Université de Reims Champagne-Ardenne

Faculté des sciences économiques, sociales et de gestion

Master 1 - Analyse et politque économique

Master 1 - Mathématiques et applications

Parcours statistique pour l'évaluation et la prévision

Déterminants du bonheur

Projet - Méthodes d'échantillonnages

Année universitaire 2020 - 2021

par

Ouael ETTOUILEB, Jacob DEMBELE, Pierre Emmanuel BADIN et Maxime ANGOULVENT

Encadrant universitaire : Amor KEZIOU

Table des matières

Liste des tableaux

Table des figures

Introduction

La recherche du bonheur constitue un objectif pour chaque individu. Il est le sens même de la vie. Certains philosophes considèrent que le bonheur est la condition nécessaire pour une vie réussie. D'après Kant : « Le pouvoir, la richesse, la considération, même la santé ainsi que le bien-être complet et le contentement de son état, est ce qu'on nomme le bonheur » (Métaphysique des mœurs). L'étude que nous réalisons est basée sur le jeu de données : «World Happiness report 2021» détaillé ultérieurement. Notre étude a pour but de déterminer si certains facteurs influencent le bonheur de la population et en quelle proportion. Pour réaliser cette étude, nous réaliserons tout d'abord un état des lieux du bonheur dans le monde à travers divers aspects par une analyse descriptive. Ensuite, nous créerons un modèle expliquant le bonheur selon différents facteurs tels que le PIB, l'espérance de vie, le soutien social, etc...

Données utilisées

Il s'agit d'un jeu de données issu de l'enquête «World Happiness report 2021» réalisée par «Gallup World Poll», dont le but est d'étudier le bonheur en 2021 dans différentes régions du monde. Il met en relation le bonheur et d'autres variables qui sont susceptibles de l'influencer. Nous avons 8 variables observées sur 149 individus. Nous présentons ci-dessous les différentes variables :

- Country name est la variable qui permet d'identifier les pays ;
- Ladder score correspond au score du bonheur ou le bien-être subjectif. Les valeurs de cette variable proviennent de la publication du 26 février 2021 du Gallup World Poll (GWP) couvrant les années 2005 à 2020. Ces valeurs sont obtenues en calculant la réponse moyenne nationale à la question sur l'évaluation de la vie. La formulation anglaise de la question est la suivante : "Imaginez une échelle, dont les échelons sont numérotés de 0 en bas à 10 en haut. Le haut de l'échelle représente la meilleure vie possible et le bas de l'échelle représente la pire vie possible. Sur quel échelon de l'échelle diriez-vous que vous vous situez personnellement à ce moment-là ?" Cette mesure est également appelée échelle de vie de Cantril.
- Logged GDP per capita correspond aux statistiques du PIB par habitant (GDP) ayant subi une transformation logarithmique. Le PIB par habitant est la valeur du PIB créé dans chaque pays, rapportée à l'effectif de leurs populations respectives. Ces valeurs sont exprimées en parité de pouvoir d'achat (PPA) en dollars constants de 2017 et sont issues de la mise à jour du 14 octobre 2020 des Indicateurs du développement mondial (WDI). Les chiffres du PIB de Taïwan, de la Syrie, de la Palestine, du Venezuela, de Djibouti et du Yémen proviennent du tableau 9.1 de PennWorld. Le PIB par habitant en 2020 n'est pas encore disponible en décembre 2020 alors les concepteurs du rapport se sont appuyés sur les perspectives économiques de l'OCDE n° 108 (décembre 2020), et de la banque mondiale (dernière mise à jour : 06/08/2020) pour prolonger la série chronologique du PIB par habitant de 2019 à 2020 en utilisant les prévisions de croissance du PIB réel par pays

en 2020. Les prévisions de croissance du PIB sont ajustées pour tenir compte de la croissance de la population en soustrayant la croissance de la population de 2018-19 de la croissance prévue pour 2019-20.

— Social support (le soutien social) ; Cette variable correspond au fait d'avoir quelqu'un sur qui compter en cas de problème. C'est la moyenne nationale des réponses binaires (0 ou 1) à la question du GWP "Si vous aviez des problèmes, avez-vous des parents ou des amis sur lesquels vous pouvez compter pour vous aider quand vous en avez besoin, ou pas ?".

— health life expectancy (Espérance de vie en bonne santé EVBS). Selon l'INSEE, "L'espérance de vie en bonne santé est la durée de vie moyenne en bonne santé - c'est-à-dire sans limitation irréversible d'activité dans la vie quotidienne ni incapacités - d'une génération fictive soumise aux conditions de mortalité et de morbidité de l'année. Elle caractérise la mortalité et la morbidité indépendamment de la structure par âge".

— Social support (le soutien social) ; Cette variable correspond au fait d'avoir quelqu'un sur qui compter en cas de problème. C'est la moyenne nationale des réponses binaires (0 ou 1) à la question du GWP "Si vous aviez des problèmes, avez-vous des parents ou des amis sur lesquels vous pouvez compter pour vous aider quand vous en avez besoin, ou pas ?".

— health life expectancy (Espérance de vie en bonne santé EVBS). Selon l'INSEE, "L'espérance de vie en bonne santé est la durée de vie moyenne en bonne santé - c'est-à-dire sans limitation irréversible d'activité dans la vie quotidienne ni incapacités - d'une génération fictive soumise aux conditions de mortalité et de morbidité de l'année. Elle caractérise la mortalité et la morbidité indépendamment de la structure par âge". C'est une mesure plus intéressante que l'espérance de vie pour comparer les niveaux de vie entre pays car comme le disait en 1997 le directeur général de l'OMS, le Dr Hiroshi Nakajima, "sans qualité de la vie, une longévité accrue ne présente guère d'intérêt (...), l'espérance de vie en bonne santé est plus importante que l'espérance de vie".Les données de l'EVBS qui figurent dans ce jeu de données sont extraites du dépôt de données de l'Observatoire mondial de la santé (OMS) (Dernière mise à jour : 2020-09-28). Les données de cette source sont disponibles pour les années 2000, 2005, 2010, 2015 et 2016. Pour la période d'échantillonnage de ce rapport (2005-2020), l'interpolation et l'extrapolation sont utilisées.

— Freedom to make life choices (la liberté de faire des choix de vie) est la moyenne nationale des réponses à la question du GWP "Êtes-vous satisfait ou insatisfait de votre liberté de choisir ce que vous faites de votre vie ?".
— Generosity (la générosité) est le résidu de la régression de la moyenne nationale des réponses à la question du GWP "Avez-vous donné de l'argent à une organisation caritative au cours du dernier mois ?" sur le PIB par habitant.
— Perceptions de corruption (Perception de la corruption) est la moyenne nationale des réponses à deux questions du GWP : "La corruption est-elle répandue au sein du gouvernement ou non" et "La corruption est-elle répandue au sein des entreprises ou non" ? La perception globale est simplement la moyenne des deux réponses binaires (0 ou 1). Si la perception de la corruption au sein du gouvernement est manquante, la perception de la corruption dans les entreprises est utilisée comme perception globale. La perception de la corruption au niveau national est simplement la réponse moyenne de la perception globale au niveau individuel.

Statistiques descriptives

Regional indicator	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices
South Asia	4.441857	8.682571	0.703429	62.681000	0.765000
Sub-Saharan Africa	4.494472	8.075194	0.696750	55.886472	0.723194
Middle East and North Africa	5.219765	9.666118	0.797647	65.609118	0.716471
Southeast Asia	5.407556	9.421444	0.820333	64.888444	0.909000
Commonwealth of Independent States	5.467000	9.401833	0.872500	65.009500	0.816917
East Asia	5.810333	10.367667	0.860500	71.252167	0.763500
Latin America and Caribbean	5.908050	9.370000	0.839500	67.076050	0.831750
Central and Eastern Europe	5.984765	10.109059	0.887412	68.338412	0.797059
Western Europe	6.914905	10.822714	0.914476	73.033095	0.858714
North America and ANZ	7.128500	10.809500	0.933500	72.325000	0.898750

FIGURE 1 – Moyennes par régions

Ici, nous allons tenter de décrire et de résumer le jeu de données par des graphiques et des indicateurs statistiques pertinents afin de comprendre les phénomènes relatifs au bonheur dans le monde. Tout d'abord nous commencerons par une analyse univariée de chaque variable pour comprendre son comportement, ensuite nous verrons une analyse bivariée afin de voir comment certaines variables interagissent entre elles deux à deux.

Nous constatons que les régions les plus riches en termes de PIB par habitant sont les plus heureuses, également elles ont l'espérance de vie la plus élevée (figure 1).

La croissance économique ne semble pas être la seule variable liée au bonheur (figure A1). Nous trouvons également le soutien social et la liberté. De même, ces derniers sont corrélés positivement avec la croissance économique.

Nous remarquons (figure A1) que le bonheur n'est pas lié à la perception de la générosité.

Résultats

Pour répondre à notre problématique nous avons commencé par tester l'influence de chaque variables sur le score de bonheur indépendamment les unes des autres (annexe B), puis nous avons créé un modèle linéaire multiple expliquant le bonheur selon différents facteurs (annexes C,D,E). On peut constater que le bonheur moyen d'une population semble être significativement lié avec la région du monde dans laquelle vie celle-ci. Ainsi, on note que vivre en Amérique du Nord, toutes choses égales par ailleurs, augmente le score de bonheur de 0.76 points en moyenne. Il s'agit de l'influence géographique la plus forte mais cet effet positif peut aussi se retrouver pour l'Amérique latine et les caraïbes (de l'ordre de 0.36 points) et pour l'Europe de l'Est (0.65 points). A l'inverse, le fait de vivre en Asie du sud semble diminuer le score de bonheur de 0.42 points en moyenne, toutes choses égales par ailleurs.

Avec ce modèle, on constate que le PIB d'un pays a une influence significative sur le niveau de bonheur de la population, en effet, plus ce PIB est élevé, plus le score augmente. En particulier, une augmentation d'une unité du logarithme du PIB se traduit par une augmentation de 0.33 point du score de bonheur. Les deux facteurs sociaux qui semblent le plus influencer le score de bonheur sont les indices de soutien social et de liberté. On constate qu'une augmentation d'un point sur l'indice de soutien social augmente le score de bonheur de 2.01 points, toutes choses égales par ailleurs. Pour ce qui est de l'indice de liberté, lorsqu'il augmente d'un point, le score de bonheur augmente de 2.60 points.

Discussions

D'après les sections précédentes nous avons constaté que les pays les plus heureux du monde ont le PIB par habitant le plus élevé. De même, la plupart des moins heureux sont très pauvres, cependant, cette corrélation est imparfaite. La Finlande est le pays le plus heureux du monde en 2021, avec un PIB par habitant de 48 782 $. Les États-Unis occupent la 19 -ème position au niveau du score du bonheur. Ils ont pourtant un PIB par habitant largement supérieur à celui de la Finlande. Comme indiqué par notre modèle précédemment, la croissance économique s'accompagne par une augmentation du bonheur général des citoyens. Le Costa Rica se trouve classé 16 ème au niveau du bonheur alors que son PIB par habitant est largement inférieur au pays du haut du classement. Cette corrélation entre croissance économique et bonheur est loin d'être parfaite, pour mieux comprendre cette liaison, il faut se poser la question sur la manière dont la richesse économique est dépensée et distribuée. Si nous essayons d'expliquer le score du bonheur uniquement avec le PIB par habitant, nous expliquerons que 62% de la variance de ce score tandis que si on ajoute d'autres facteurs tels que le soutien social, la liberté et l'espérance de vie nous expliquons 80% de la variance du bonheur. Par conséquent, l'augmentation du bonheur dépend d'autres facteurs qui sont également corrélés avec la croissance économique. Autrement dit, si cette dernière est dépensée sans prendre en compte le bien-être collectif, le score du bonheur n'augmentera pas, par contre, si elle est dépensée pour servir le soutien social ou la santé des individus, l'augmentation du bonheur sera plus importante.

Conclusion

En conclusion, nous pouvons dire que le bonheur dépend de plusieurs facteurs à savoir, la croissance économique, le soutien social, la liberté, l'espérance de vie en bonne santé et la région où nous vivons. En effet, l'augmentation de ces derniers se traduit par une augmentation du bonheur. Aujourd'hui, ce sont les pays les plus riches qui sont les plus heureux dans le monde (Finlande, Suisse, Luxembourg), mais d'autres pays sont heureux sans forcément être riches (Costa Rica). Pour aller plus loin, il serait intéressant de savoir s'il existe une relation circulaire entre le bonheur et les variables citées précédemment.

A

Statistiques descriptives

A.1 Matrice de corrélation

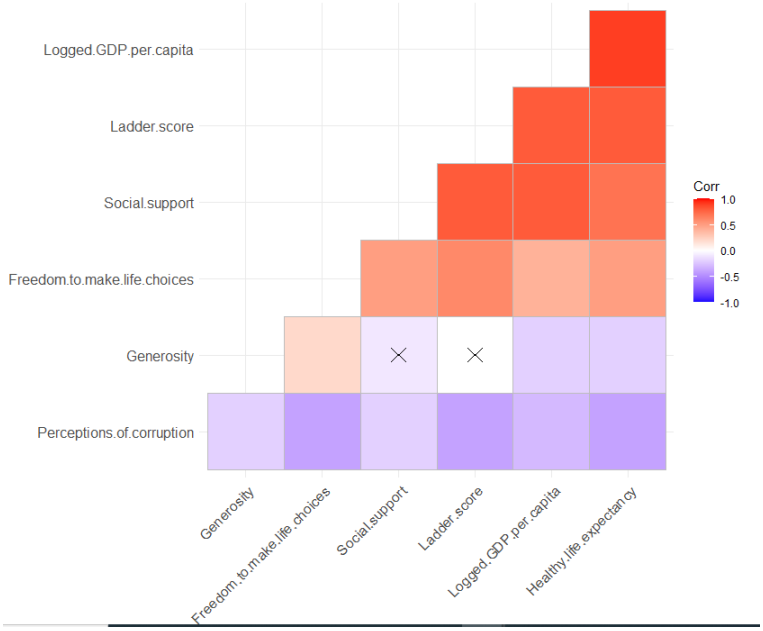

FIGURE A.1 – Matrice de corrélation

Nous constatons que le score de bonheur est fortement corrélé avec toutes les variables dont on dispose sauf la variable générosité. Les variables PIB par habitant, support social, espérance de vie, liberté de choix et perception de la corruption peuvent être des variables expliquant le score de bonheur, cependant, ces variables sont corrélées entre elles.

B

Test de lien de dépendance avec régression linéaire simple

Les variables PIB par habitant, support social, espérance de vie en bonne santé, liberté de choix et perception de la corruption ont un lien statistiquement significatif avec le score de bonheur.

Variables	R2	Pvalue de T.Student	Pvalue T.Fisher
Generosity	0.00	0.829	0.8294
Social.support	0.5729	<2e-16	2.2e-16
Logged.GDP.per.capita	0.6237	2e-16	2.2e-16
Freedom.to.make.life.choices	0.3694	2e-16	2.2e-16
Perceptions.of.corruption	0.1774	8.88e-08	8.881e-08
Regional.indicator	0.6213	.	< 2.2e-16

TABLE B.1 – Test de lien avec le bonheur par régression linéaire simple

C

Choix du modèle optimal

Tout d'abord nous avons réalisé un calcul exhaustif de tous les modèles possibles à partir des variables dont on dispose selon plusieurs critères (AIC, BIC, Cp et R2 ajusté). Également, nous avons utilisé la régression Lasso (dont le lambda est choisi en minimisant l'erreur de prévision théorique calculée par le one leave out cross validation) pour le choix de variables explicatives. Ensuite, nous avons utilisé la validation croisée (one leave out) pour estimer l'erreur de prévision théorique pour chaque modèle. Finalement, nous avons retenu le modèle ayant l'erreur de prévision la plus faible qui correspond au modèle minimisant le critère Cp de Mallow.

Critère	Variables incluses	Erreur théorique de prévision
Modèle complet	Toutes les variables	0.27722
R2 ajusté	etats_indp, latine_caraibe, a_amerique, europe_ouest, Logged.GDP.per.capita, Social.support, Freedom.to.make.life.choices	0.26761
BIC	latine_caraibe, a_amerique, europe_ouest, Logged.GDP.per.capita, Social.support, Freedom.to.make.life.choices	0.26577
AIC	latine_caraibe, a_amerique,sud_asie, sud_est_as, europe_ouest, Logged.GDP.per.capital, Social.support, Healthy.life.expectancy, Freedom.to.make.life.choices et Generosity	0.26485
CP	latine_caraibe, a_amerique, sud_asie, sud_est_asie, europe_ouest, Logged.GDP.per.capita, Social.support, Freedom.to.make.life.choices	0.26241
Lasso	latine_caraibe, a_amerique, sud_asie, sud_est_asie, europe_ouest, Logged.GDP.per.capita, Social.support, Healthy.life.expectancy,Freedom.to.make.life.choices, Perceptions.of.corruption	0.26823
Modèle Lasso	Toutes les variables (Modèle biasé)	0.25915

TABLE C.1 – Différents critères du choix du modèle optimale

D

Vérification des hypothèses du modèle linéaire

D.1 Non corrélation des résidus

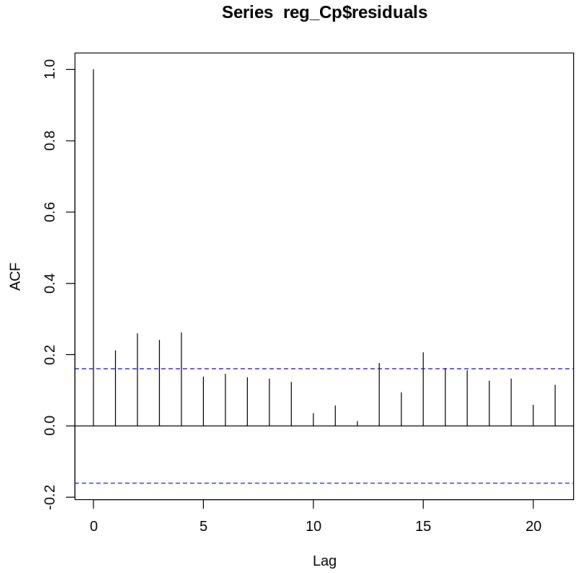

FIGURE D.1 – Autocorrélation des suites de résidus

Selon le graphique des auto-corrélations des erreurs nous remarquons l'existence de plusieurs corrélations significatives, les résidus peuvent être corrélés.

D'après le résultat du test de Durbin-Watson (DW = 1.6093) , nous rejetons l'hypothèse nulle (p-value = 0.0128) qui correspond à la non-corrélation des résidus.

D.2 Linéarité

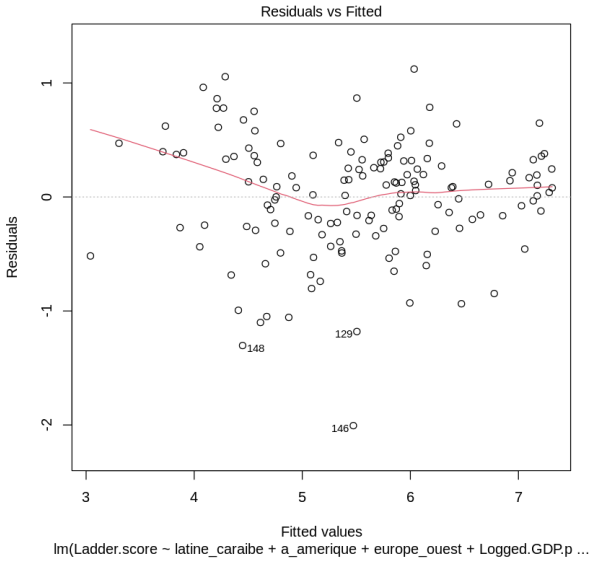

FIGURE D.2 – Graphiques des résidus en fonction des valeurs ajustées

Sur le graphique des résidus nous constatons que la courbe de régression locale n'est pas horizontale, le lien entre variable réponse et variables explicatives pourrait être non linéaire.

D'après le test de Rainbow (Rain = 3.8458) nous rejetons l'hypothèse de linéarité (p-value = 3.586e-08).

D.3 Homoscédasticité

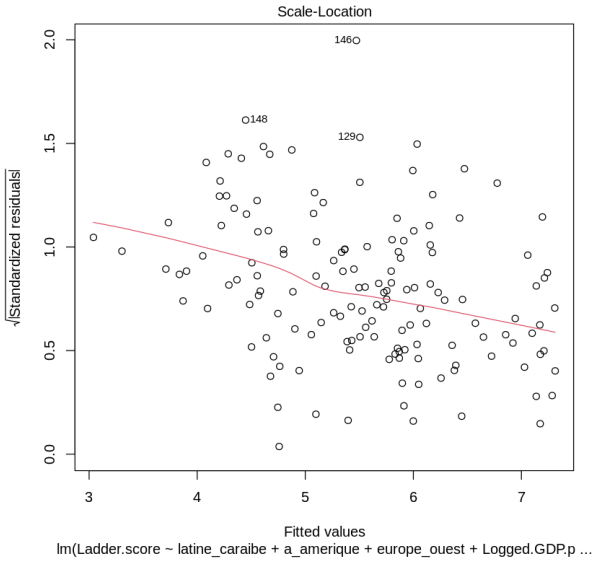

FIGURE D.3 – Graphiques des résidus standardisés en fonction des valeurs ajustées

Sur le graphique des résidus standardisé, nous ne remarquons pas de structure particulière ou une tendance. L'hypothèse d'homoscédasticité est vérifiée.

De même, d'après le test de Breusch-Pagan (BP = 12.796) nous confirmons l'hypothèse d'homoscédasticité (p-value = 0.1191).

D.4 Normalité des résidus

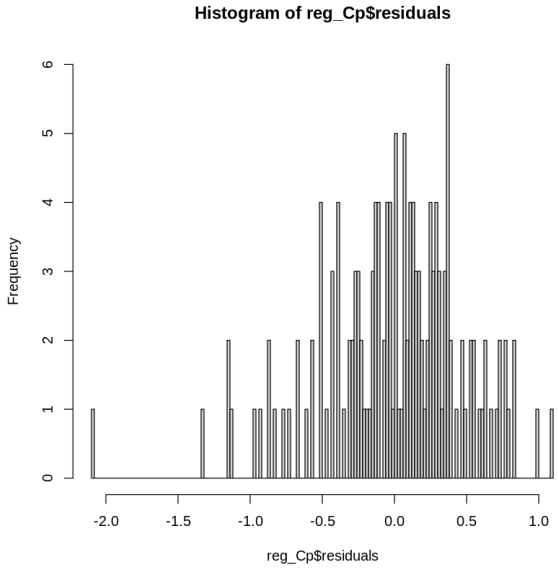

FIGURE D.4 – Histogramme des résidus

D'après l'histogramme, nous constatons que les résidus ont une distribution quasi gaussienne avec légère asymétrie.

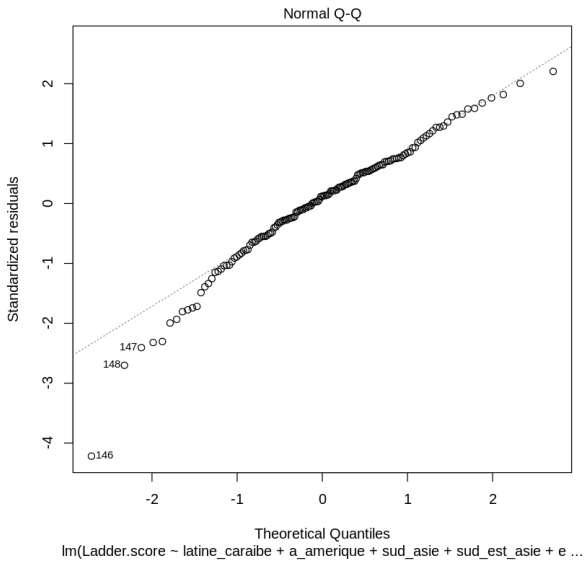

FIGURE D.5 – Graphique QQplot

Sur le qqplot nous remarquons une grande partie des résidus standardisés alignés sur

les quantiles de la loi normale, cependant, une petite partie des résidus standardisées dévie des quantiles de la loi normale.

Selon le test de Shapiro Wilk (W = 0.96544) nous rejetons l'hypothèse de la gaussianité des résidus (p-value = 0.0008422). Nous pouvons nous passer de la gaussianité des résidus vu que nous disposons d'un grand échantillon.

D.5 Outliers et points leviers extrêmes

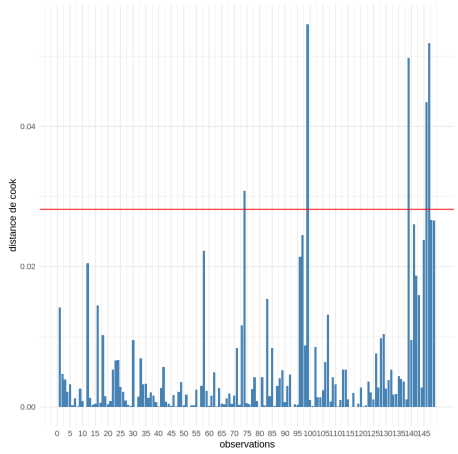

FIGURE D.6 – Graphique de la distance de Cook par observation

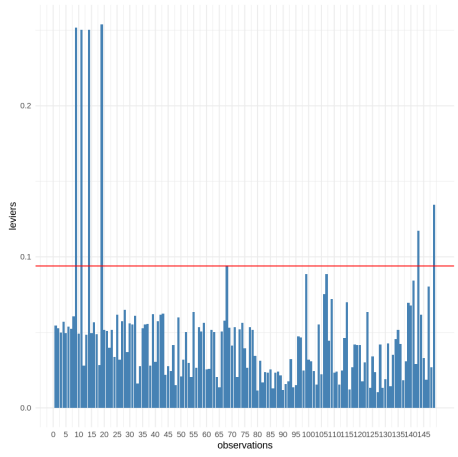

FIGURE D.7 – Graphique d'effet levier par observation

Nous remarquons la présence de quelques points influents.

E

Interprétation du modèle

E.1 Validation du modèle

```
Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                -1.49001    0.43758  -3.405 0.000863 ***
latine_caraibe              0.36434    0.12866   2.832 0.005311 **
a_amerique                  0.73276    0.26846   2.729 0.007159 **
sud_asie                   -0.42093    0.20056  -2.099 0.037632 *
sud_est_asie               -0.31598    0.18720  -1.688 0.093651 .
europe_ouest                0.65714    0.14351   4.579 1.02e-05 ***
Logged.GDP.per.capita       0.33931    0.06266   5.415 2.59e-07 ***
Social.support              2.01372    0.61320   3.284 0.001293 **
Freedom.to.make.life.choices  2.60155    0.45253   5.749 5.40e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4998 on 140 degrees of freedom
Multiple R-squared:  0.7951,    Adjusted R-squared:  0.7834
F-statistic:  67.9 on 8 and 140 DF,  p-value: < 2.2e-16
```

FIGURE E.1 – Coefficients et significativé du modèle

D'après le test de significativité globale de Fisher, nous constatons que notre modèle est globalement significatif.

D'après les tests de Student touts les coefficients de régression sont significativement non nuls sauf pour la variable « sud_est_asie »

Notre modèle explique 79.51 % de la variance de la variable réponse « ladder.score »

E.2 Interprétation

— Toutes choses égales par ailleurs, le fait d'être de l'Amérique latine ou des Caraïbes augmente le score du bonheur de 0.36 points ;
— Toutes choses égales par ailleurs, le fait d'être nord-américain augmente le score du bonheur de 0.76 points ;
— Toutes choses égales par ailleurs, le fait d'être du sud de l'Asie diminue le score du bonheur de 0.42 points ;
— Toutes choses égales par ailleurs, le fait d'être du sud-est de l'Asie n'a pas d'effet significatif ;
— Toutes choses égales par ailleurs, le fait d'être en Europe de l'Est augmente le score du bonheur de 0.65 points ;
— Toutes choses égales par ailleurs, quand le logarithme du PIB par habitant augmente (log(dollar) le score du bonheur augmente de 0.33 points ;
— Toutes choses égales par ailleurs, quand l'indice du soutien social augmente d'un point le score du bonheur augmente de 2.01 points ;
— Toutes choses égales par ailleurs, quand l'indice de liberté des choix individuels augmente d'un point le score du bonheur augmente de 2.60 points.

F

Code

Le code du projet, est disponible sur le notebook dans le lien suivant : https://colab.research.google.com/drive/1s7Ji1cBhU1oJUXuABPQ0slzNS22Mn579?usp=sharing