Régression linéaire avec R

Analyse des déterminants de l'espérance de vie en 2015

Table des matières

Listes des figures

Listes des Tableaux

INTRODUCTION

L'espérance de vie est un indicateur clé permettant d'évaluer le niveau de développement et la qualité de vie d'un pays reflétant à la fois les conditions de vie, l'accès aux soins de santé et les politiques publiques mises en place pour améliorer le bien-être de la population. Assurer une longévité accrue et une bonne santé est un enjeu fondamental pour les gouvernements, qui cherchent à réduire la mortalité prématurée et à garantir un accès équitable aux soins. Plusieurs études ont mis en évidence le rôle déterminant des dépenses de santé, du niveau de revenu, des conditions environnementales et des comportements de santé dans la variation de l'espérance de vie entre les pays (Preston, 1975 ; Cutler, Deaton & Lleras-Muney, 2006). Toutefois, ces facteurs ne se manifestent pas de manière uniforme à travers le monde.

En effet L'augmentation de l'espérance de vie dans le monde est largement attribuée aux progrès médicaux, aux améliorations des conditions de vie et aux politiques de santé publique. Cependant, de nombreux facteurs socio-économiques, environnementaux et comportementaux influencent également cet indicateur. Alors que certains pays disposent de systèmes de santé robustes et d'une protection sociale étendue, d'autres sont confrontés à des inégalités marquées en matière d'accès aux soins, de nutrition et d'infrastructures sanitaires. De plus, des éléments tels que l'éducation, la pollution, les maladies infectieuses et les modes de vie jouent un rôle clé dans la détermination de la longévité des populations (World Bank, 2019).

Cet état de choses justifie l'intérêt à mener notre recherche sur **« l'analyse des déterminants de l'espérance de vie en 2015 »**. De façon spécifique Cette étude vise à analyser les principaux déterminants de l'espérance de vie à l'échelle mondiale. Le présent document s'est donc intéressé à un ensemble de données compilées par l'Organisation mondiale de la santé (OMS) et les Nations Unies (ONU). Il couvre l'espérance de vie dans 193 pays entre 2000 et 2015, en prenant en compte plusieurs variables économiques, sanitaires et démographiques. Ce jeu de données est particulièrement utile pour des analyses statistiques visant à comprendre les déterminants de l'espérance de vie.

L'étude de ces données permettra d'examiner l'impact de divers facteurs sur l'espérance de vie.

I. Pistes d'analyse du Jeu de données :

Quelques questions d'analyse possibles avec ce jeu de données :

- Quels sont les facteurs les plus influents sur l'espérance de vie ?
- Un pays avec une espérance de vie inférieure à 65 ans devrait-il augmenter ses dépenses de santé ?
- Quel rôle jouent les habitudes de vie (régime alimentaire, exercice, tabagisme, alcool) ?
- Quelle est l'influence de la scolarisation sur l'espérance de vie ?
- Existe-t-il une corrélation entre la densité de population et l'espérance de vie ?
- Comment la couverture vaccinale influence-t-elle l'espérance de vie ?

II. Objectif:

L'objectif général de cette étude est de faire ressortir les déterminants les plus influents sur l'espérance de vie.

III. Prise en main du jeu de données

a) Contenu du jeu de données

Le jeu de données couvre 193 pays de 2000 à 2015 et comprend plusieurs indicateurs socio-économiques, sanitaires et démographiques liés à l'espérance de vie. Il a été compilé à partir des données de l'OMS et de l'ONU. Mais nous nous contenterons de faire l'étude en prenant en compte une unité temporelle (2015), la dernière année du jeu de données à notre disposition. Donc notre étude portera donc sur 183 pays. Cette sélection réduit notre échantillon à 183 pays, en raison de l'absence de données pour 10 pays sur cette période (et une absence de valeur pour notre variable dépendante qu'est l'espérance de vie).

Tableau 1: Pays causant la chute du nombre observations

Country	Year	Life Expectancy
Cook Islands	2013	NA
Dominica	2013	NA
Marshall Islands	2013	NA
Monaco	2013	NA
Nauru	2013	NA
Niue	2013	NA
Palau	2013	NA
Saint Kitts and Nevis	2013	NA
San Marino	2013	NA
Tuvalu	2013	NA

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

b) Présentation des variables

La liste des variables pouvant intervenir dans notre analyse comporte des variables quantitatives et des variables qualitatives. Nous allons en faire une présentation.

Tableau 2 : Presentation de variables

Nom (Français)	Nom (Anglais)	Description	Nature de la variable
Pays	Country	Nom du pays	Qualitative (Nominale)
Année	Year	Année d'observation	Quantitative (Discrète)
Statut	Status	Pays en développement ou développé	Qualitative (Nominale)
Espérance de vie	Life expectancy	Espérance de vie à la naissance (en années)	Quantitative (Continue)
Mortalité adulte	Adult Mortality	Taux de mortalité adulte (probabilité de mourir entre 15 et 60 ans pour 1000 habitants)	Quantitative (Continue)
Décès infantiles	Infant deaths	Nombre de décès d'enfants de moins d'un an pour 1000 habitants	Quantitative (Discrète)
Consommation d'alcool	Alcohol	Consommation d'alcool enregistrée par habitant (15 ans et plus) en litres	Quantitative (Continue)

Dépenses de santé en pourcentage	Percentage expenditure	Dépenses de santé en pourcentage du PIB par habitant	Quantitative (Continue)
Vaccination contre l'hépatite B	Hepatitis B	Pourcentage des enfants de 1 an vaccinés contre l'hépatite B	Quantitative (Continue)
Rougeole	Measles	Nombre de cas signalés de rougeole pour 1000 habitants	Quantitative (Discrète)
IMC	BMI	Indice de masse corporelle moyen de la population	Quantitative (Continue)
Décès avant 5 ans	Under-five deaths	Nombre de décès d'enfants de moins de 5 ans pour 1000 habitants	Quantitative (Discrète)
Vaccination contre la polio	Polio	Pourcentage des enfants de 1 an vaccinés contre la polio	Quantitative (Continue)
Dépenses totales de santé	Total expenditure	Dépenses totales de santé en pourcentage du PIB	Quantitative (Continue)
Vaccination contre la diphtérie	Diphtheria	Pourcentage des enfants de 1 an vaccinés contre la diphtérie	Quantitative (Continue)
VIH/SIDA	HIV/AIDS	Décès pour 1 000 naissances vivantes VIH/SIDA (0-4 ans)	Quantitative (Discrète)
PIB	GDP	Produit intérieur brut par habitant (en USD)	Quantitative (Continue)
Population	Population	Population totale du pays	Quantitative (Discrète)
Maigreur 10-19 ans	Thinness 10-19 years	Prévalence de la maigreur chez les adolescents (10-19 ans)	Quantitative (Continue)
Maigreur 5-9 ans	Thinness 5-9 years	Prévalence de la maigreur chez les enfants (5-9 ans)	Quantitative (Continue)
Composition du revenu	Income composition of resources	Indice du développement humain (IDH) mesurant la composition des revenus (de 0 à 1)	Quantitative (Continue)
Scolarisation	Schooling	Nombre moyen d'années de scolarisation	Quantitative (Continue)

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

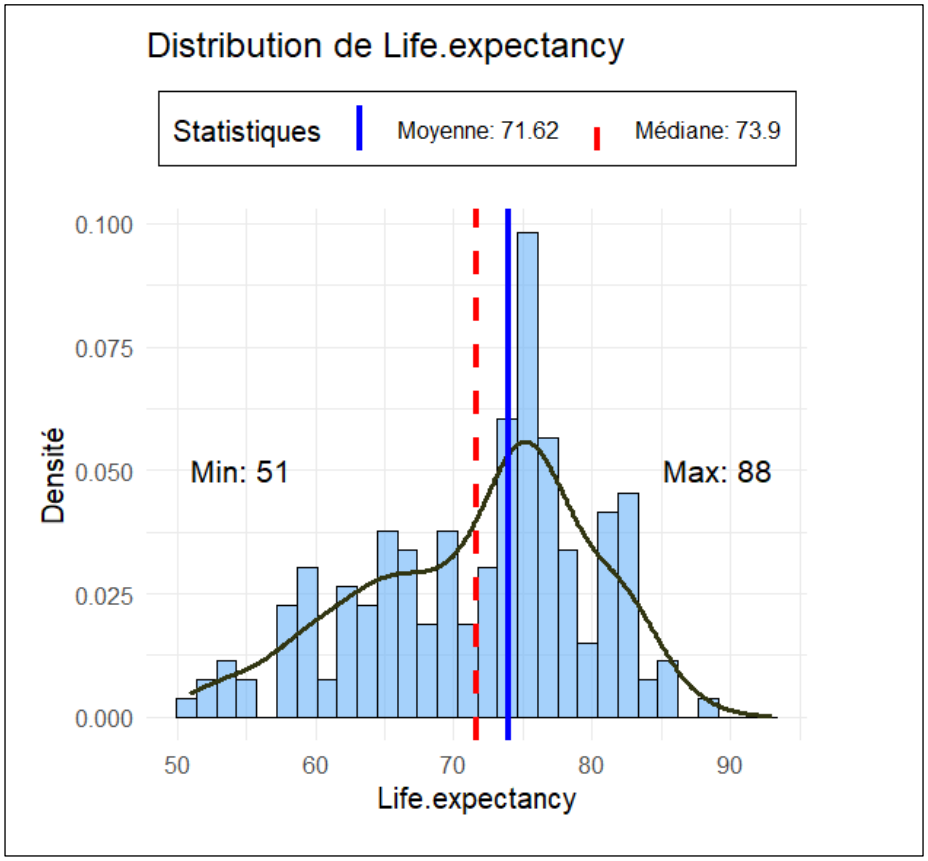
c) Analyse visuelle des variables

L'analyse descriptive des variables ne se limite pas aux statistiques numériques telles que la moyenne, la médiane ou l'écart-type. Une étape clé avant d'effectuer une régression linéaire consiste à examiner la distribution des variables à l'aide de graphiques. Cela permet d'identifier des tendances, des asymétries, des valeurs extrêmes et d'évaluer la nécessité d'une transformation des données.

- Variable dépendante (Esperance de vie)

L'image ci-dessous présente la distribution de plusieurs variables de notre jeu de données. Chaque histogramme montre la répartition des valeurs d'une variable avec une courbe de densité pour une meilleure visualisation. De plus, la moyenne et la médiane sont indiquées pour observer la symétrie des distributions.

Figure 1 : Distribution de la variable "Life Expectancy"



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Cette distribution de l'espérance de vie met en évidence une **hétérogénéité significative entre les pays.** En effet l'histogramme et la courbe de densité indiquent une distribution légèrement asymétrique à gauche. En moyenne, les gens pouvaient espérer vivre jusqu'à environ **72 ans**.

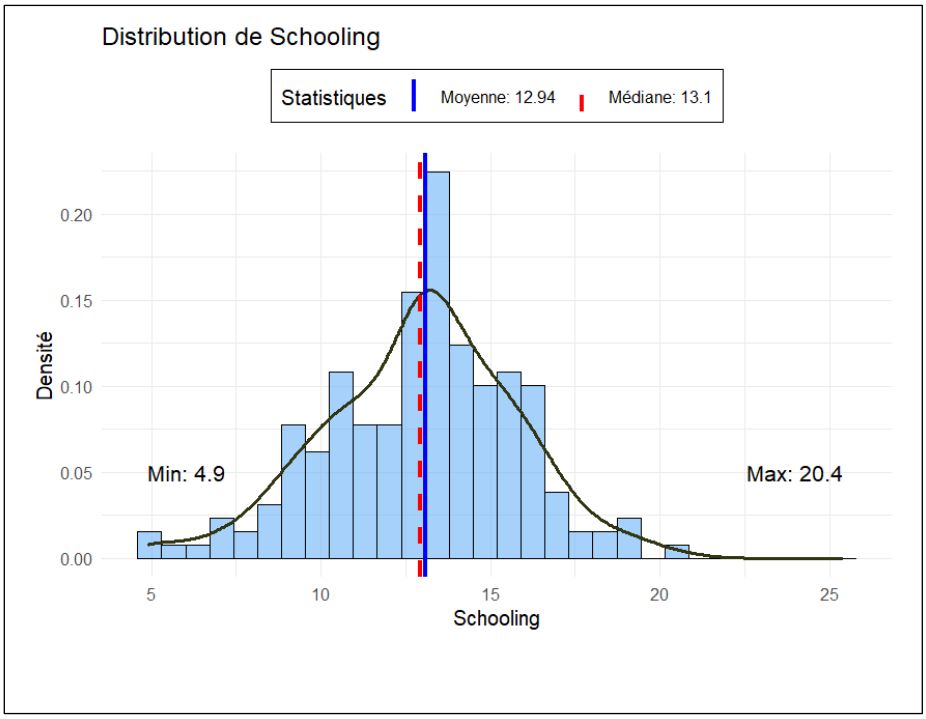
Cependant, ce chiffre variait considérablement : certains avaient la chance d'atteindre **près de 88 ans en 2015**, tandis que d'autres dépassaient à peine **les 51 ans**.

- Variables indépendantes

Nous allons en présenter une liste non exhaustive.

- **Scolarisation**

Figure 2 : Distribution de la variable "Scolarisation"

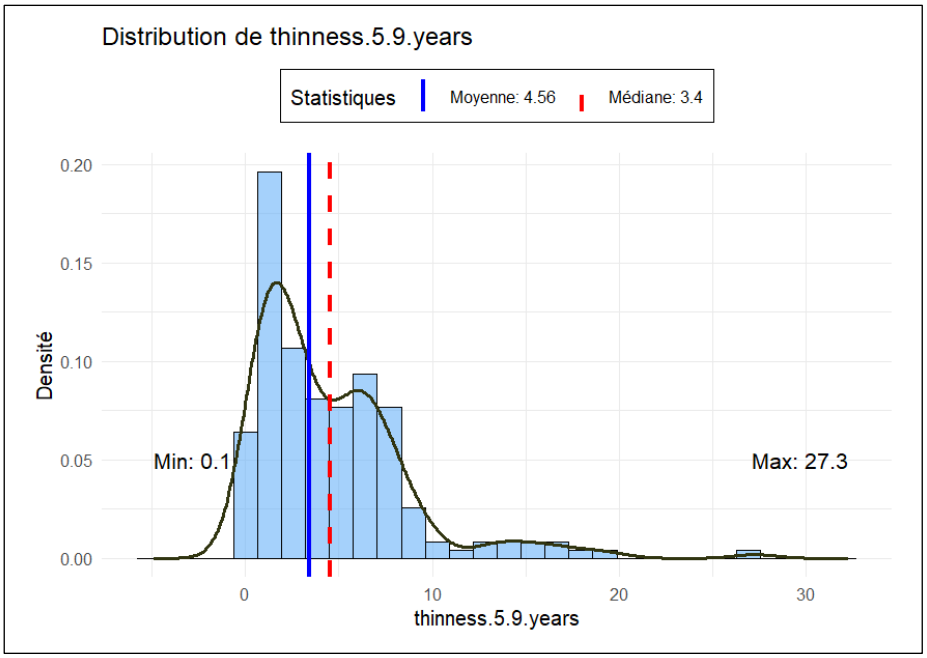


Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Cette distribution de la variable **Schooling** met en évidence une légère asymétrie vers la gauche. L'histogramme et la courbe de densité montrent une répartition des valeurs relativement proche d'une distribution normale, avec une moyenne de **12.94 ans** et une médiane légèrement supérieure (**13.1 ans**), suggérant une faible asymétrie. En moyenne, les individus ont suivi environ **13 années de scolarité**, bien que cette durée varie considérablement : certains n'ont eu que **4.9 années**, tandis que d'autres ont atteint jusqu'à **20.4 années**.

- **Maigreur 5-9 ans (Thinness 5-9 years)**

Figure 3 : Distribution de la variable "Maigreur"



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Cette distribution de la variable **thinness (5-9 years)** révèle une forte asymétrie à droite. En moyenne environ **4.56%**, des enfants de 5 à 9 ans était considérés comme maigres, bien que cette prévalence varie considérablement : certains pays enregistrent des taux extrêmement bas (**0.1%**), tandis que d'autres atteignent jusqu'à **27.3%**.

Chapitre 1 : Régression linéaire

La régression linéaire est une méthode statistique essentielle permettant de modéliser la relation entre une variable dépendante et une ou plusieurs variables indépendantes. Son objectif principal est d'identifier une fonction mathématique qui décrit au mieux cette relation, facilitant ainsi l'interprétation et la prédiction des variations de la variable étudiée.

- **Régression Linéaire Simple**

Une analyse de régression est dite **simple** lorsqu'elle vise à prédire les valeurs d'une variable dépendante (**Y**, ou variable expliquée) en fonction d'une seule variable indépendante (**X**, ou variable explicative). Ce modèle permet de résumer la relation entre les deux variables à l'aide d'une droite d'ajustement, dont l'équation générale est donnée par :

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

Où :

- **Y** : Variable dépendante (ou réponse), dont on cherche à expliquer les variations.
- **X** : Variable indépendante (ou explicative), supposée influencer Y.
- β_0 : Ordonnée à l'origine, représentant la valeur de Y lorsque X=0.
- β_1: Coefficient de régression, indiquant la variation de Y pour une unité d'augmentation de X.
- ε : Terme d'erreur, représentant les fluctuations aléatoires ou les facteurs non pris en compte par le modèle.

La régression linéaire simple repose sur plusieurs hypothèses fondamentales, telles que la normalité des résidus, l'homoscédasticité (variance constante des erreurs) et l'absence d'autocorrélation.

- **Régression Linéaire Multiple**

Lorsque la variable dépendante (**Y**) est influencée par plusieurs variables explicatives ($X_1, X_2, ..., X_p$), on parle de **régression linéaire multiple**. Ce modèle généralise la régression simple et est exprimé par l'équation suivante :

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \varepsilon$$

Où :

- $X_1, X_2, ..., Xp$: Variables explicatives contribuant à expliquer la variabilité de Y.
- $\beta_0, \beta_1, ..., \beta p$: Coefficients de régression, estimés à partir des données disponibles.

Critères d'évaluation du modèle

L'évaluation de la qualité d'un modèle de régression repose sur plusieurs indicateurs, parmi lesquels :

- **Le coefficient de détermination R^2** : Il mesure la proportion de variance de Y expliquée par le modèle. Plus R^2 est proche de 1, meilleur est l'ajustement du modèle aux données.
- **Le test de significativité des coefficients** : Il permet de vérifier si chaque variable explicative a un effet significatif sur Y à l'aide de tests statistiques comme le test de Student.
- **L'analyse des résidus** : Elle sert à examiner si les hypothèses sous-jacentes à la régression (normalité, homoscédasticité, indépendance) sont respectées.

L'intérêt de la régression multiple réside dans sa capacité à modéliser des relations complexes en tenant compte de plusieurs facteurs simultanément. Cependant, elle nécessite une attention particulière à des phénomènes comme **la multicolinéarité** (corrélation excessive entre variables explicatives) et **la sélection des variables** pertinentes afin d'éviter un surajustement du modèle.

Mettons donc en place le cadre idéal pour sa réalisation avec notre jeu de données.

A. Etude et choix des variables pertinentes

Avant de procéder à l'analyse statistique et aux modélisations, une exploration approfondie des données a été réalisée afin de comprendre leur structure et identifier d'éventuelles incohérences. Par exemple l'exploration des valeurs manquantes, en calculant le nombre de valeurs manquantes pour chaque variable et en observant leur proportion.

Tableau 3 : Analyse des valeurs manquantes

Variables	Valeurs manquantes	Valeurs manquantes (%)
Status	0	0.0
Life.expectancy	0	0.0
Adult.Mortality	0	0.0
infant.deaths	0	0.0
Alcohol	177	0.96
percentage.expenditure	0	0.0
Hepatitis.B	9	0.049
Measles	0	0.0
BMI	2	0.010
under.five.deaths	0	0.0
Polio	0	0.0
Total.expenditure	181	0.98
Diphtheria	0	0.0
HIV.AIDS	0	0.0
GDP	29	0.15
Population	41	0.22
thinness.1.19. years	2	0.01
thinness.5.9. years	2	0.01
Income.composition.of.resources	10	0.05
Schooling	10	0.05

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

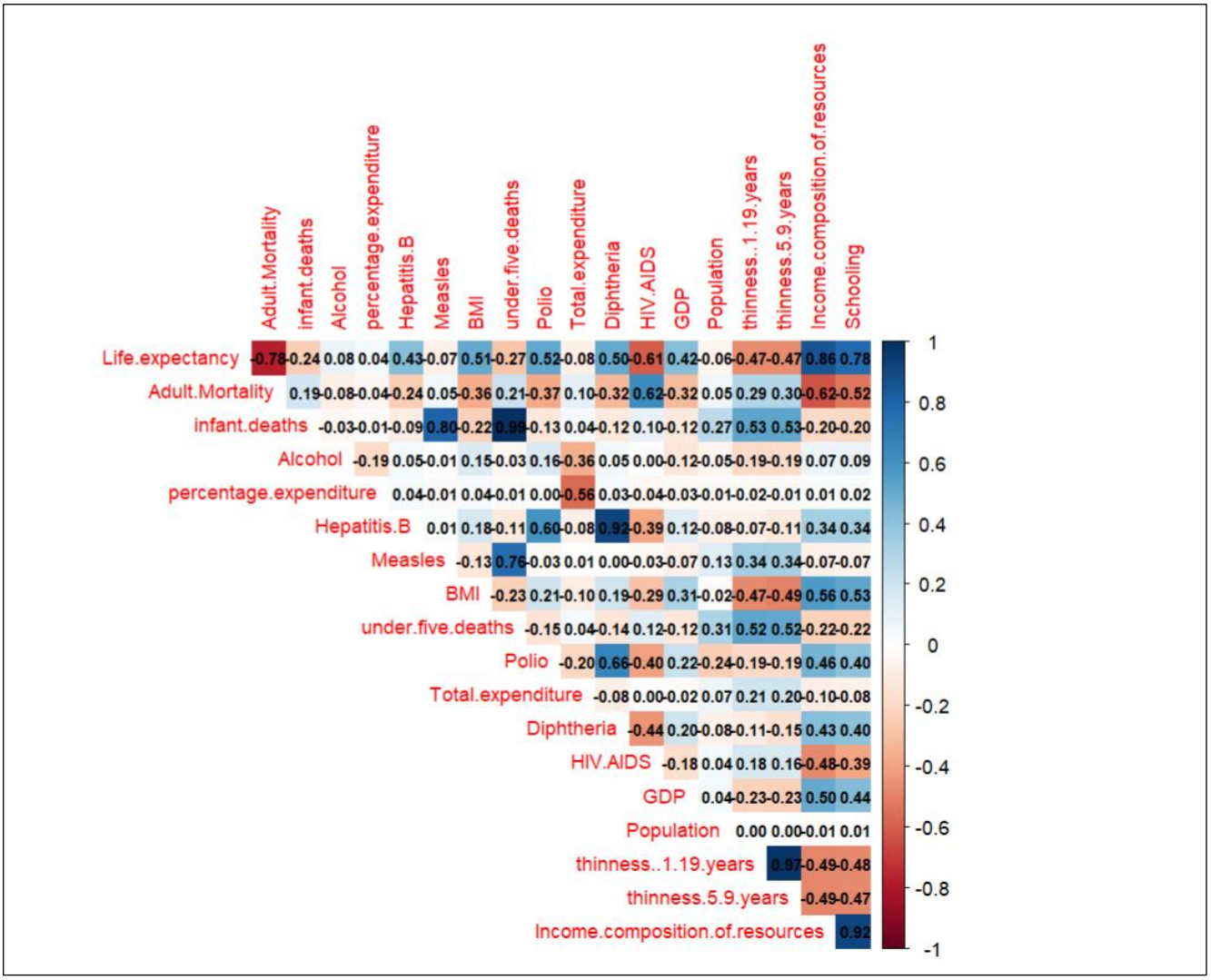
Pour traiter ces valeurs manquantes, une approche d'imputation par la médiane a été adoptée pour les variables numériques. Cette méthode a été privilégiée par rapport à la moyenne car elle est moins sensible aux valeurs aberrantes (outliers). L'existence de doublons a également été vérifiée.

De plus afin de mieux comprendre la répartition des différentes variables du jeu de données, nous avons réalisé des histogrammes et de courbes de densité dont les résultats ont été présentés plus haut. Ces visualisations permettent d'identifier la forme de la distribution, la présence éventuelle d'asymétries et de valeurs extrêmes, ainsi que les tendances générales des données.

- **Analyse de multicolinéarité**

Une matrice de corrélation a été calculée afin d'évaluer les relations entre les différentes variables numériques du jeu de données. Cette matrice permet d'identifier les variables fortement corrélées entre elles, pouvant induire un problème de multicolinéarité dans le modèle de régression.

Figure 4 : Corrplot des corrélations



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Une forte corrélation (≥ 0.7 ou ≤ -0.7) entre les variables indépendantes est un signal d'alerte. Voici les paires de variables qui présentent une forte corrélation et qui risquent de créer un problème de multicolinéarité :

- **Schooling et Income composition of resources (+0.92)** : Ces deux variables mesurent des aspects économiques et éducatifs similaires. Il serait préférable d'en garder une seule dans le modèle.
- **Measles et Infant deaths (+0.76)** : Un lien évident existe entre la rougeole et la mortalité infantile.
- **HIV/AIDS et Adult Mortality (+0.62)** : La prévalence du VIH influence la mortalité adulte, ce qui pourrait entraîner de la colinéarité.

- **Thinness 1-19 years et Thinness 5-9 years (+0.97) :** Il est logique que ces deux variables soient fortement corrélées, car elles mesurent le même phénomène (la prévalence de la maigreur) mais sur des tranches d'âge légèrement différentes. Les enfants qui souffrent de sous-nutrition à 5-9 ans sont déjà inclus dans ceux qui en souffre de 1 à 19 ans.

L'objectif étant d'identifier les variables fortement corrélées avec ***Life.expectancy*** afin de les intégrer dans le modèle de régression seules les variables présentant une corrélation en absolue supérieure ou égale à **0.55** avec elle seront retenues pour une analyse plus approfondie.

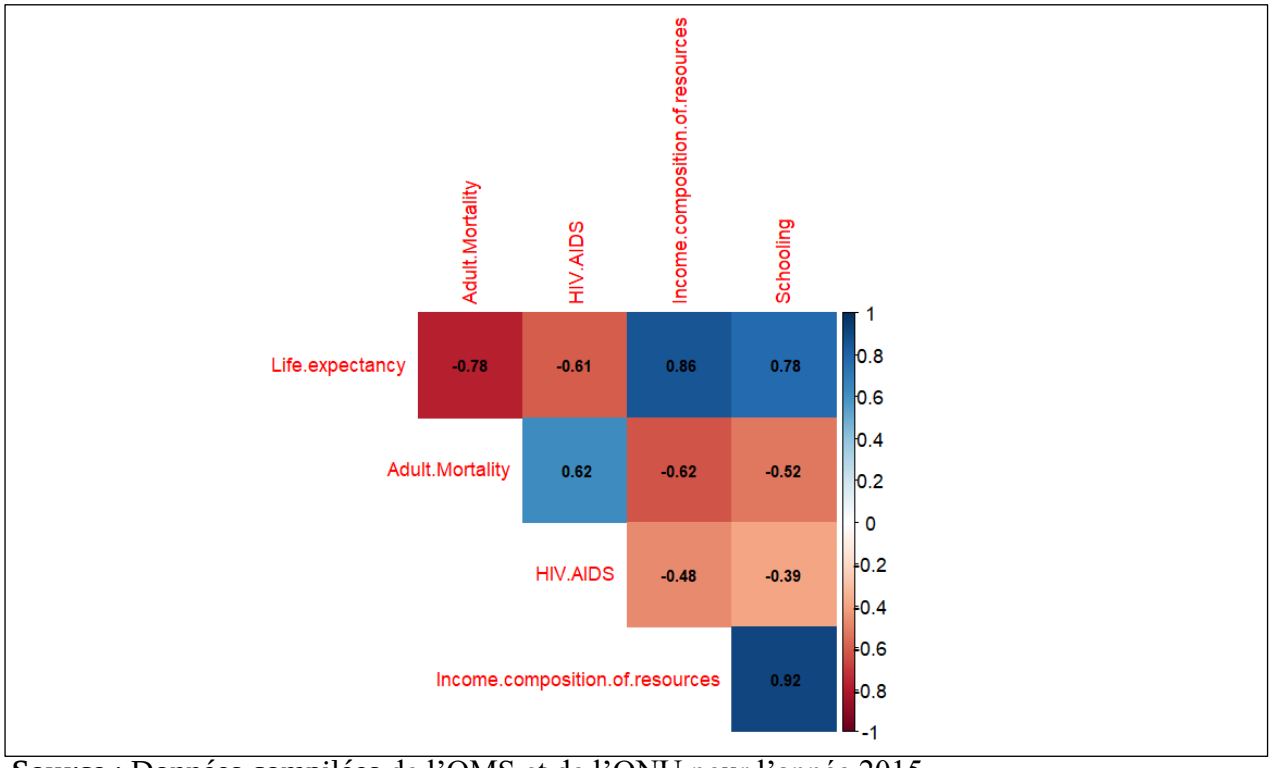
Tableau 4 : Corrélations avec Life Expectancy

Variable	Corrélation avec Life.expectancy
Adult.Mortality	-0.78
HIV.AIDS	-0.61
Income.composition.of.resources	0.86
Schooling	0.78

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Après avoir sélectionné les variables les plus pertinentes, une seconde matrice de corrélation a été réalisée uniquement avec ces variables afin d'identifier d'éventuelles corrélations trop fortes entre elles.

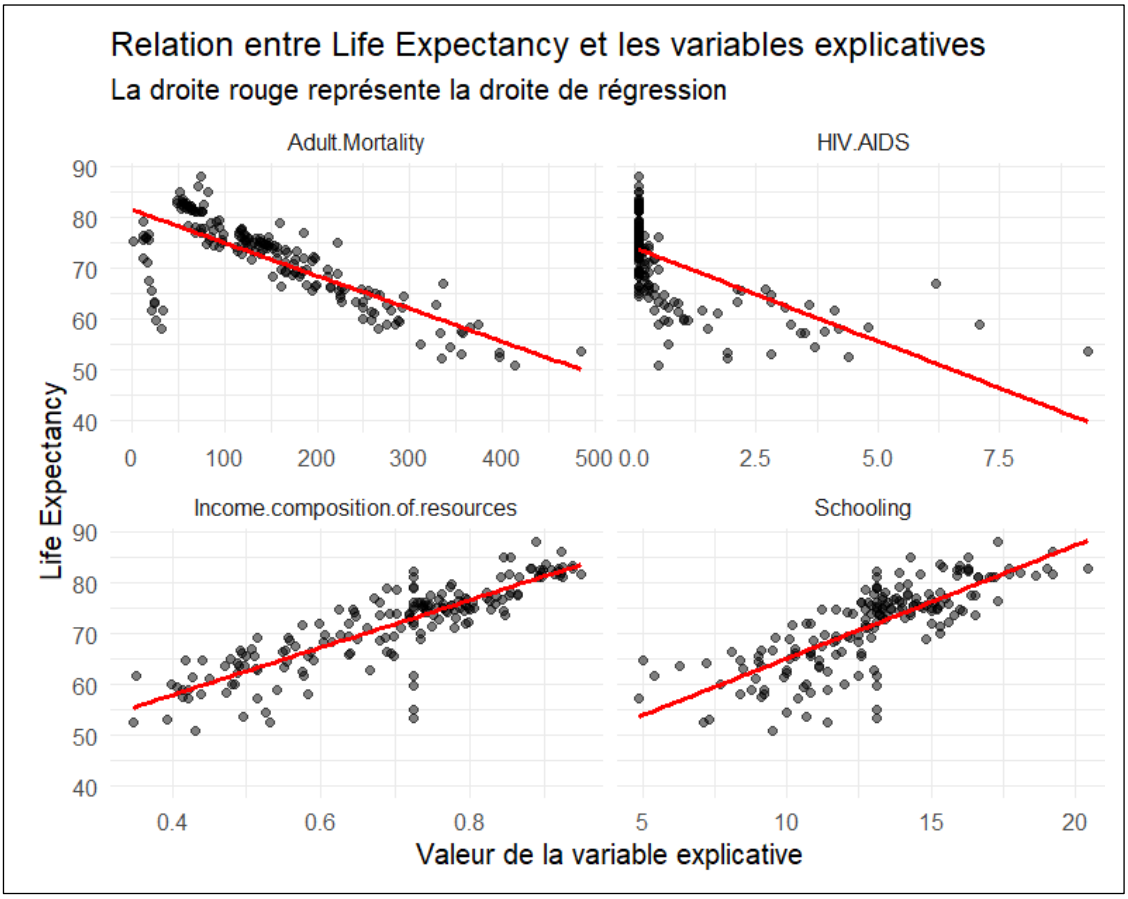
Figure 5 : Corrplot des corrélations



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Ensuite avant d'appliquer des tests de diagnostic tels que l'**indice de facteur d'inflation de variance (VIF)**, il est essentiel de vérifier si les variables indépendantes présentent une relation linéaire avec la variable dépendante en visualisant la relation pour confirmer les coefficients de corrélations.

Figure 6 : Relation entre Life et les autres variables explicatives



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Il semble y avoir une relation linéaire entre la variable dépendante et chacune des variables explicatives. Les variables **Adult Mortality** et **HIV/AIDS** exercent une influence négative sur l'espérance de vie, tandis que **Income Composition of Resources** et **Schooling** ont une influence positive. Nous pouvons maintenant confirmer la multicolinéarité en utilisant l'indice de facteur d'inflation de variance (*Variance Inflation Factor - VIF*). Il a été calculé afin de nous aider à mieux percevoir le degré de colinéarité. Les résultats du VIF indiquent que :

Tableau 5 : VIF pour différentes combinaisons de variables

Variables indépendantes	VIF - Variable 1	VIF - Variable 2
Income.composition.of.resources + Schooling	6.288123	6.288123
Income.composition.of.resources + HIV.AIDS	1.292921	1.292921
Income.composition.of.resources + Adult.Mortality	1.638571	1.638571
Schooling + HIV.AIDS	1.181394	1.181394
Schooling + Adult.Mortality	1.372316	1.372316
HIV.AIDS + Adult.Mortality	1.627751	1.627751

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

La colinéarité la plus forte est entre « Schooling » et « Income.composition.of.resources » (VIF = 6.29). Les autres paires de variables ont des VIF inférieurs à 2, indiquant une faible colinéarité.

Remarque : La variable *« Income Composition of Resources »* est un indicateur composite qui mesure le niveau de développement humain d'un pays en fonction des ressources disponibles pour améliorer la qualité de vie des citoyens. Elle est calculée à partir de plusieurs facteurs économiques et sociaux parmi lesquels l'accès aux services de base : Santé, éducation, logement, etc. La décision a été prise de conserver la variable « Schooling »

- **Justification du choix de la variable explicative "Schooling"**

Lors de l'analyse des relations entre l'espérance de vie et les variables explicatives, nous avons constaté une forte colinéarité entre **Income Composition of Resources** et **Schooling** (VIF = 6.288). Cette redondance indique que ces deux variables capturent des informations similaires.

Afin d'éviter les problèmes de multicolinéarité et d'améliorer l'interprétation du modèle, nous avons choisi de conserver **"Schooling" et d'exclure "Income Composition of Resources"**.

Ce choix est motivé par plusieurs raisons :

1. **Une variable plus spécifique et compréhensible** :
    - "Schooling" représente directement le nombre moyen d'années de scolarisation dans un pays est facile à interpréter.
2. **Un indicateur clé du développement humain** :
    - Les individus instruits sont plus susceptibles d'adopter des comportements favorables à la santé (ex. : vaccinations, alimentation équilibrée, prévention des maladies).
3. **Éviter la redondance avec "Income Composition of Resources"** :
    - "Income Composition of Resources" est un indicateur composite qui inclut déjà le niveau d'éducation.

De plus d'un autre côté nous avons :

Adult_Mortality qui est un indicateur de la mortalité des adultes dans une population. Une augmentation de ce taux est associée à une réduction de l'espérance de vie moyenne, car un plus grand nombre d'adultes meurent prématurément.

En d'autres termes, ce n'est pas tant que **Adult Mortality "affecte" Life Expectancy**, mais plutôt que **les deux variables capturent des dimensions différentes du même phénomène** (la longévité d'une population). Nous allons alors le retirer.

Les variables finales retenues après cette analyse sont entre autres :

- Life.expectancy
- HIV.AIDS et
- Schooling

Notre modèle initial peut s'écrire sous la forme :

$$Life.Expectancy=\beta_0+\beta_1\times HIV.AIDS+\beta_2\times Schooling+\epsilon$$

Où :

- β_0 est l'ordonnée à l'origine (constante du modèle).
- β_1 sont les coefficients de régression associés à chaque variable explicative.
- ϵ représente l'erreur résiduelle.

1. **Régression linéaire simple (RLS)**

Il sera fait ici l'analyse des résultats du modèle de régression linéaire simple, permettant d'évaluer l'effet individuel des différentes variables explicatives sur l'espérance de vie (*Life.Expectancy*). Dans un premier temps, nous présentons les résultats des corrélations entre chacune des variables dépendantes et (*Life.Expectancy*). Ensuite, nous analysons les résultats des **tests de normalité** et d'**hétéroscédasticité** afin de vérifier la validité du modèle. Enfin, nous discutons des coefficients estimés et leur signification.

Pour chaque régression les modèles s'écrivent donc :

- **Modèle 1 : Effet du VIH/SIDA**

  $$Life.Expectancy=\beta_0+\beta_1\times HIV.AIDS+\epsilon$$

- **Modèle 2 : Effet du niveau d'éducation**

  $$Life.Expectancy=\beta_0+\beta_1\times Schooling+\epsilon$$

Nous avons obtenu les résultats suivants :

Tableau 6 : Régression linéaire simple

Variable indépendante	R^2	p-value	Shapiro-Wilk (Normalité)	Breusch-Pagan (Hétéroscédasticité)
Schooling	0,60	< 2.2e-16	0,0000623	0,0519
HIV.AIDS	0,37	< 2.2e-16	0,1161	0,0048

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Nous procédons ici à la vérification et à l'analyse de l'hypothèse d'homoscédasticité des résidus de nos modèles de régression linéaire simple et d'autres part au test de normalité des erreurs. À cet effet, le test de **Breusch-Pagan** a été utilisé pour déterminer si la variance des erreurs est constante à travers toutes les observations, c'est-à-dire si l'hypothèse d'homoscédasticité est respectée et d'un autre côté le test de **Shapiro-Wilk** pour la vérification de l'hypothèse de normalité des résidus.

Les résultats obtenus indiquent que la probabilité associée à la statistique du test de Breusch-Pagan est significative au seuil de 5 % pour les modèles estimés avec **HIV.AIDS** (p = 0.0048) et suggérant ainsi la présence d'hétéroscédasticité. En revanche, pour le modèle utilisant **Schooling**, la valeur p est de 0.0519, ce qui indique une absence d'hétéroscédasticité au seuil de 5 %, mais une possible présence au seuil de 10 %.

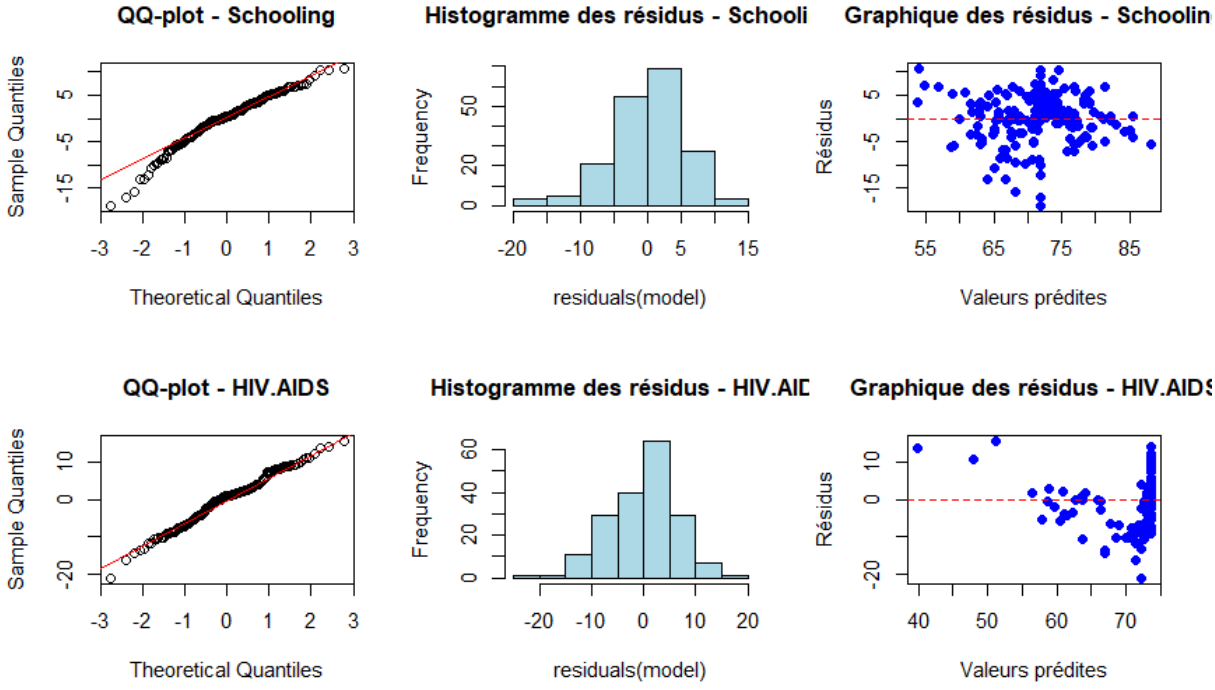
Face à ces résultats, il convient d'envisager des corrections pour atténuer l'impact de l'hétéroscédasticité. Une approche classique consiste à recourir aux **Moindres Carrés**

Généralisés (MCG), afin d'obtenir des estimations plus robustes et non biaisées des coefficients du modèle.

D'autres part Les résultats montrent que la normalité des résidus est rejetée pour les modèles utilisant **Schooling** (p = 6.23e-05), indiquant une distribution des résidus significativement différente d'une loi normale. En revanche, pour le modèle avec **HIV.AIDS**, la valeur p est de 0.1161, ce qui ne permet pas de rejeter l'hypothèse de normalité au seuil de 5 %. Lorsque la normalité des résidus n'est pas respectée pour corriger ce problème on peut faire une **transformation des variables** : L'application d'une transformation logarithmique, racine carrée ou Box-Cox sur la variable dépendante peut aider à rapprocher la distribution des résidus d'une loi normale.

Le graphique suivant confirme nos résultats par rapport aux différents tests :

Figure 7: Vérification graphique des hypothèses



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Problème de Normalité

Dans notre analyse, nous testerons plusieurs transformations « logarithmique », « racine carrée » et « inverse » sur les variables indépendantes afin d'évaluer leur impact sur la correction des biais dans les hypothèses. Ensuite, nous appliquerons deux transformations, Box-Cox et Yeo-Johnson, à la variable dépendante afin de déterminer laquelle permet de mieux satisfaire les hypothèses du modèle. Si nécessaire, nous sélectionnerons la transformation la plus appropriée en fonction des résultats obtenus.

Tableau 7 : Transformations pour la variable Schooling

Transformation	R^2	Shapiro-Wilk (p)	Breusch-Pagan (p)	Hétéroscédasticité corrigée ?	Normalité améliorée ?
Logarithmique	0,56	5,80E-05	0,0006	Non	Non
Racine	0,58	6,66E-05	0,008	Non	Non
Inverse	0,46	0,0001	8,51E-07	Non	Non

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Nous avons testé trois transformations (logarithmique, racine carrée et inverse) sur la variable Schooling qui respectais l'hypothèse d'homogénéité. Les tests n'ont pas été concluant. Aucune transformation ne nous permet d'obtenir les résultats souhaités.

Tableau 8 : Transformations pour la variable HIV.AIDS

Transformation	R^2	Shapiro-Wilk (p)	Breusch-Pagan (p)	Hétéroscédasticité corrigée?	Normalité améliorée?
Logarithmique	0,63	0,33	0,5728	Oui	Oui
Racine carrée	0,5164	0,2069	0,0467	Non	Oui
Inverse	0,5989	0,2234	0,5021	Oui	Oui

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

On peut observer que toutes les transformations donnent des p-valeurs supérieures à 0,05 suggérant que les résidus suivent approximativement une distribution normale.

La transformation logarithmique est la meilleure suivie de l'inverse. La transformation logarithmique est la plus performante, avec un R^2 le plus élevé (0,63), des résidus normalement distribués et une absence d'hétéroscédasticité.

- **Analyse des résultats après transformation Boxcox**

La transformation de Box-Cox dépend d'un paramètre λ (lambda) qui ajuste la forme de la distribution des données. Pour chaque variable explicative, nous avons estimé λ et nous avons retenu la valeur de λ la plus optimale :

- **Schooling** : λ = **2**
- **HIV.AIDS** : λ = **1.51**

Tableau 9: Lambda

λ trouvé	Transformation recommandée
λ = 2	Y^2 (carré de la variable)
λ = 1	Aucune transformation (modèle linéaire classique)
λ = 0.5	√Y (racine carrée)
λ = 0	log(Y) (logarithme naturel)
λ < 0	1/Y (inverse) est utilisé

Source : Par les auteurs à travers les recherches

Tableau 10 : Régression linéaire simple Boxcox

Régression	R^2	Shapiro-Wilk (p)	Breusch-Pagan (p)	Hétéroscédasticité corrigée?	Normalité améliorée?
Schooling	0,61	0,0099	0,1698	Oui	Non
HIV.AIDS	0,3581	0,1851	0,0111	Non	Oui

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Remarque : Cette transformation ne corrige toujours pas le problème des hypothèses. Et vue la valeur du lambda la transformation log et racine carrée sur Life.Expectancy n'auront aussi aucun effet.

- **Analyse des résultats après transformation Yeo-Johnson**

Tableau 11: Yeo-Johnson

Régression	R^2	Shapiro-Wilk (p)	Breusch-Pagan (p)	Hétéroscédasticité corrigée?	Normalité améliorée?
Schooling	0,6237	0,20	0,43	Oui	Oui
HIV.AIDS	0,3251	0,08	0,10	Oui	Oui

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Pour "Schooling" et HIV_AIDS, la transformation Yeo-Johnson a bien corrigé la normalité et l'hétéroscédasticité.

- **Transformations finales retenues pour chaque modèle**

L'analyse des résultats montre que la **transformation logarithmique** est la plus adaptée pour la variable HIV.AIDS elle offre le meilleur **R^2**, elle améliore la normalité des résidus avec un **test de Shapiro-Wilk non significatif (0.33)** et corrige l'hétéroscédasticité (0.57), garantissant la stabilité des erreurs.

De même dans la première régression simple l'analyse des résultats montre que la **transformation Yeo-Johnson** sur la variable dépendante est la plus adaptée pour la variable **Schooling.**

Résultat de l'estimation des modèles RLS

➢ Régression linéaire simple avec Schooling

Tableau 12: Résultats régression linéaire 1

Variables indépendantes	Coefficient	Erreur type	t value	Probabilité
Constante	-5458.2	4632.4	-1.178	0.240
Schooling	6059.5***	349.8	17.320	<2e-16
F	300			
p-value	<2.2e-16			
R-squared	0.6237			
Adjusted R-squared	0.6216			

*** Significatif à 1% (P<0,01) ; ** Significatif à 5% (P<0,05) ; * Significatif à 10% (P<0,10)

Les coefficients estimés dans ce modèle ne peuvent pas être interprétés directement en raison de l'application de la transformation de Yeo-Johnson sur la variable dépendante. Cette transformation modifie l'échelle d'origine des données afin d'améliorer la normalité des résidus et la stabilité de la variance. Pour obtenir une interprétation en unités réelles, il est nécessaire d'appliquer la transformation inverse aux prédictions et aux coefficients estimés. Cela permet de revenir à l'échelle d'origine et d'exprimer les effets des variables explicatives de manière plus réel. Après transformation « Une année supplémentaire de scolarisation est associée à une augmentation moyenne de **2.14 années** d'espérance de vie, après avoir corrigé l'effet de la transformation Yeo-Johnson. »

Détransformation :

$$T(y) = \frac{(y+1)^{\lambda} - 1}{\lambda}$$

Multiplions par λ des deux côtés :

$$\lambda T(y) + 1 = (y+1)^{\lambda}$$

Prenons la racine λ-ième :

$$y + 1 = (\lambda T(y) + 1)^{\frac{1}{\lambda}}$$

Donc :

$$y = (\lambda T(y) + 1)^{\frac{1}{\lambda}} - 1$$

L'effet marginal recherché et donc le coefficient estimé est donné par la dérivée de **Life.expectancy** par rapport à la variable dépendante (X) :

En remplaçant T(Y) par sa forme normal ($\beta_0 + \beta_1 X + \varepsilon$) on obtient :

$$\frac{dy}{dX} = \frac{d}{dX}\left((\lambda(\beta_0 + \beta_1 X + \varepsilon) + 1)^{\frac{1}{\lambda}} - 1\right)$$

La dérivée d'une fonction puissance u^p est :

$$\frac{d}{dx}u^p = pu^{p-1} \cdot \frac{du}{dx}$$

Avec :

- $u = \lambda(\beta_0 + \beta_1 X + \varepsilon) + 1$
- $p = \frac{1}{\lambda}$

On applique la règle :

$$\frac{dy}{dX} = \frac{1}{\lambda}(\lambda(\beta_0 + \beta_1 X + \varepsilon) + 1)^{\frac{1}{\lambda}-1} \times \lambda\beta_1$$

Simplifions :

$$\frac{dy}{dX} = (\lambda(\beta_0 + \beta_1 X + \varepsilon) + 1)^{\frac{1}{\lambda}-1} \times \beta_1$$

Tableau 13 : Résultats régression linéaire 2

Variables indépendantes	Coefficient	Erreur type	t value	Probabilité (p-value)
Constante	63,88	0,57	111,67	< 2e-16 ***
HIV_AIDS_log	-5,15	0,29	-17,59	< 2e-16 ***
F-statistique	309,4			
p-value	< 2.2e-16			
R-squared	0,630			
Adjusted R-squared	0,628			

*** Significatif à 1% (P<0,01) ; ** Significatif à 5% (P<0,05) ; * Significatif à 10% (P<0,10)

Dans ce modèle, la variable **HIV.AIDS** a été transformée en utilisant « log » avant d'être utilisée dans la régression. Par conséquent, le coefficient « -5,15 » ne peut pas être interprété directement comme une variation linéaire.

Notre modèle actuel est sous la forme :

$$Life.Expectancy = 63{,}88 - 5{,}15 \times log(HIV.AIDS)$$

L'effet marginal et donc le coeefficient estimé est donné par la dérivée de **Life.expectancy** par rapport à **HIV.AIDS** :

$$\frac{d(Life.expectancy)}{d(HIV.AIDS)} = \beta_1 \times \frac{1}{HIV.AIDS}$$

Pour des valeurs de HIV l'effet marginal est de -51 quand HIV_AIDS est très faible cela signifie qu'une augmentation d'une unité (1 décès supplémentaire pour 1 000 naissances vivantes) est associée à une baisse très forte de l'espérance de vie.
Pour des valeurs de HIV l'effet marginal passe à -5, cela signifie qu'une augmentation supplémentaire d'un décès pour 1 000 naissances a toujours un impact négatif sur l'espérance de vie.
Tout ceci se traduit par le fait que dans les pays où HIV_AIDS est faible (très peu de décès infantiles liés au VIH/SIDA), une petite augmentation peut indiquer une détérioration rapide des conditions sanitaires, ce qui impacte fortement l'espérance de vie.
Tandis que dans les pays où HIV_AIDS est déjà très élevé, une augmentation supplémentaire a un effet moindre, car la mortalité infantile est déjà structurellement élevée, donc l'impact relatif est plus faible.

2. Régression linéaire multiple (RLM)

La régression linéaire multiple est une extension de la régression linéaire simple qui permet de modéliser la relation entre une variable dépendante (ou réponse) et plusieurs variables indépendantes (ou explicatives). Contrairement à la régression linéaire simple qui utilise une seule variable explicative, la RLM permet d'évaluer l'effet simultané de plusieurs prédicteurs sur la variable d'intérêt tout en contrôlant l'influence des autres facteurs.

- **Écriture du modèle**

Dans notre cas, nous souhaitons analyser l'impact de la scolarisation (Schooling) et du taux de mortalité dû au VIH/SIDA (HIV.AIDS) sur l'espérance de vie (Life.expectancy). Le modèle de régression linéaire multiple s'écrit sous la forme :

$$Life.expectancy = \beta_0 + \beta_1 \times Schooling + \beta_2 \times HIV.AIDS + \varepsilon$$

Ce modèle nous permettra de quantifier comment l'éducation et l'impact du VIH/SIDA influencent l'espérance de vie, tout en tenant compte des effets conjoints des deux variables explicatives.

- **Statistiques générales**

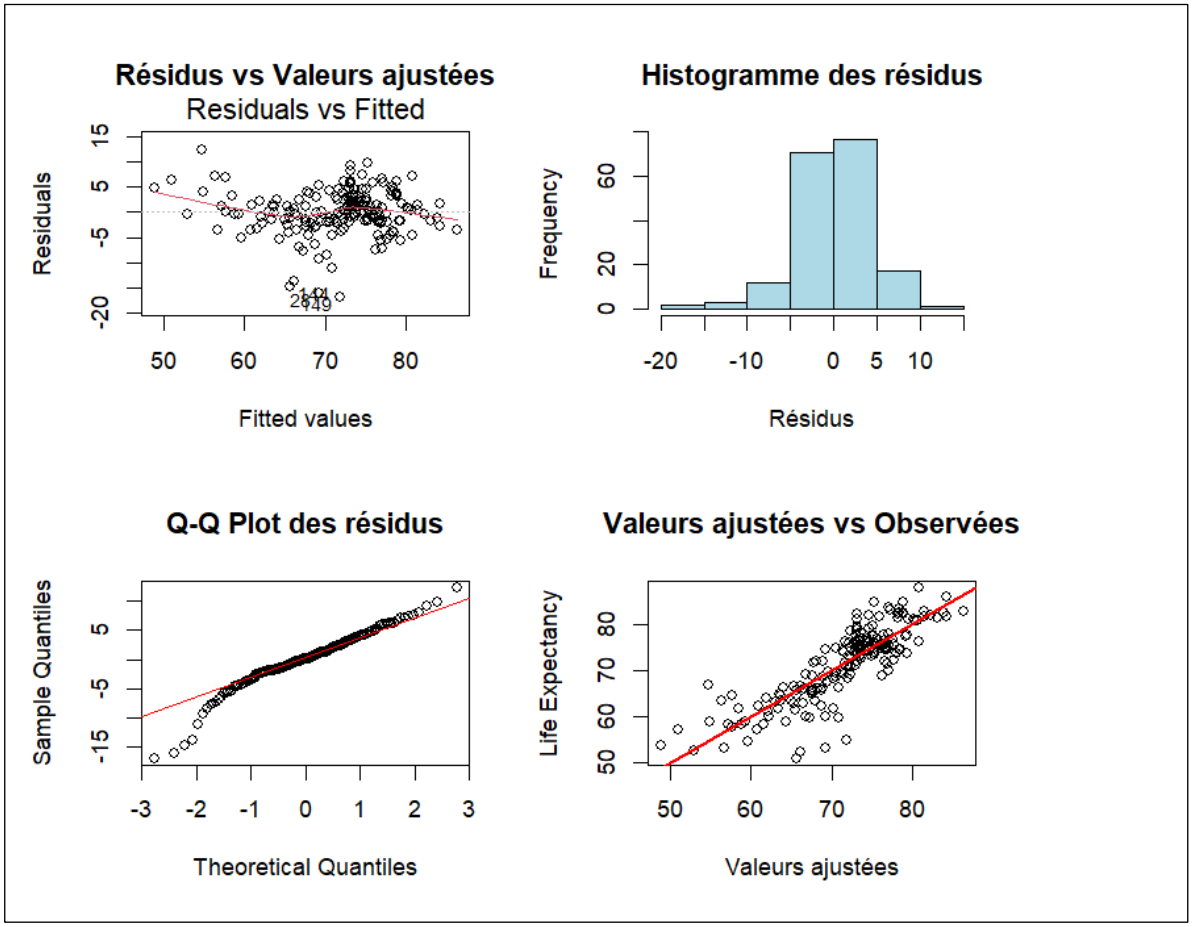
Tableau 14: Statistiques

Modèle	**Life.expectancy ~ Schooling + HIV.AIDS**
Nombre d'observations	183 (182 degrés de liberté)
R^2 multiple	0.7115
R^2 ajusté	0.7083
F-statistique	222, p-value < 2.2e-16
Erreur standard des résidus	4.387

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

- **Analyse des résidus**

Figure 8 : Tests hypothèses du modèle



Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Nous pouvons également le confirmer à travers les tests ci-dessous.

Tableau 15 : Tests de confirmation

Test	Statistique	p-value	Interprétation
Shapiro-Wilk	W = 0.95413	1.176e-05	Résidus non normaux
Breusch-Pagan	BP = 3.556	0.169	Pas d'hétéroscédasticité

Source : Données compilés de l'OMS et de l'ONU pour l'année 2015

La p-value très faible (< 0.05) indique un rejet de H_0. Les résidus ne suivent pas une distribution normale. Il sera nécessaire d'effectuer une transformation des variables. De plus la p-value étant supérieure à 0.05 pour le test de Breusch Pagan, nous ne rejetons pas H_0. L'hypothèse d'homoscédasticité est donc respectée.

- **Correction de la non-normalité des résidus**

Plusieurs transformations ont été effectué afin de pouvoir valider cette hypothèse. Les résultats suivants ont été obtenus :

Tableau 16 : Résultats des différentes transformations

Modèle	R^2 ajusté	Erreur résiduelle	Normalité des résidus (Shapiro-Wilk, p)	Hétéroscédasticité (Breusch-Pagan, p)
Log(Schooling) + Log(HIV)	0.7577	3.998	0.0228	0.039
Log(Schooling) + Sqrt(HIV)	0.7228	4.277	0.00023	0.0228
Log(Schooling) + Inverse(HIV)	0.7427	4.121	0.5607	0.0105
Sqrt(Schooling) + Log(HIV)	0.7711	3.887	0.0121	0.0587
Sqrt(Schooling) + Sqrt(HIV)	0.7396	4.145	0.00014	0.0495
Sqrt(Schooling) + Inverse(HIV)	0.7541	4.029	0.3616	0.0104
Inverse(Schooling) + Log(HIV)	0.7203	4.296	0.1422	0.0073
Inverse(Schooling) + Sqrt(HIV)	0.6714	4.657	0.006	0.0016
Inverse(Schooling) + Inverse(HIV)	0.7119	4.36	0.7668	0.0057

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Aucune des transformations n'as pu nous permettre de valider notre hypothèse. Nous passons alors aux transformations qui ont été utilisées lors de la régression linéaire simple. Soit logarithmique pour la variable HIV.AIDS et Yeo Johnson pour Life.Expectancy dans le cas de la variable schooling. Notre modèle de régression multiple s'écrira donc :

$$Life.Expectancy\_YJ = \beta_0 + \beta_1 \log(HIV\_AIDS) + \beta_2 Schooling + \varepsilon$$

Voici les résultats obtenus :

Tableau 17: Tests hypotheses

Tests	**p-value**
Shapiro-Wilk (Normalité des résidus)	0.3065795
Breusch-Pagan (Hétéroscédasticité)	0.5365632

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Tableau 18 : Regression multiple finale

Variables indépendantes	**Coefficient**	**Erreur type**	**t value**	**Probabilité (p-value)**
Constante	8833.1	3943.8	2.24	0.0263 *
HIV_AIDS_log	-7919.4	771.8	-10.26	<2e-16 ***
Schooling	4035.6	341.4	11.82	<2e-16 ***
F-statistique	289.1			
p-value	<2.2e-16			
R-squared	0.7626			
Adjusted R-squared	0.7599			

*** Significatif à 1% (p < 0.01), ** Significatif à 5% (p < 0.05), * Significatif à 10% (p < 0.10).

Source : Données compilées de l'OMS et de l'ONU pour l'année 2015

Les coefficients estimés dans ce modèle ne peuvent pas être interprétés directement en raison de l'application des différentes transformation effectués. Ses transformations modifient l'échelle d'origine des données afin d'améliorer la normalité des résidus et la stabilité de la variance. Pour obtenir une interprétation en unités réelles, il est nécessaire d'appliquer une détransformation correspondante. Elle reste la même que précédemment ici on obtient le modèle initial suivant :

$$Life.Expectancy\_YJ = ((\beta_0 + \beta_1 Schooling + \beta_2 \log(HIV.AIDS)) \times \lambda + 1)^{1/\lambda} - 1$$

Pour obtenir les coefficients correspondants on passe à la dérivée par rapport à chaque X du modèle de la même manière que précédemment. Après détransformation on constate que :

- Chaque année supplémentaire de scolarisation est associée à une augmentation moyenne de 1,57 an de l'espérance de vie.
- Chaque unité supplémentaire de VIH/SIDA est associée à une réduction moyenne de 4,67 ans de l'espérance de vie.

Conclusion

L’objectif principal de cette étude était d’identifier les déterminants les plus influents sur l’espérance de vie en 2015. Pour ce faire, nous avons d’abord présenté les différentes approches théoriques et empiriques liées à cette thématique avant d’exposer la méthodologie adoptée.

Nous avons procédé à une vérification et à une correction des hypothèses de multicolinéarité, d’hétéroscédasticité et de normalité, puis à l’estimation des modèles de régression linéaire simple et multiple. Les résultats obtenus montrent que la scolarisation a un effet positif et significatif sur l’espérance de vie, tandis que la prévalence du VIH/SIDA a un effet négatif.

L'effet marginal indique qu’une année supplémentaire de scolarisation augmente l’espérance de vie de 1,57 an en moyenne, tandis qu’une augmentation d’une unité de la prévalence du VIH/SIDA entraîne une baisse moyenne de 4,67 ans de l’espérance de vie.

Ces résultats confirment l'importance des politiques publiques en matière d'éducation et de santé pour améliorer l'espérance de vie. L’augmentation de l’accès à l’éducation pourrait être une stratégie efficace pour accroître l’espérance de vie, tout comme la réduction de l’impact des maladies transmissibles, en particulier le VIH/SIDA. Les décideurs politiques devraient donc renforcer les programmes d’éducation et de sensibilisation à la santé pour améliorer les conditions de vie et la longévité des populations.

Références bibliographiques

Cutler, D. M., Deaton, A. S., & Lleras-Muney, A. (2006). The determinants of mortality. *Journal of Economic Perspectives*, *20*(3), 97-120. https://doi.org/10.1257/jep.20.3.97

Preston, S. H. (1975). The changing relation between mortality and level of economic development. *Population Studies*, *29*(2), 231-248. https://doi.org/10.1080/00324728.1975.10410201

World Bank. (2019). World Development Indicators. *The World Bank Group*. Consulté sur https://databank.worldbank.org/source/world-development-indicators

Organisation mondiale de la santé (OMS). (2018). World Health Statistics 2018: Monitoring health for the SDGs. *Organisation mondiale de la santé*. Consulté sur https://www.who.int/gho/publications/world_health_statistics/en/