N°257
Mai 2015

STAT & CO

Lin Cécile, Sergent Margaux, Célia Salgueiro

On vous répond

Chers lecteurs, puisque le numéro du mois dernier sur l'espérance de vie vous a conquis, nous revenons sur ce sujet ce mois-ci. En effet, nous constatons que l'espérance de vie, comme vu dans notre précédent article, est de plus en plus élevée au fur et à mesure des années. Cependant, même si celle-ci évolue, il reste toujours de grandes différences d'espérance de vie selon les pays. Notre étude va ainsi tenter de démontrer quels facteurs influencent cette dernière. Pour orienter notre étude, nous avons donc choisi quatre facteurs qui peuvent expliquer la différence d'espérance de vie selon les pays. Ces facteurs sont : le VIH/SIDA, la couverture vaccinale contre la polio, la maigreur chez les enfants et adolescents et enfin la consommation d'alcool. Notre étude va donc essayer de répondre à la problématique suivante : est-ce que les facteurs internes au pays et donc plus globalement le pays où nous naissons déterminent notre espérance de vie ?

Nous avons sélectionné une variable cible qui est Y correspondant à l'espérance de vie. Le but de notre étude sera d'expliquer l'espérance de vie moyenne par pays en fonction d'autres variables choisies. Tout d'abord, nous remarquons que l'espérance de vie peut être très variable d'un pays à l'autre. En effet, la plus faible espérance de vie sur l'ensemble des pays de notre étude s'élève à 48.1 ans tandis que la plus élevée est de 89 ans. Cependant, nous constatons que 50 % des pays ont une espérance de vie supérieure ou égale à 73.9 ans.

Notre première variable explicative choisie est X1 correspondant au VIH/SIDA (nombre de décès pour 1000 habitants). De nos jours, le VIH/SIDA provoque encore des morts mais moins nombreuses. En effet, 75 % des pays ont un nombre de décès inférieur ou égal à 0.3 pour 1 000 habitants. Cependant, il y a encore des pays atteints plus fortement par cette maladie étant donné que le pays le plus touché à un ratio de 7.3 morts pour 1000 habitants.

Notre deuxième variable explicative choisie est X2 la consommation d'alcool par habitant en litre d'alcool pur. Encore une fois, la consommation d'alcool peut être très variable selon les pays. En effet, 25% des pays ont une consommation d'alcool inférieure ou égale à 0.01 litre d'alcool pur tandis que 75% des pays ont une consommation d'alcool inférieure ou égale à 6.58 litres d'alcool pur. Et le pays ayant la plus grosse consommation d'alcool s'élève à 15.19 litres.

Notre troisième variable explicative est X3 le taux de vaccination contre la polio chez les enfants de 1 an en pourcentage. La plupart des pays ont une couverture vaccinale contre la polio chez les enfants de 1 an très importante. En effet, 25% des pays ont un pourcentage de vaccination inférieur ou égal à 79% et 75% des pays ont un taux de vaccination inférieur ou égal à 97%. Cependant, le pays avec la plus faible couverture vaccinale s'élève à 8 % ce qui montre encore des fortes inégalités entre les pays.

Et finalement, notre dernière variable explicative X4 correspond au taux de maigreur chez les enfants et adolescents de 10 à 19 ans en pourcentage. Il y a de grandes disparités entre les pays concernant ce taux. En effet, le pays avec le plus faible taux de maigreur s'élève à 0.1 % tandis que le taux le plus élevé est de 26.8%. Avec nos données, on voit également que 75% des pays ont un taux inférieur ou égal à 6.6%

Notre étude portera donc sur un échantillon de 161 pays avec Y notre cible et X1, X2, X3 et X4 nos variables explicatives. On pose un modèle théorique qui est : $Y=a_01+a_1X1+a_2X2+a_3+X3+a_4X4+u$ avec les hypothèses sur les erreurs qui doivent être :

- Centrées : $\forall_{i=1,..,n}\ E(u_i)=0$
- Décorrélées : $\forall_{i\neq j}\ E(u_iu_j)=0$
- Homoscédastiques : $\forall_{i=1,..,n}\ Var(u_i)= \sigma^2$

Par une régression linéaire, on obtient le modèle calculé suivant :
$\hat{Y}=67.3738\times1-3.0999X1+0.6282X2+0.0782X3-0.5103X4+\hat{u}$
où $\hat{u}$ est appelé résidu et qui vérifient les mêmes conditions que celles sur les erreurs.

La performance du modèle utilisé est montrée par le R^2. Le modèle calculé choisi avec les variables X1, X2, X3 et X4 explique à 62% Y, l'âge moyen d'espérance de vie des pays. Intéressons-nous plus globalement au modèle et à sa pertinence. Nous allons donc réaliser un test de significativité globale de la régression. Pour cela, nous réalisons le test de Fisher qui suit le problème suivant :

> H_0 : Le modèle n'est pas pertinent pour expliquer Y
>
> H_1 : Le modèle est pertinent pour expliquer Y

On choisit ici un $\alpha = 1\%$ comme risque d'erreur. Si on obtient une p-value plus petite que notre risque d'erreur alors on prend H_1. À l'inverse, si la p-value est supérieure au risque d'erreur, on choisit H_0. En effectuant le test, on obtient une p-value très petite, très proche de 0. On a donc une p-value inférieure à notre risque d'erreur que nous nous sommes fixées, nous décidons alors de choisir H_1 et d'accepter la pertinence du modèle choisi pour expliquer l'espérance de vie.

Ensuite, nous allons également nous intéresser à la pertinence de chaque régresseur pour expliquer Y. Nous allons donc réaliser un test de significativité locale des régresseurs. Pour cela, nous réalisons le test de Student qui suit le problème suivant :

H_0 : $X_{i=1,\dots,4}$ n'explique pas Y

H_1 : $X_{i=1,\dots,4}$ explique Y

On choisit ici un $\alpha = 1\%$ comme risque d'erreur. De nouveau, si on obtient une p-value plus petite que notre risque d'erreur alors on prend H_1. À l'inverse, si la p-value est supérieure au risque d'erreur, on choisit H_0.

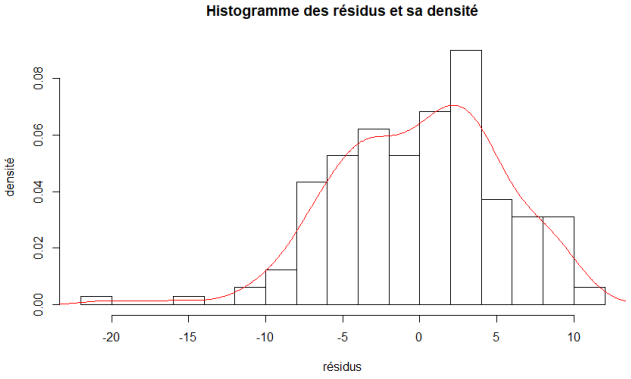
En effectuant le test pour X1, X2, X3 et X4, tous les quatre obtiennent une p-value très petite, très proche de 0. On a donc une p-value inférieure à notre risque d'erreur. Ainsi nous décidons de choisir et d'accepter que chaque X_i explique Y.

Nous pouvons donc conclure dans un premier temps que le VIH/SIDA, la consommation d'alcool, le taux de vaccination contre la polio et le taux de maigreur sont des facteurs qui permettent d'expliquer l'espérance de vie.

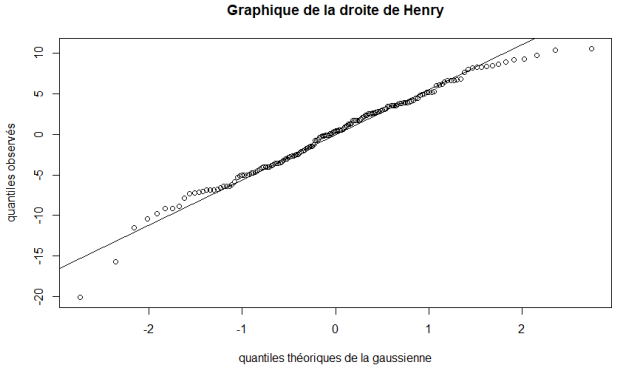
Nous allons maintenant faire une étude des résidus pour confirmer l'efficacité du modèle choisi. En effet, les résidus doivent répondre à quatre hypothèses :

- Centrage : $\forall_{i=1,..,n}\ E(\hat{u}_i)=0$
- Homoscédasticité : $\forall_{i=1,..,n}\ Var(\hat{u}_i)= \sigma^2$
- Non corrélation : $\forall_{i\neq j}\ E(\hat{u}_i\hat{u}_j)=0$
- Gaussianité : $\forall_i$, $\hat{u}_i$ suit une loi gaussienne

Pour commencer et confirmer que les résidus suivent une loi gaussienne, nous allons regarder si les résidus suivent une loi symétrique. Tout d'abord, nous constatons que la moyenne est de 0 et la médiane est également très proche de 0, elle est de 0.4. On peut aussi observer que le premier et le troisième quartiles sont plus ou moins symétriques. En effet, le premier quartile est à -3.799 et le troisième à 3.703. De plus, la densité des résidus est représentée par la ligne rouge sur le graphique ci-dessous :



Comme vous pouvez le remarquer, la densité et l'histogramme se superposent, il n'y a pas de grandes disparités entre les deux. On peut également remarquer l'allure gaussienne que prend la densité, malgré le fait qu'elle ne soit pas parfaitement symétrique. Pour vérifier si notre observation s'avère juste, il faut en premier lieu tracer la droite de Henry. Cette droite produit un graphe de points plus ou moins alignés si la densité est gaussienne. Si nous obtenons un alignement approximatif des points, cela appuiera notre intuition et nous pourrons alors entreprendre un test statistique qui confirmera pour de bon la gaussianité de la densité.



Avec les données récoltées sur les différents pays en 2014, nous obtenons le graphe ci-dessus. On peut noter que tous les points sont globalement alignés, mis à part quelques exceptions que l'on peut remarquer en haut à droite et en bas à gauche. Ainsi, nous pouvons modéliser la loi des résidus par une gaussienne. Voyons à présent ce que va nous dire le test de Shapiro que nous allons effectuer.

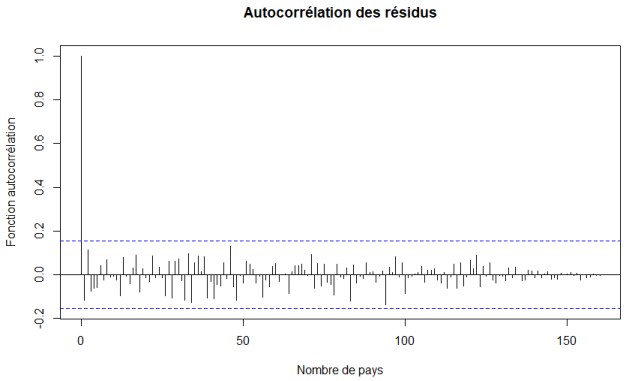
Pour cela, il faut tout d'abord poser le problème de test qui est le suivant :

H_0 : $\hat{u}$ suit une loi gaussienne

H_1 : $\hat{u}$ ne suit pas une loi gaussienne

Avec $\alpha = 1\%$ comme risque d'erreur. Si on obtient une p-value plus grande que notre risque d'erreur alors on choisit H_0. Au contraire, si la p-value est inférieure à notre risque d'erreur, alors on rejette H_0 et on accepte H_1. En effectuant ce test, on trouve une p-value de 3%. La p-value est donc supérieure au risque d'erreur que nous nous sommes fixées, nous décidons alors de choisir H_0 et d'accepter la gaussianité de nos résidus.

Toutes les conditions sur les résidus sont donc vérifiées, nous allons donc pouvoir maintenant procéder à une étude de l'autocorrélation de ces derniers. Pour cela, nous allons représenter le graphique suivant :

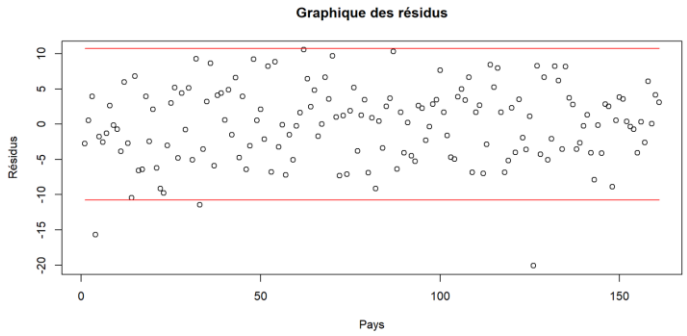


Pour que notre modèle soit correct, l'autocorrélation des résidus entre chaque pays doit être quasiment nulle et donc ne doit pas dépasser les deux lignes bleues. Nous pouvons donc observer ici qu'aucune des barres, à part la première, ce qui est normal, ne dépasse ces lignes. Donc nous pouvons dire que les résidus ne sont pas auto-corrélés. Pour vérifier si notre observation s'avère juste, nous allons réaliser un test statistique qui est le test de Dubin-Watson. Tout d'abord, posons le problème de test :

H_0 : les résidus ne sont pas auto-corrélés
H_1 : les résidus sont auto-corrélés

Avec α =1% comme risque d'erreur. Si on obtient une p-value plus grande que notre risque d'erreur alors on choisit H_0. Au contraire, si la p-value est inférieure à notre risque d'erreur, alors on accepte H_1. En effectuant ce test, on trouve une p-value de 92%. La p-value est donc supérieure au risque d'erreur que nous nous sommes fixées, nous décidons alors de choisir H_0 et d'accepter que les résidus ne soient pas auto-corrélés.

Après cette étude sur les résidus, faisons maintenant une étude sur les individus, afin de détecter les différents pays atypiques par rapport à nos quatre variables explicatives. Voici le graphe représentant les résidus obtenus par la régression :



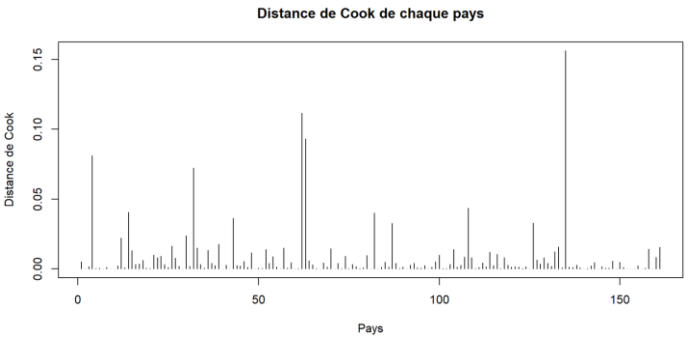
On voit que les points, représentant les résidus, ont une répartition équilibrée, c'est-à-dire qu'environ 50 % sont dans la zone positive au-dessus de la droite y = 0 et 50 % sont en dessous. En ayant ajouté les deux bandes rouges qui représentent le double de l'estimation de l'écart type, cela nous facilite le repérage des pays ayant des caractéristiques atypiques par rapport à nos variables X1, X2, X3, X4 choisies. Ainsi, on voit que quelques pays ressortent :

- L'Angola possède un niveau de mortalité, dû au VIH/SIDA, très élevé par rapport aux autres pays. Le niveau est de 2 morts pour 1000 habitants, sachant que le 3ème quartile est de 0.3 mort pour 1000 habitants donc 75 % des pays ont un niveau de mortalité du VIH/SIDA inférieur ou égale à 0.3/1000. De même, ce pays possède un taux de vaccination contre la polio assez faible, elle est de 68 % alors que la médiane se situe à 94 % donc 50 % des pays ont une couverture vaccinale contre la polio supérieure ou égale à 94 %. On peut de nouveau noter, un taux de maigreur, chez les 10-19 ans, assez important : il est de 8.5 %. Le 3ème quartile de cette variable étant de 6.6 % donc l'Angola a un taux presque 2 points au-dessus par rapport à 75 % des pays. Enfin, on peut noter que l'Angola ayant une espérance de vie de 51.7 ans se démarque avec la variable X2, niveau de consommation d'alcool. En effet, au vu de nos données, nous avons pu remarquer que le niveau d'espérance de vie et le niveau de consommation d'alcool ont une corrélation positive. L'Angola ayant une espérance de vie faible devrait avoir un niveau de X2 d'autant plus faible, or ce n'est pas le cas : ce pays a une variable X2 de 8.33 litres contre une médiane de 0.4 et une moyenne de 3.3.

- La Côte d'Ivoire est aussi à remarquer, avec une espérance de vie de 52.8 ans, le niveau de mortalité du VIH/SIDA est de 2 morts pour 1000 habitants, soit 1.7 point au-dessus de la médiane des pays de notre étude. Ajouté à cela, sa couverture vaccinale contre la polio est de 76% ce qui est également assez faible puisque cela représente 18 points en-dessous de la médiane de cette variable X3. Le taux de maigreur est également atypique avec un taux de 5.6 % ce qui est donc plus important que chez 50 % des pays de notre étude puisque la médiane de cette variable est de 3 %.

- La Sierra Leone est également à souligner avec un niveau de mortalité du VIH/SIDA de 0.6 mort pour 1000 habitants ce qui est supérieur à 75 % des pays de notre étude. Ce pays, possédant une espérance de vie de 48.1 ans, a également un taux de maigreur de 7.5 % ce qui est donc 0.9 point de plus que 75 % des autres pays et 4.5 points plus important que chez 50 % de nos sujets d'étude. Enfin la couverture vaccinale de la Sierra Leone est de 83 %, situant ainsi le pays dans la moitié inférieure des pays par rapport à cette variable, la médiane étant de 94 %.

On peut compléter ce repérage avec un second graphe, celui représentant les distances de Cook.



Avec ce graphique, on peut voir que cinq pays ont des distances de Cook relativement grandes par rapport aux autres pays. Ces cinq pays sont : l'Angola, le Congo, l'Islande, l'Inde et le Swaziland avec une espérance de vie respective de 51.7, 64.2, 82.5, 68.0 et 58.4 ans. Voici un extrait de nos données sur ces différents pays :

Pays	VIH/ SIDA (X1)	Alcool (X2)	Couverture vaccinale contre la polio (X3)	Taux de maigreur chez les 10-19 ans (X4)
Angola	2.0	8.33	68	8.5
Congo	3.0	0.01	9	7.6
Islande	0.1	7.45	9	1.0
Inde	0.2	3.07	84	26.8
Swaziland	7.3	0.01	98	4.3

L'Angola se distingue pour les mêmes raisons citées précédemment. On peut remarquer que les quatre autres pays ont une ou plusieurs variables qui se distinguent. Par exemple, l'Inde se démarque avec un taux très important de maigreur chez les 10-19 ans, 26.8 %, ce qui est la valeur maximale des valeurs prises par l'ensemble des pays pour cette variable X4, contre une médiane de 3 %. L'Islande et le Congo ont un taux très faible de vaccination contre la polio, un taux de 9 % contre le premier quartile de 79 % soit 70 points de moins que 75 % des pays de notre étude. Bien que le niveau d'espérance de vie de l'Islande soit élevé, le pays se distingue par son taux anormal de vaccination contre la polio, pour un pays à forte espérance de vie. De même, on peut également noter le niveau important de mortalité pour 1000 habitants du VIH/SIDA, du Swaziland et du Congo qui ont respectivement un niveau de 7.3 et de 3 alors que la médiane est de 0.1. Enfin, l'Inde se différencie également avec un niveau de consommation d'alcool de 3.07 litres ce qui représente une valeur importante pour un pays à faible espérance de vie.

Par cette étude des individus, nous avons pu étudier certains pays aux caractéristiques atypiques et dans quelle mesure ils le sont. Cela nous permet ainsi de comprendre les différences entre ces pays, amenant à une divergence importante de leur variable Y, l'espérance de vie. Par ailleurs avec leurs spécificités, on peut noter qu'ils ont donc une influence assez importante sur notre modélisation.

Faisons maintenant une étude sur les régresseurs, c'est-à-dire sur les variables explicatives. Il est important que les variables explicatives choisies ne soient pas corrélées de manière trop importante entre elles. En effet, expliquer l'espérance de vie avec deux ou plusieurs variables similaires n'est pas pertinent. Assurons-nous que ces variables explicatives sont peu corrélées et calculons donc la tolérance de chaque variable Xi par rapport aux autres variables Xj. Pour cela, commençons par faire les régressions linéaires sur chaque Xi en fonction de toutes les autres variables.

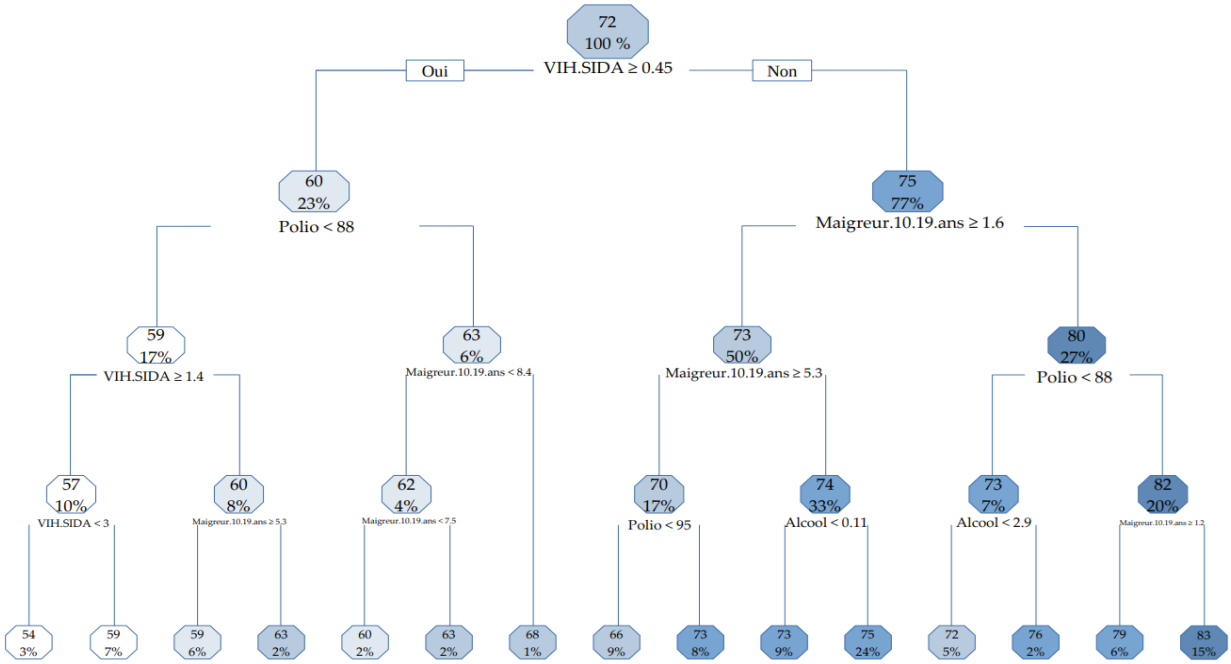
On note donc qu'il est nécessaire d'obtenir un R^2 petit, montrant ainsi que chaque variable explicative ne peut être expliquée par les autres. On obtient avec chaque régression, le résultat suivant :

- VIH/SIDA : $R^2 = 0.09385$
- alcool : $R^2 = 0.1944$
- couverture vaccinale contre la polio : $R^2 = 0.06356$
- taux de maigreur chez les 10-19 ans : $R^2 = 0.1677$

On a donc un niveau de tolérance $tol(Xi) = 1 - R^2$ pour chaque variable :

- tol(X1) = 0.90615 soit 90.61 %
- tol(X2) = 0.8056 soit 80.56 %
- tol(X3) = 0.93644 soit 93.64 %
- tol(X4) = 0.8323 soit 83.23 %

Les tolérances de chaque variable par rapport aux autres sont toutes assez grandes donc les régresseurs sont bien décorrélés les uns des autres. On peut donc affirmer que les choix des variables et du modèle sont bien pertinents pour expliquer la différence du niveau d'espérance de vie entre les pays.



Pour finir cette édition sur l'espérance de vie, nous vous avons concocté un petit jeu, sous forme d'arbre, afin que vous puissiez déterminer par vous-même votre propre niveau d'espérance de vie.

Voici un tableau de données avec une liste de pays qui pourront vous aider :

Pays	VIH/ SIDA (X1)	Alcool (X2)	Couverture vaccinale contre la Polio (X3)	Taux de maigreur chez les 10-19 ans (X4)
Afghanistan	0.1	0.01	58	17.5
Bénin	1.1	0.01	74	7.1
Brésil	0.1	7.32	96	2.7
Canada	0.1	8.1	91	0.5
Cameroun	3.7	0.01	86	5.7
Chine	0.1	5.78	99	3.7
Equateur	0.1	3.82	84	1.2
France	0.1	11.5	98	0.6
Guinée	0.5	0.01	42	7.5
Italie	0.1	7.56	95	0.6
Japon	0.1	0.01	99	2
Nigeria	3.9	0.01	49	1.1
Philippines	0.1	4.52	77	1
Fédération de Russie	0.1	10.12	97	2.3
Ukraine	0.2	8.06	45	2.3
Etats-Unis d'Amérique	0.1	8.82	93	0.8
Vietnam	0.1	4.09	96	14.3
Zimbabwe	6.3	6.5	92	5.9

Avec ces données, nous pouvons constater que certaines tendances se démarquent :

- les pays occidentaux qui ont généralement une espérance de vie élevée, ont également un taux de maigreur chez les 10-19 ans très faible et un niveau de consommation d'alcool assez important, ce qui indique qu'une bonne alimentation peut partiellement corriger les impacts négatifs d'une consommation d'alcool. De plus, ces pays dits développés sont ceux avec un niveau de richesses beaucoup plus important que les pays à faible espérance de vie, on peut donc en déduire qu'ils ont les capacités économiques de soigner leur population.

- les pays peu développés voire non développés ont généralement une espérance de vie faible. Ils possèdent également une couverture vaccinale contre la polio assez faible par rapport aux pays développés, ainsi qu'un niveau de mortalité pour 1000 habitants du VIH/SIDA plus important. Cette tendance indique le fait que ces pays, dits pauvres ont peu de moyens économiques dédiés au secteur de la santé.

Enfin, si vous aviez un doute concernant la lecture de nos données, voici quelques exemples :
Si vous vivez en Italie, votre pays possède un niveau de mortalité du VIH/SIDA de 0.1%, un taux de maigreur des 10-19 ans de 0.6%, une couverture vaccinale contre la Polio de 95%, et un niveau de consommation d'Alcool de 7.56 litres. Dans ce cas de figure, votre espérance de vie est d'environ 83 ans.

De même, si vous vivez en Guinée, votre pays a un niveau de mortalité du VIH/SIDA de 0.5%, un taux de maigreur des 10-19 ans de 7.5%, une couverture vaccinale contre la Polio de 42%, et un niveau de consommation d'alcool de 0.01 litres. Dans ce cas de figure, votre espérance de vie est d'environ 59 ans.

Un dernier exemple, avec l'Equateur, ce pays a un niveau de mortalité du VIH/SIDA de 0.1%, un taux de maigreur des 10-19 ans de 1.2%, une couverture vaccinale contre la Polio de 84%, et un niveau de consommation d'alcool de 3.82 litres. Dans ce cas de figure, l'espérance de vie est d'environ 76 ans.

En conclusion, à travers ces différentes études, nous avons donc pu constater que le pays où nous naissons et vivons détermine bien notre espérance de vie. En effet, la situation de notre pays de naissance joue un rôle important que ce soit en termes d'éducation impactant par la suite le niveau de vie comme on l'a vu dans l'édition précédente ou encore en termes de santé. Notons que les pays, développés en termes de richesse économique, qui ont généralement une espérance de vie élevée, se caractérisent par un taux de maigreur chez les 10-19 ans très faible et un niveau de consommation d'alcool assez important, un taux extrêmement faible de mortalité dû au VIH/SIDA, ainsi qu'une protection vaccinale contre la polio d'environ 90%. Au contraire, les pays dits pauvres, tels que beaucoup de pays d'Afrique, ont une espérance de vie globalement faible, bien que leur consommation d'alcool soit faible, ce qui s'explique par un taux de maigreur assez important, dû au manque d'accès à une nourriture de qualité, par une couverture vaccinale très faible et par un taux de mortalité du VIH/SIDA très élevé, causés par un manque d'accès au soin.

ANNEXE

Pour effectuer notre projet, nous avons pris nos données sur le site Kaggle et nous avons également fait des recherches pour compléter les quelques données manquantes.

Voici la liste des pays de notre étude :

Afghanistan
Afrique du Sud
Albanie
Algérie
Allemagne
Angola
Antigua-et-Barbuda
Arabie saoudite
Argentine
Arménie
Australie
Autriche
Azerbaijan
Bahamas
Bangladesh
Barbade
Biélorussie
Belgique
Bélize
Bénin
Bhutan
Bolivie
Bosnie Herzégovine
Bulgarie
Burkina Faso
Burundi
Cameroun
Canada
Chili
Chine
Colombie
Comores
Congo
Cote d'Ivoire
Croatie
Cuba
Chypre
Danemark
Equateur
Egypte
Emirats-Arabes-Unis
Érythrée
Espagne
Estonie
Etats-Unis d'Amérique
Ethiopie
Fidji
Finlande
Fédération de Russie
France
Gambie
Géorgie
Ghana
Grèce
Grenade
Guatemala
Guinée
Guinée-Bissau
Guinée équatoriale
Haiti
Honduras
Hongrie
Iles Salomon
Inde
Indonésie
Iran
Iraq
Irlande
Islande
Israël
Italie
Japon
Jordanie
Kazakhstan
Kenya
Kiribati
Koweit
Kyrgyzstan
Laos
Lettonie
Liban
Liberia
Libye
Lituanie
Luxembourg
Macédoine du Nord
Madagascar
Malawi
Malaisie
Maldives
Mali
Malte
Maroc
Mauritanie
Mexique
Micronésie
Mongolie
Montenegro
Mozambique
Myanmar
Namibie
Népal
Nicaragua
Niger
Nigéria
Norvège
Nouvelle Zélande
Oman
Ouganda
Ouzbékistan
Pakistan
Panama
Papouasie Nouvelle Guinée
Paraguay
Pays-Bas
Péru
Philippines
Pologne
Portugal
Qatar
République Centre Afrique
République de Corée
République Démocratique du Congo
République dominicaine
Roumanie
Royaume-Uni
Rwanda
Salvador
Samoa
Sao Tomé-et-Principe
Sénégal
Serbie
Seychelles
Sierra Leone
Singapour
Slovaquie
Slovénie
Sri Lanka
Suède
Suisse
Suriname
Swaziland
Syrie
Tajikistan
Tchad
Thailande
Timor-Leste
Togo
Tonga
Trinité et Tobago
Tunisie
Turquie
Turkménistan
Ukraine
Uruguay
Vanuatu
Venezuela
Viet Nam
Yémen
Zambie
Zimbabwé