquoi une transformation en log ?

- C’est l’une des transformations de données possibles parmi d’autres :
  - 1/x
  - Exp(x)
  - Etc.
- L’intuition du log :
  - J’ai représenté la répartition des revenus à l’aide d’une échelle logarithmique
  - Ça semblait être la transformation de données la plus logique

2 facteurs : Gini et revenu moyen

- **Revenu d'un individu** en fonction du **coefficient de gini (x1)** et du **revenu moyen du pays (x2)**
- Y = -3,822 * *x1* + 1,009 * *x2* + 3,927
- R^2 (variance expliquée) = 0,4954
- P-value ≈ 0

- **Revenu d'un individu (log)** en fonction du **coefficient de gini (x1)** et du **revenu moyen du pays(log) (x2)**
- Y = -1,634 * *x1* + 0,986 * *x2* + 0,471
- R^2 = 0,728
- P-value ≈ 0

3 facteurs : Gini, revenu moyen et classe parents

- **Revenu d'un individu** en fonction de la **classe parent (x1)**, du **revenu moyen du pays (x2)** et du **coefficient de gini (x3)**
- **Y = 51,382 * *x1* + 1,009 * *x2* -6,472 * x3 -2589,472**
- **R^2 = 0,52**
- **P-value ≈ 0**

- **Revenu d'un individu (log)** en fonction de la **classe parent (x1)**, du **revenu moyen du pays (log)(x2)** et du **coefficient de gini (x3)**
- **Y = 0,01 * *x1* + 0,986 * *x2* + -1,635 * x3 - 0,082**
- **R^2 = 0,78**
- **P-value ≈ 0**

Analyse des résultats

Tests	Régression n°2	Régression n°4
Significativité des paramètres	Les p-values de toutes les variables sont inférieures au seuil alpha	
Atypicité des données : Variables explicatives *Calcul des leviers*	1363563 observations sont supérieurs aux 3 seuils	1363959 observations sont supérieurs aux 3 seuils
Atypicité des données : Variable à expliquer : *Calcul des résidus studentisés*		
L'influence des données *Distance de Cook*		
Colinéarité (*Variance Inflation Factor*)	Aucune variable n'a un VIF > 10	
Homoscédasticité (*Test de Breusch Pagan*)	Le test n'est pas concluant	
Normalité des résidus (*Jarque-Bera*)	Le test n'est pas concluant	

L’indice de Gini

- Le coefficient de régression associé à l’indice de Gini est négatif dans ces 4 régressions.
- Cela signifie que le revenu d’un individu est inversement corrélé au coefficient de Gini.

- *Peut-on affirmer que le fait de vivre dans un pays plus inégalitaire favorise plus de personnes qu’il n’en défavorise ?*
  - *Non, ce coefficient négatif est influencé par une minorité de très riche*

Conclusion

- Le revenu d'un individu est expliqué à 78% par :

Son pays d'origine(revenu moyen et mobilité intergénérationnelle) et les revenus de ses parents

Recommandations pour le ciblage des nouveaux clients (maximisation des profits) :

- Cibler des pays dont la mobilité intergénérationnelle est faible pour limiter le facteur aléatoire.
- Au sein de ces pays cibler les individus dont les parents sont issus des classes de revenu les plus élevés.