Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Métodos Numéricos

2do cuatrimestre 2021

Trabajo Práctico 3

Integrantes

Integrante	LU	Correo electrónico
Sebastián Bocaccio	287/18	sebastianbocaccio16@gmail.com
Oscar Alvarez	619/18	oscar_algu@hotmail.com
Andrés Barbuto	113/12	andresbarbuto@gmail.com
Martín Federico Mallol	208/20	martinmallolcc@gmail.com

Resumen

En este trabajo realizamos el estudio de distintas características de los países y su relación con su expectativa de vida. Primero, realizamos un análisis exploratorio de los datos brindados por la cátedra para poder entender mejor y saber distinguir entre las características que se relacionan con la calidad de vida y aquellas que no. También, agregamos algunos características a los países tales como tasa de suicidios, enfermedades virales, información geográfica, etc.(Información provista por la Organización Mundial de la Salud) para la realización de nuestros experimentos con el objetivo de poder observar como estas características impactan sobre la expectativa de vida de la población de los países.

Palabras Claves

Regresión Lineal; Cuadrados mínimos; Análisis de datos;

Reservado para la cátedra

Instancia	Docente	Nota
Primera entrega		
Segunda entrega		

Índice

1. Introducción teórica

En este trabajo nos dedicamos a investigar sobre el problema de tratar de explicar la esperanza de vida de los países según la relación con las demás variables. Nuestro objetivo es usar esta información para luego realizar una regresión sobre los datos y ver cual es la relación de cada variable con la esperanza de vida. Las regresiones pueden ser de suma importancia en casos donde se busque encontrar la manera de explicar datos que en principio se desconozca la familia de funciones que los responden

Como dataset utilizamos inicialmente un conjunto de datos otorgado por la cátedra con una amplia cantidad de países y features de los mismos, incluida la esperanza de vida. Pero esta dataset no fue el único incluido, para algunos experimentos utilizamos información otorgada por la organización mundial de la salud la cual ampliaba nuestra data anterior. Toda la información recogida pertenece siempre a años no posteriores a 2015.

Como todo trabajo en donde se trabaje con datos, es fundamental realizar un análisis exploratorio primero para poder luego hacerles un mejor uso, ya que se lograra conocer mejor su estructura. Además, este estudio preliminar nos ayuda a identificar información errónea o datos atípicos los cuales deben ser tomados en cuenta ya que pueden perturbar enormemente los resultados de cualquier análisis si no se tienen en consideración.

1.1. Cuadrados Minimos Lineales(Regresion Lineal)

El método de regresión lineal es un método estadístico que nos permite resumir y estudiar la relación entre una variable continua, denotada y, a la cual llamamos respuesta.º "variable dependientez otra variable o muchas variables x_i a las cuales llamaremos "predictores.º "variables independientes". La idea es modelar estas relaciones mediante ecuaciones normales $A^tAx = A^tb$, donde las columnas de A son los "predictores", b es la respuestaz x vendría a ser la recta que mejor ajusta nuestros datos. Este método nos podría resultar interesante por ejemplo cuando tenemos datos desconocidos, es decir, más ecuaciones que variables(matriz no cuadrada con mas filas que columnas). A^tA es cuadrada, así que me va a poder brindar una predicción "suficientemente buena"sobre los datos faltantes.

2. Desarrollo

2.1. Análisis exploratorio de datos

En esta sección nos dedicaremos a explorar el dataset entregado por la cátedra con el objetivo de lograr un mejor entendimiento de la estructura del mismo. Esto lo hacemos dado que un mejor entendimiento de los datos nos permite entender mejor que características de los paıses se relacionan con la expectativa de vida de sus habitantes y cuales no.

2.1.1. Información general del dataset

En el dataset se encuentran en total 183 países donde se puede obtener los siguiente tipos de datos sobre cada país

Feature	Breve descripción	Tipo de variable
Status	Si el pais es considerado desarrollado o en desarrollo	categórica
Life expectancy	Esperanza de vida en años	numérica acotada
Adult mortality	Cantidad de muertos entre 15 y 60 años cada 1000 habitantes	numérica porcentaje
Infant deaths	Cantidad de infantes muertos cada 1000 habitantes	numérica porcentaje
Alcohol	Cantidad de consumo de alcohol en litros per capita	numérica no acotada
percentage expenditure	porcentaje del pbi empleado en salud	numérica porcentaje
Hepatitis B	inmunización entre bebes menores a 1 año habitantes	numérica porcentaje
Measles	Cantidad de casos de sarampion cada 1000	numérica porcentaje
BMI	Promedio de la base corporal de la población	numérica acotada
under-five deaths	Número de muertes de menores de 5 años cada 1000 habitantes	numérica porcentaje
Polio	Porcentage de inmunización en menores de 1 año	numérica porcentaje
Total expenditure	Porcentage empleado en salud sobre el porcentage de GE	numérica porcentaje
Diphtheria	inmunización entre bebes menores a 1 año	numérica porcentaje
HIV/AIDS	Cantidad de muertes cada 1000 nacimientos	numérica porcentaje
GDP	PBI per capita (en USD)	numérica no acotada
Population	Cantidad de población	numérica no acotada
thinness 1-19 years	Prevalencia de delgadez en población entre 10 y 19	numérica porcentaje
thinness 5-9 years	Prevalencia de delgadez en población entre 5 y 9	numérica porcentaje
Income composition of resources	Que tan productivamente se utilizan los recursos	numérica porcentaje
Schooling	Cantidad de años en la escuela	numérica acotada

Cuadro 1: Informacion sobre los tipos de variables que se encuentran en el dataset

Aclaración: Adicionalmente a los features previamente mencionados, había en el dataset una variable llamada Unnamed 0 la cual decidimos descartar su uso en todo el trabajo debido a que nos es imposible inferir cual es su significado y creemos que no nos va a poder aportar herramientas para un mejor análisis del fenómeno.

En caso de desconocer la nomenclatura sobre los tipos de variables, se puede encontrar información al respecto [1]

2.1.2. Visualización de la esperanza de vida de los países

Tengamos un primer vistazo para observar como es la distribución de años de vida en nuestra muestra de países.

En la figura 1 podemos observar que efectivamente la cantidad de años de vida esperados difiere enormemente; pudiendo tomando valores que pueden diferir hasta de 40 años. Esto es un dato, que si bien es ampliamente conocido, que haya países con estas diferencias en la esperanza de vida es información alarmante.

Por otro lado, podemos visualizar 3 posibles grupos de países. Para los que tienen esperanza menor a 70 años, parece haber una distribución equilibrada entre ellos. Alrededor de los 74 años hay un segundo grupo con distribución mas similar al de una normal y es en donde se encuentra el bin mayoritario en muestras. Por ultimo, a partir de los 77 hay un tercer grupo con una distribución que parece ser la mitad izquierda de una normal.

Con el objetivo de poder segmentar mas la muestra y ver si estos comportamientos se siguen reflejando para una partición mas minuciosa, decidimos realizar un gráfico similar pero esta vez con 50 bins. El resultado puede visualizarse en la figura 2

Efectivamente el comportamiento anterior se mantiene. Lo que nos resulta muy llamativo es el comportamiento del 3 grupo. Esperábamos ver que al aumentar la cantidad de años hubiese un decrecimiento gradual de la esperanza de vida y no un crecimiento que es lo que reflejan los datos.

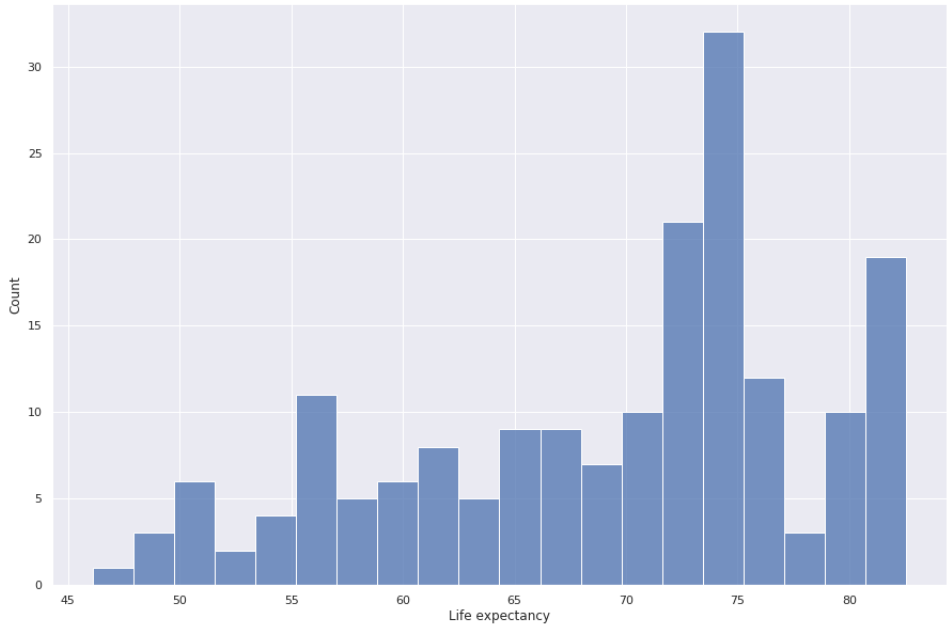


Figura 1: Distribución de la esperanza de vida de los países separándolo en 20 bins

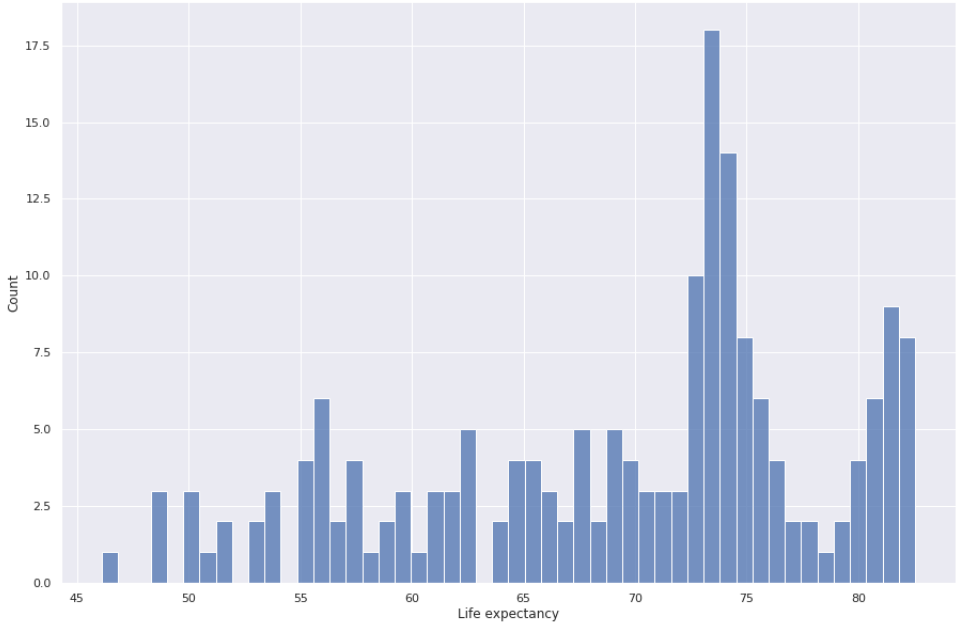


Figura 2: Distribución de la esperanza de vida de los países separándolo en 50 bins

Ahora bien, omo están distribuidos los valores de la esperanza en el mundo? Existen zonas del espacio que se caracterizan por tener mayor o menor esperanza? Observemos la figura 3

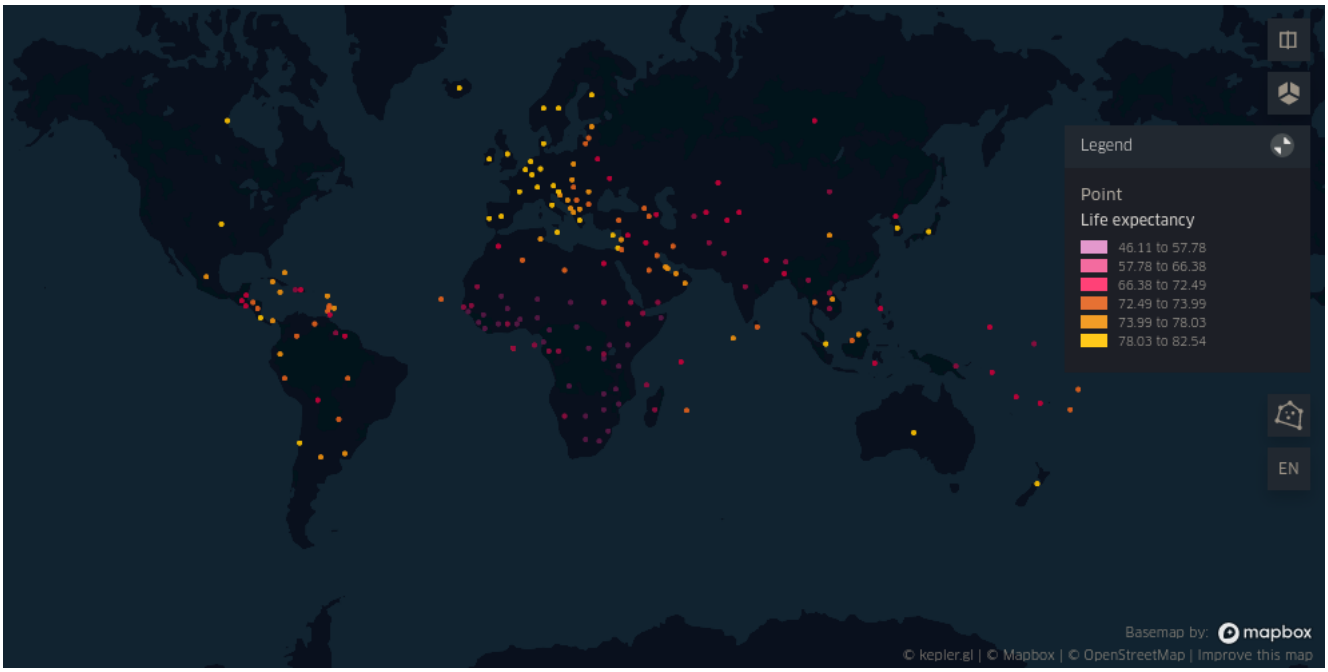


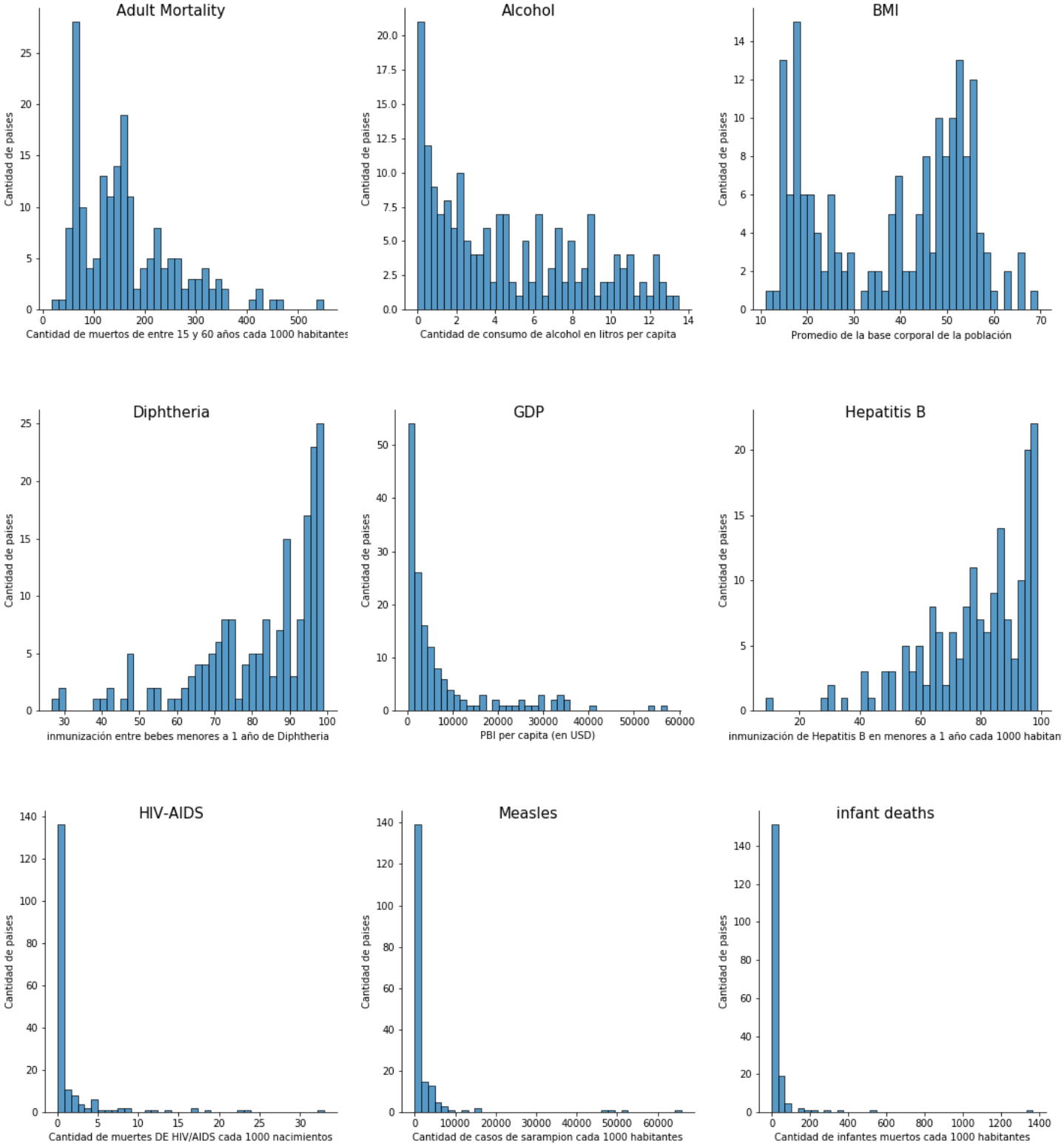
Figura 3: Esperanza de vida de los países según su ubicación geográfica

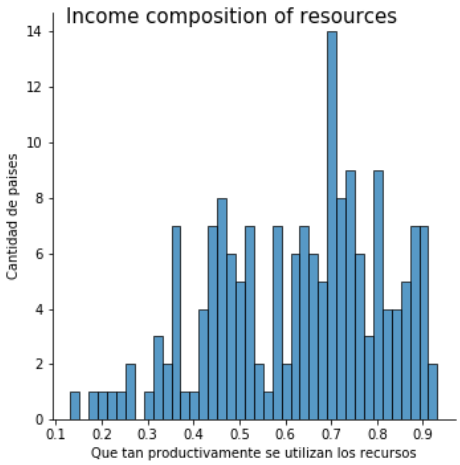
Efectivamente sucede que no esta distribuido uniformemente, sino que existen zonas donde hay una predominancia a una baja expectativa de vida que se evidencia muy rápidamente al observar África. También, existen otras zonas

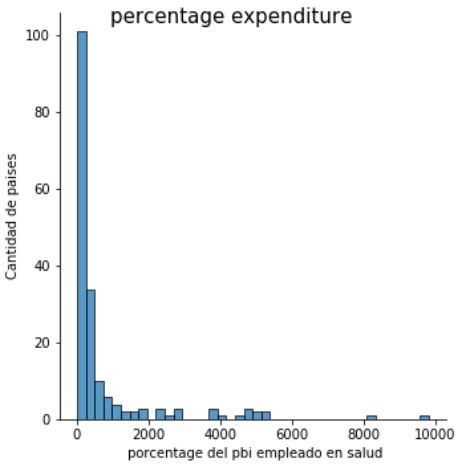
donde sucede lo contrario como en Europa occidental donde hay una gran predominancia a una alta expectativa.

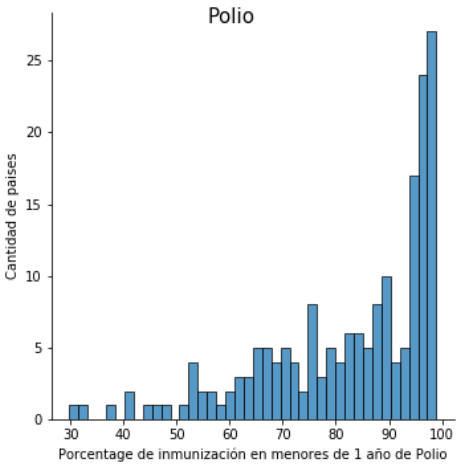
2.1.3. Visualización de las demás variables

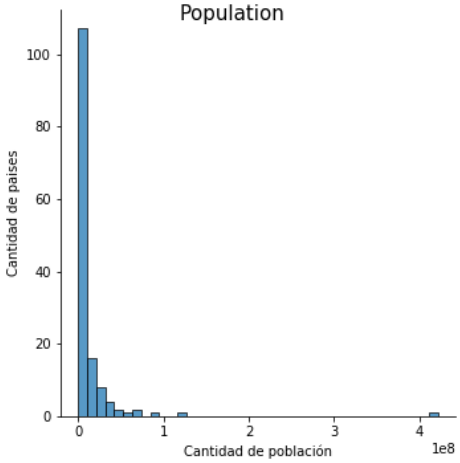
Veamos ahora como es la distribución de las demás variables. Nuestro objetivo es tratar de aprender un poco mas sobre el dataset y también poder visualizar si existen comportamientos anómalos, como por ejemplo valores de porcentaje que sean mayores a 100 o menores a 0, y los datos faltantes en alguno de los países. La distribución de los features pueden visualizarse en el conjunto de imágenes de la figura 4

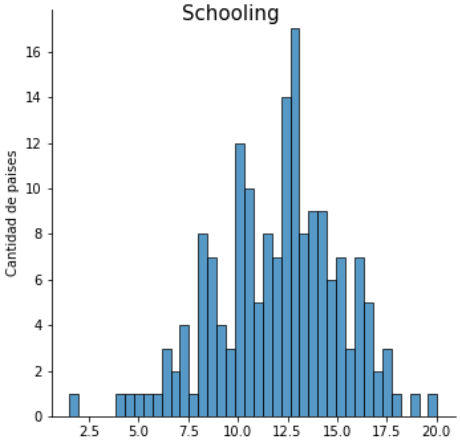


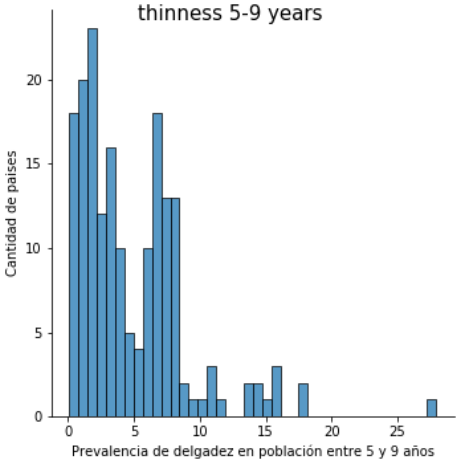

Income composition of resources
Cantidad de paises
Que tan productivamente se utilizan los recursos

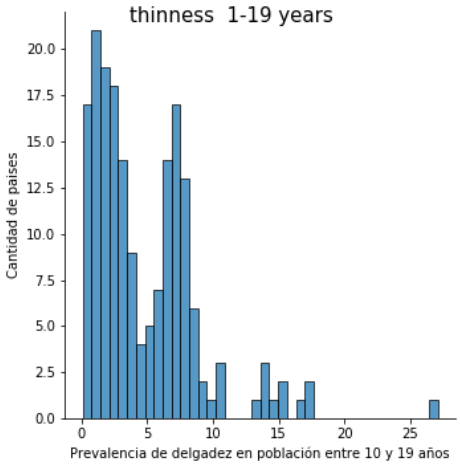

percentage expenditure
Cantidad de paises
porcentage del pbi empleado en salud

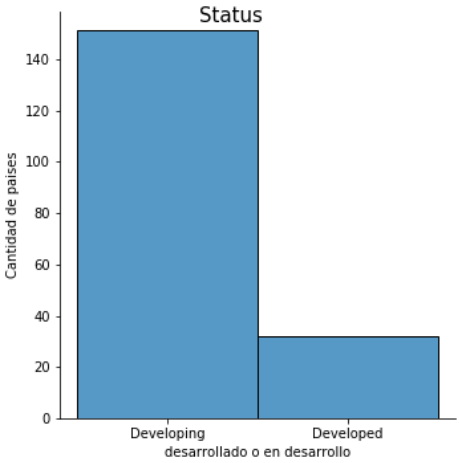

Polio
Cantidad de paises
Porcentage de inmunización en menores de 1 año de Polio

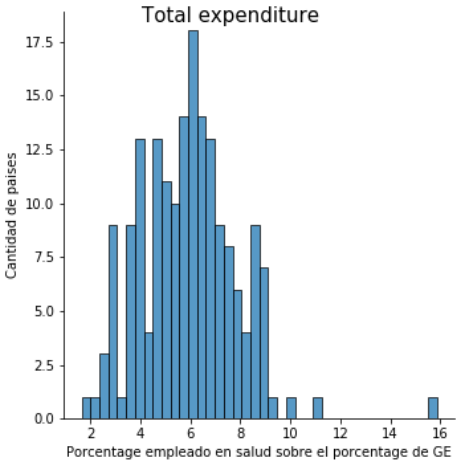

Population
Cantidad de paises
Cantidad de población
1e8


Schooling
Cantidad de paises
Cantidad de años en la escuela


thinness 5-9 years
Cantidad de paises
Prevalencia de delgadez en población entre 5 y 9 años


thinness 1-19 years
Cantidad de paises
Prevalencia de delgadez en población entre 10 y 19 años


Status
Cantidad de paises
Developing
Developed
desarrollado o en desarrollo


Total expenditure
Cantidad de paises
Porcentage empleado en salud sobre el porcentage de GE

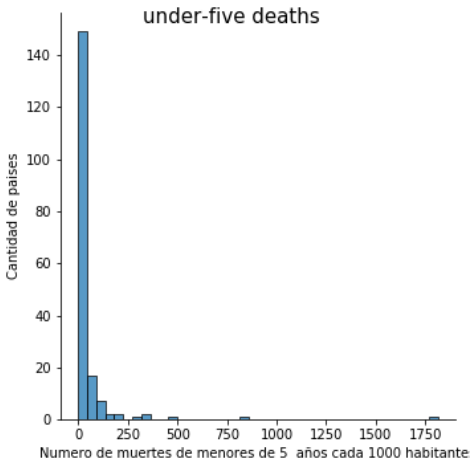


Figura 4: Distribución de las demás variables en el dataset

En lo queda de la sección, nos dedicaremos a revisar el comportamiento de cada feature individualmente y ver en cada una los comportamientos extraños que se pueden encontrar, además de una breve descripción de su distribución.

- Adult Mortality: En este caso el comportamiento general es que a medida que crece la mortalidad, disminuye la cantidad de países con esa tasa de mortalidad. Lo que nos pareció llamativo del gráfico son los primeros dos bins, que son los países con menor Adult Mortality y quisimos averiguar que países eran. Estos son Tunisia, Albania y Iceland; sus valores correspondientes de mortalidad son 18.7500, 45.0625 y 49.3750. Por otro lado, el caso con mas muertes sucede en Lesotho con 550.0625 muertos.

  Países sin información:
- Alcohol:

  En este caso nos resulto sorpresivo la cantidad de países con cantidad de consumición tan baja, quienes son mayoría a nivel global. Muchos de estos países resultaron ser los países musulmanes donde esto tiene sentido ya que el Islam prohíbe el consume de alcohol.

  Países sin información: South Sudan
- BMI:

  En este caso encontramos una distribución bimodal, lo cual nos sorprendió. Lo que nos llamo altamente la atención es que todos los países de la segunda moda poseen valores que se categorizan como obesidad [3].

  Países sin información: Sudan y South Sudan
- Diphtheria:

  En este caso podemos observar que la mayoría de los países tienen una muy alta inmunización ante esta enfermedad. Nos llamo la atención sobre los casos con los valores mas bajos y estos son Chad, Equatorial Guinea y Somalia.

  Países sin información:
- GDP:

  En este caso encontramos lo que esperábamos, muchos mas países con menos ingresos y pocos con valores significativamente muy superiores.

  Países sin información: Bahamas, Bolivia (Plurinational State of), Congo, Czechia', Côte d'Ivoire, Democratic People's Republic of Korea, Democratic Republic of the Congo, Egypt, Gambia, Iran (Islamic Republic of), Kyrgyzstan, Lao People's Democratic Republic, Micronesia (Federated States of), Republic of Korea, Republic of Moldova, Saint Lucia, Saint Vincent and the Grenadines, Slovakia, The former Yugoslav republic of Macedonia, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania, United States of America, Venezuela (Bolivarian Republic of), Viet Nam, Yemen
- Hepatitis B:

  En este caso también observamos lo que esperábamos, que la inmunización de la hepatitis B sea muy alta y que a medida que la cobertura baje, también bajen la cantidad de países. El caso mas anómalo es el de Equatorial Guinea que posee tan solo un 9 %.

Países sin información: Switzerland, Denmark, Finland, Norway, Hungary, Slovenia, Iceland, Japan y United Kingdom

- HIV-AIDS, Measles, infath Deaths y under-five deaths:

  Decidimos estudiar todas estas características de la población en conjunto ya que obtuvieron resultados ampliamente similares y esperábamos esto. En todas las features los números son muy cercanos a 0 en la mayoría de los casos, pero se encuentran países con valores altos. Los países mas atípicos para cada categoría son en su respectivo orden Swaziland (donde mas de un tercio de su población adulta esta infectada de HIV [4]), China y por ultimo India donde se reportan valores de 1366.6875 y 1812.5 respectivamente de muertes cada 1000 habitantes. Estos últimos evidentemente son datos erróneos por lo que deben ser eliminados durante los análisis. En el caso de Measles, tenemos numeros superiores a 1000, por lo cual no parece tener sentido estos datos. Existe la posibilidad de que haya en el dataset y esten ingresados los casos totales en vez del porcentaje.

  Países sin información:

- Income Composition of resources:

  En este caso no esperábamos ningún comportamiento particular ya que nuestro conocimiento sobre esta feature sobre los países era nula. Su distribución, si bien difiere, tienen ciertas semejantes a una normal centrada en el 0.7 %. Lo que mas nos sorprendió es la cantidad de países que no incluían información al respecto.

  Países sin información: Czechia, Côte d'Ivoire, Democratic People's Republic of Korea, Democratic Republic of the Congo, Republic of Korea, Republic of Moldova, Somalia, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania y United States of America

- Percentage expenditure:

  En este caso tampoco esperábamos ningún comportamiento particular ya que también nuestro conocimiento sobre esta feature era nula. Por desgracia nos esperamos con un set de datos bastante extraños con respecto a esta feature ya que muchos valores sobrepasan el valor posible. A la hora de utilizar esta feature esto va tener que ser tomado en cuenta ya que es tan solo un dato, sino bastantes.

  Países sin información:

- Polio:

  En este caso encontramos una curva como la que esperábamos ver. La cantidad de países con el porcentaje de inmunización en menores de 1 año crece a medida que crece el porcentaje. Donde se encuentran los valores minimos es alrededor del 30 % para Somalia y Chad.

  Países sin información:

- Population:

  Este caso tiene un comportamiento muy parecido al de HIV-AIDS, Measles, infath Deaths y under-five deaths ya que un valor atípico modifica la visualización en el gráfico. En este caso el país con

  La cantidad de países que tenemos sin información resulta sorpresivo dado que la cantidad de población es un dato elemental de un país. Esta cantidad de ausencias claramente va a tener que ser consideraba a la hora de querer usar a la población como medio para explicar la esperanza de vida.

  Países sin información: Antigua and Barbuda, Bahamas, Bahrain, Barbados, Bolivia , Brunei Darussalam, Congo, Cuba, Czechia, Côte d'Ivoire, Democratic People's Republic of Korea, Democratic Republic of the Congo, Egypt, Gambia', Grenada, Iran (Islamic Republic of), Kuwait, Kyrgyzstan, Lao People's Democratic Republic, Libya, Micronesia (Federated States of), New Zealand, Oman, Qatar, Republic of Korea, Republic of Moldova, Saint Lucia, Saint Vincent and the Grenadines, Saudi Arabia, Singapore, Slovakia, Somalia, The former Yugoslav republic of Macedonia, United Arab Emirates, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania, United States of America, Venezuela (Bolivarian Republic of) , Viet Nam, Yemen

- Schooling

  En este caso la cantidad de años que se dedican en la educación entre los países claramente es un comportamiento normal. El caso mas anómalo es el de South Sudan cuyo valor es de tan solo 1.53125 años. Lastimosamente, si bien esta puede no ser el numero exacto este pais es el que tiene la mayor proporción de chicos fuera de la escuela [5] por lo cual no creemos que podamos considerarlo como un dato erróneo.

  Países sin información: Czechia, Côte d'Ivoire, Democratic People's Republic of Korea, Democratic Republic of the Congo, Republic of Korea, Republic of Moldova, Somalia, United Kingdom of Great Britain and Northern Ireland, United Republic of Tanzania, United States of America

- thinness 1-19 years y thinness 5-9 years

  En este caso también decidimos estudiar ambos en conjunto debido al comportamiento similar que obtuvieron y que son dos features que al menos de forma intuitiva creemos están bastante relacionada. Encontramos una distribución bimodal concentrada entre el 4 % y el 8 %. Con respecto al grupo de países en el centro, Tenemos el caso mas atípico en donde ese porcentaje es muy alto, superando el 25 % para India. Para el grupo de países que están en el centro con valores entre 13 y 20 identificamos que son los mismos para ambas features siendo estos Afghanistan, Bangladesh, Bhutan, Maldives, Myanmar, Nepal, Pakistan, Sri Lanka, Viet Nam, Yemen.

  Países sin información: Sudan y South Sudan

- Status

  Se realiza un análisis de mayor profundidad de esta feature en la sección 2.1.4

  Países sin información:

- Total expenditure

  En este caso, vemos que claramente encontramos una distribución normal entre los datos centrada en el 6 %. Nos llamo particularmente la atención identificar si los países con menor porcentaje muestran una esperanza de vida menor a los de mayor. Tomamos los países que utilizan menos del 2.8 % y a los que utilizan mas del 9 %. En el cuadro 2 podemos observar que una mayor inversión en salud no necesariamente conlleva a una mayor esperanza de vida. Tenemos el caso de Sierra Leone donde tienen una esperanza bajísima de tan solo 46 años y por otro lado tenemos a Qatar que destina menos del 30 % de la proporciona que Sierra Leone destina y tiene valores hasta mayores al promedio. Aun así, entre esta submuestra de países se ve que exista una tendencia de que aquellos países que destinan mas fondos a salud, tienen mayor esperanza de vida.

  Países sin información: Democratic People's Republic of Korea, Somalia

País	Total expenditure	Life expectancy
Indonesia	2.663	67.5562
Myanmar	2.0126	64.2
Qatar	2.6013	77.0312
South Sudan	2.7100	53.875
Timor-Leste	1.6466	64.75625
Greece	9.0386	81.2187
Micronesia	11.056	68.2
Norway	9.086	81.7937
Sierra Leone	9.218	46.1125
Sweden	9.9326	82.5187
United States of America	15.8633	78.06250

Cuadro 2: Comparación entre los países con menor y mayor Total expenditure con su Life expectancy

2.1.4. Diferenciación entre países desarrollados y subdesarrollados

Dado nuestro dominio en este problema, creemos que un factor que va a influir muy fuertemente en la esperanza de vida de un país es el hecho de estar desarrollado o no, siendo estarlo un gran estimador de que la esperanza de vida sea mayor. A lo largo de esta sección vamos a explorar si efectivamente esto es algo que sucede en nuestros datos.

En nuestro dataset tenemos en total 32 países desarrollados y 151 en subdesarrollo. Es importante tener esta información en cuenta ya que hay un des balance bastante grandes entre las clases. Aun así, creemos que la cantidad de muestras de los países desarrollados a priori son suficientes para continuar con el análisis.

En la figura 5 podemos ver como se distribuye la esperanza de vida bajo esta categorización de los países.

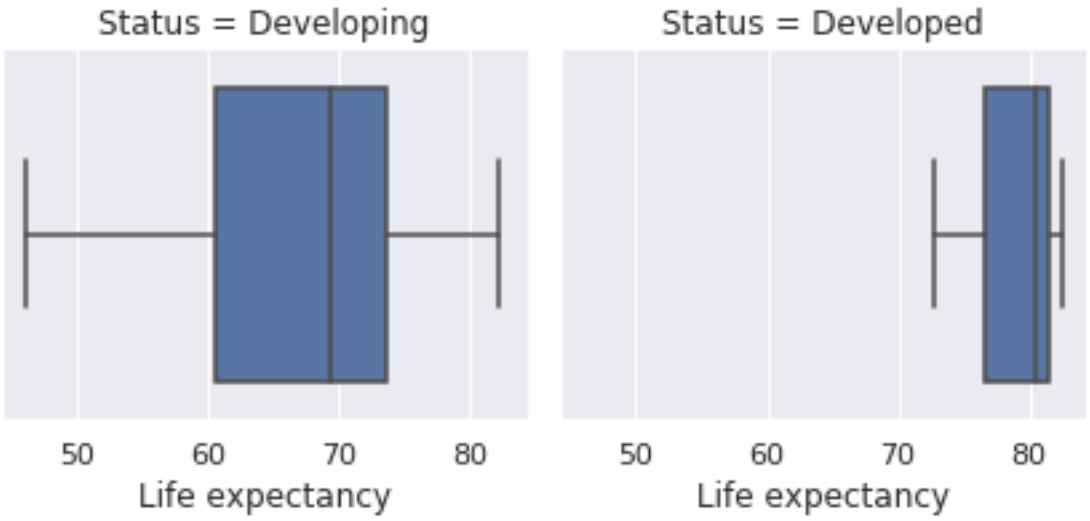


Figura 5: Distribución de la esperanza de vida en años según categorización de Status

Sobre este gráfico nos pareció resultante destacar que:

- Efectivamente el promedio de la esperanza de vida de los países desarrollados es mayor al de los países en desarrollo.
- La diferencia es significativamente grande, habiendo más de 10 años de separación entre la mediana de los dos grupos.
- La alta dispersión que se observa entre los países en desarrollo.
- Los países con mayor esperanza de vida pertenecen al grupo de los subdesarrollados.

Nos llamo sorpresivamente la atención el hecho de que si bien los países desarrollados tienen mayor esperanza de vida en promedio, el máximo no se encuentra entre uno de ellos. Decidimos investigar cuales eran esos países que tenían tal cualidad.

Dichos países son Canadá, Finlandia, Francia, Grecia, Israel y República de Korea. Todos estos pertenencen a los mas desarrollados [2]. Sin embargo, por alguna razón aparecen como *Developing* en el dataset.

Teniendo esto en cuenta, queremos observar si sucede lo inverso entre el grupo de desarrollados. Osea, que los países con menos esperanza de vida entre ellos no pertenecen realmente a esa categoría. Estos son Bulgaria, Hungría, Lithuania, Rumanía y Eslovaquia. Estos pertenecían al bloque socialista europeo y se consideran menos desarrollados que Canadá, Finlandia, Francia, Grecia, Israel y Corea del Sur.

Con estos análisis, creemos que es cuestionable la forma en la cual se definió el atributo de *Status* y consideramos importante hablar con la fuente de estos datos para poder generar un mejor entendimiento de porque el dataset esta categorizado de esta forma.

Países sin información:

2.1.5. Correlaciones entre variables no categóricas

En esta sección nos dedicaremos a estudiar la correlación entre grupos de features en búsqueda de algunos que se correlacionen fuertemente entre sí, para evitar lo mayor posible problemas de multicolinearidad. El proceso o término de multicolinealidad en econometría es una situación en la que se presenta una fuerte correlación entre variables explicativas del modelo. Con esto en mente, nos es importante resaltar que la correlación tiene que ser fuerte, ya que siempre existirá correlación entre dos variables explicativas en un modelo, es decir, la no correlación de dos variables es un proceso idílico, que sólo se podría encontrar en condiciones de laboratorio.

Decidimos visualizar las correlaciones entre las variables utilizando un heatmap el cual puede verse en la figura 6 Optamos por dejar de lado el feature *Life Expectancy* de este gráfico para analizarlo en profundidad luego.

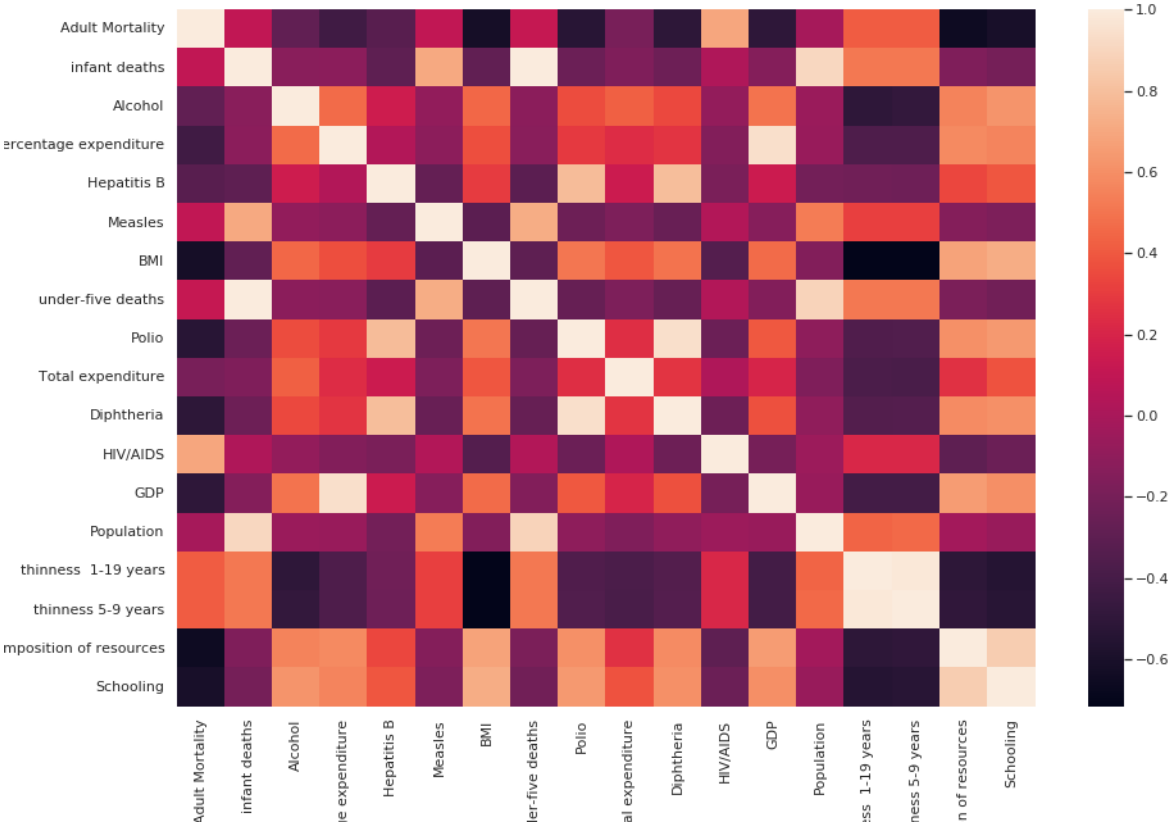


Figura 6: Correlación entre las variables no categoricas

A continuación, mostraremos cuales son los grupos mas correlacionados y sus valores exactos.

- Under-five deaths , Infant deaths , Population

Feature	Under-five deaths	Infant deaths	Population
Under-five deaths	1	0.996959	0.891045
Infant deaths	0.996959	1	0.906096
Population	0.891045	0.906096	1

  Tiene mucho sentido el que exista una amplia correlación entre las features de las muertes. Lo que nos llamo la atención es que la población también este muy correlacionado a este fenómeno.

- Thinnes 1-19 years, Thinnes 5-9 years y su correlación es 0.985098

  También parece bastante intuitivo que estos valores esten altamente correlacionados.

- GDP, Percentage expenditure y su correlación es 0.942375

  En este caso, podemos observar que a medida de que los países con mayor PBI tienden a destinan mas porcentaje del mismo a la salud.

2.1.6. Correlaciones con la esperanza de vida

En el cuadro 3 puede observarse cual es la correlación de cada variable con la esperanza de vida.

A continuación, charlaremos acerca de nuestras impresiones y hallazgos encontrados en esta tabla que nos parecieron importantes de destacar.

- Adult mortality es el feature que tiene más correlación, particularmente negativa. Nos parecio natural que este sea el factor mas correlacionado.

Feature	Correlación
Adult Mortality	-0.896441
infant deaths	-0.199914
Alcohol	0.461720
percentage expenditure	0.524320
Hepatitis B	0.429536
Measles	-0.201076
BMI	0.723824
under-five deaths	-0.225785
Polio	0.679231
Total expenditure	0.290713
Diphtheria	0.672322
HIV/AIDS	-0.587153
GDP	0.611808
Population	-0.039915
thinness 1-19 years	-0.523989
thinness 5-9 years	-0.515970
Income composition of resources	0.817545
Schooling	0.794457

Cuadro 3: Correlación de cada variable numérica con la esperanza de vida

- A continuación le sigue el Income Composition of Resource lo cual nos dice que la alta utilización de nuestros recursos hace que sea mas probable que la población viva mas tiempo.
- De forma similar a lo anterior, existe una alta correlación entre la expectativa de vida y la escolaridad de un país.
- Nos llamo la atención la poca correlación que existe entre expectativa de vida y las variables sobre la mortalidad infantil (infant deaths y under-five deaths). Esperábamos encontrar valores muchos mas altos.
- La correlación entre el índice de masa corporal y la esperanza de vida es significativamente menor al del GDP y la esperanza de vida. Creemos que estamos observando estos resultados debido a la distribución no equitativa que hay en los países.
- La correlación positiva no despreciable entre el consumo de alcohol y esperanza de vida. Sabemos que de hecho existe una relación inversa en realidad; mayor consumo de alcohol es dañino para la salud.

Es importante siempre tener en cuenta que la correlación examina la relación entre dos variables. Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso se suele decir que "la correlación no implica causalidad".

Lo anterior se puede entender fácilmente teniendo en cuenta el ejemplo del alcohol.

2.2. Experimentación

2.2.1. Relación entre expectativa de vida e información geográfica de los países con otras features

En este caso de experimentación nos centraremos en utilizar algunas de las características principales de los países, como su ubicación, continente, estatus y porcentaje de gasto en salud para explicar su esperanza de vida. Para lograrlo agregamos 3 nuevas variables a nuestro dataset: Longitud, Latitud y Continente. Creemos que el continente donde se ubica un país puede explicar varias cosas sobre el. Si pensamos en un país ubicado en el hemisferio norte del mundo, es probable que demos con uno desarrollado. Mientras que si, por ejemplo, pensamos en un país ubicado en África, es probable que su expectativa de vida sea menor a la media. Vamos a ver como esto se condice con los datos y el análisis de regresión.

Decidimos observar como era esta regresión originalmente sin outliers, para luego al final ver como afecta a los resultado la eliminación de los mismos.

Aclaraciones: Los datos adicionales sobre la ubicación de donde fue sacada esta información puede encontrarse en [6]

2.2.2. Relación entre expectativa de vida e indicadores sobre enfermedades virales

Estudiamos la relación entre expectativa de vida y los siguientes indicadores contenidos en el dataset con el que trabajamos:

- Cantidad de casos reportados de sarampión (*measles*) cada mil habitantes.
- Cantidad de muertes por HIV en niños menores de 5 años por cada mil nacimientos.
- Porcentaje de inmunización en bebés de un año frente al virus de la Hepatitis B.
- Porcentaje de inmunización en bebés de un año frente al poliovirus.

No consideramos el otro indicador relacionado con enfermedades (*Diphteria, tetanus toxoid...*), que se relaciona con inmunidad frente a enfermedades de otro tipo (bacteriales) y además correlaciona fuertemente con el porcentaje de inmunización frente al polio.

Los 4 features considerados anteriormente los resumimos en uno de la siguiente manera. Normalizamos cada feature restando cada valor por la media y dividiendo por el desvío estándar, y obtenemos para cada país p el valor del nuevo indicador sumando los valores normalizados de los features que miden cantidades de casos reportados y restando los que miden inmunidad.

$$Virus(p) = \frac{Measles(p) - \mu_{Measles}}{\sigma_{Measles}} + \frac{HIV(p) - \mu_{HIV}}{\sigma_{HIV}} - \frac{HepB(p) - \mu_{HepB}}{\sigma_{HepB}} - \frac{Polio(p) - \mu_{Polio}}{\sigma_{Polio}}$$

Hacemos el análisis de regresión entre la expectativa de vida y el feature *Virus*. Luego removemos los outliers para tratar de lograr un mejor ajuste y volvemos a realizar el análisis. Finalmente, agregamos *Status* como predictor. Dado que en nuestro dataset *Status* es una variable categórica, la convertimos en *float* asignándoles 1 a los países desarrollados y 0 a los países en vías de desarrollo.

2.2.3. Relación entre expectativa de vida y tasas de suicidios

En nuestro último caso de experimentación decidimos por incluir data externa a nuestro dataframe. Con el análisis sobre enfermedades virales ya habíamos podido trabajar sólo con data del archivo .csv original. Ahora nos pareció interesante ver cómo se puede explicar la expectativa de vida de un país agregando data nueva. Primero nos pareció interesante hacerlo agregando data sobre "portación de armas por país" [7], pero no encontramos información previa al 2015. Por lo tanto descartamos ese caso. Entonces se nos ocurrió tomar como feature externa la tasa de suicidios cada cien mil habitantes por país. Este set de datos es del 2015, por lo tanto entra en el rango de años del que proviene la data de la cátedra. Agregamos tres casos en particular de este nuevo dataset:

- Cantidad de suicidios de mujeres cada cien mil habitantes.
- Cantidad de suicidios de hombres cada cien mil habitantes.
- Cantidad de suicidios totales (ambos sexos) cada cien mil habitantes.

Nos pareció interesante ver si dividir la información en estas tres categorías nos aportaría alguna revelación. ¿Estamos seguros que el ajuste hecho entre los suicidios totales y la expectativa de vida será mayor si se realiza con el caso de un solo sexo? ¿Las personas de qué sexo cometen más suicidios en general? ¿En qué países hay más suicidios?¿En los que están encasillados con el estatus "*Developing*"?¿O en los de estatus "*Developed*"? Y por último, tal vez lo que más nos interesaba antes de realizar la experimentación, era saber si, realmente, se puede inferir la expectativa de vida de un país teniendo sólo en cuenta la cantidad de suicidios en este. A mayor cantidad de suicidios en un país, menor la expectativa de vida? ¿O son dos características que poco tienen que ver?

La experimentación incluye un análisis exploratorio de los datos del nuevo dataframe, para saber con que nos encontramos. Al tenerlo en claro, procedimos a hacer el análisis con las variables elegidas y luego, en base a los gráficos que nos permiten detectar outliers, rehicimos el análisis del ajuste pero esta vez sin outliers (como en las experimentaciones anteriores).

Aclaraciones: Los datos adicionales sobre la ubicación de donde fue sacada esta información los tomamos de la página de la Wold Health Organization. El link se puede encontrar en [8].

3. Resultados y Discusión

3.1. Información geográfica de los países con otras features

3.1.1. Utilizando solo longitud y latitud

La calidad de la regresión en donde lo único que se tiene en cuenta es la latitud y longitud de los países nos dio resultados donde los estadísticos R^2 y R^2 ajustado nos dieron valores bastante bajos con valores 0.236 y 0.227 respectivamente y rmse = 8.009. Evidentemente, estimar la esperanza de vida de un país en base a la longitud y latitud de un país no es el mejor modelo. En la figura 7 se puede observar como fueron los residuos que se generan al utilizar este modelo para explicar nuestra variable target.

Solamente de manera anecdótica y para ampliar información sobre los resultados obtenidos, la regresión nos dice que cambiar tanto la latitud como la longitud en una unidad genera una variación de la esperanza de vida de 0.1826 y -0.0168 años respectivamente

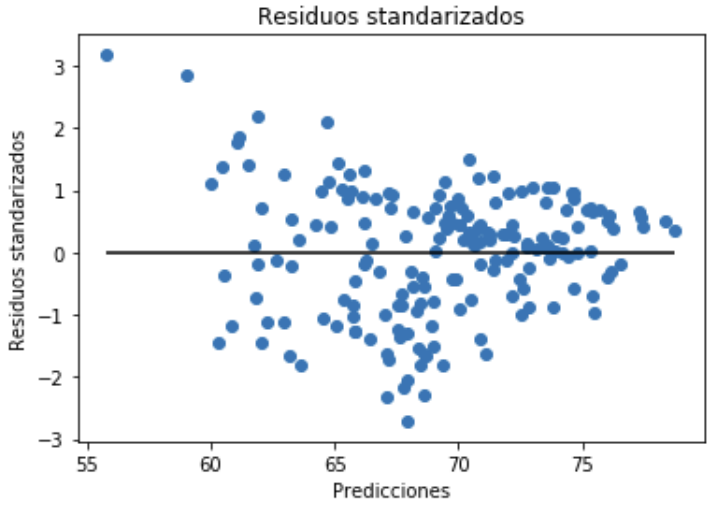


Figura 7: Gráfico de residuos en las predicciones utilizando solamente longitud y latitud

3.1.2. Utilizando solamente el continente

En búsqueda de otros modelos con mayor capacidad explicativa de nuestro fenómeno, decidimos probar si tener en cuenta la información del continente del país, en vez de la locación, podía mejorar nuestros resultados. Dado que obtuvimos bajos resultados, decidimos pasar a hacer una regresión en donde solo se tenga en cuenta el continente. Debido a que el continente de un país es un tipo de variable categórica, no existe una relación de orden entre estas, por lo cual hay que ser cuidadoso sobre como representamos esta información en nuestro dataset. La opción que nos pareció mas oportuna en este caso es el uso de variables dummies [9] para cada continente donde se indica con 1 o 0 si el país pertenece a tal continente.

En este caso los resultados fueron significativamente mejores ya que obtuvimos que $R^2 = 0.538$ y R^2 ajustado = 0.5229 rmse = 6.22225012433972. Alcanzamos a explicar la mitad de la varianza, aunque el error cuadrático medio sigue siendo alto.

En la figura 3.1.3 se puede observar como fueron los residuos que se generan al utilizar este modelo.

En la tabla 4 podemos ver los resultados de nuestra regresión donde nos explica como varia la cantidad de años de vida según el continente en donde se ubica el país. Como vimos en la exploración de datos, los países africanos tendían a tener esperanza de vida mucho menores a los demás continentes y podemos ver como esto se refleja en la regresión.

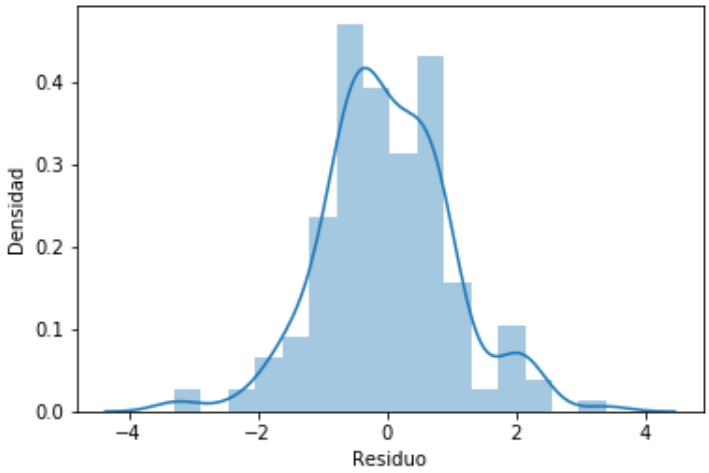


Figura 8: Gráfico de residuos en las predicciones utilizando el continente

Continente	Variación de esperanza de vida
const	60.5926
Africa	-1.2161
Asia	8.9309
Europe	16.2993
North America	13.8074
Oceania	10.6218
South America	12.1493

Cuadro 4: Variación en la diferencia de cantidad de años vida estimados según el continente

3.1.3. Utilizando continente y status

En este caso se agrega la variable de status al análisis. Se obtuvieron mejoras, pero creemos que no fueron muy significativas ya que obtuvimos que $R^2 = 0.648$ y R^2 ajustado 0.636 rmse = 5.4387, lo cual es una mejora a lo obtenido anteriormente.

De hecho, si observamos la figura y 9 podemos ver sus residuos son bastante similares.

Creemos que obtuvimos este tipo de resultados ya que la división entre países desarrollados y no desarrollados esta muy ligada a la división entre continentes.

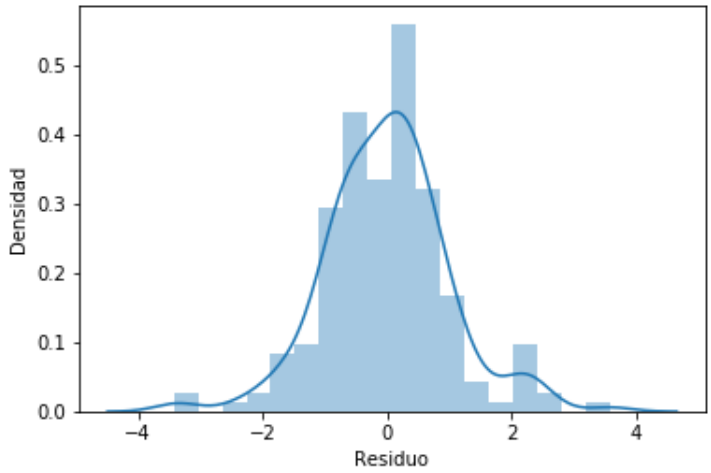


Figura 9: Gráfico de residuos en las predicciones utilizando el continente y status

Aclaración: Re-categorizamos de forma que considerábamos correcta a los paises que vimos mal clasificados en la sección 2.1.4 en este experimento.

3.1.4. Utilizando continente y percentage expendure

en este experimento. Esta vez optamos por agregar información sobre los gastos de los países en salud ya que creemos que si bien probablemente exista una relación entre este gasto y el continente, creemos que utilizarlo puede darnos una explicación mejor que tan solo usando el status.

Lamentablemente, esto no sucedió ya que obtuvimos que $R^2 = 0.545$, R^2 ajustado $= 0.5225$ y rmse $= 5.6637$. No pudimos hallar una explicación de porque obtuvimos peores resultados.

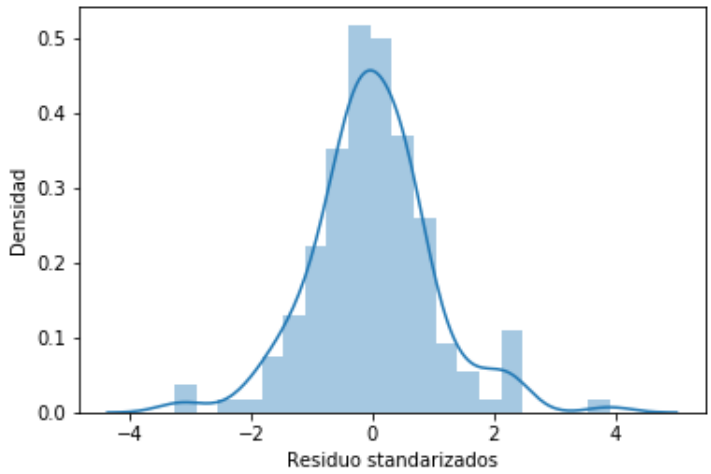


Figura 10: Gráfico de residuos en las predicciones utilizando el continente y percentage expenditure

Aclaración: Debido a que en la sección 2.1.2 vimos que esta feature tenia datos mas alla del limite, decidimos sacar del análisis a aquellos países cuyo valor sea mayor a 1000 ya que los consideramos erróneo.

3.1.5. Utilizando continente y percentage expendure pero sin outliers

Con los malos resultados de la sección 3.1.4, quisimos ver que tanto afectaban los datos outliers en la calidad de nuestra regresión. Por lo tanto, vamos a repetir el experimento pero esta vez sacando los datos que consideramos outliers.

Veamos primero con un Influence plot que se puede visualizar en la figura 11 que resume los 3 estadísticos que vamos a utilizar para quitar los outliers: Residuos studentizados, Leverage y distancia de Cook

Nuestro criterio de corte para quitar datos fue eliminar aquellos elementos que cumplan alguna de estas condiciones:

- El residuo standarizado sea mayor en modulo a 2 desviaciones standard.
- El valor del H levarage sea mayor a 0.11
- Su distancia de cook sea mayor a 0.05.

Después de hacer este filtrado, 22 países fueron eliminados.

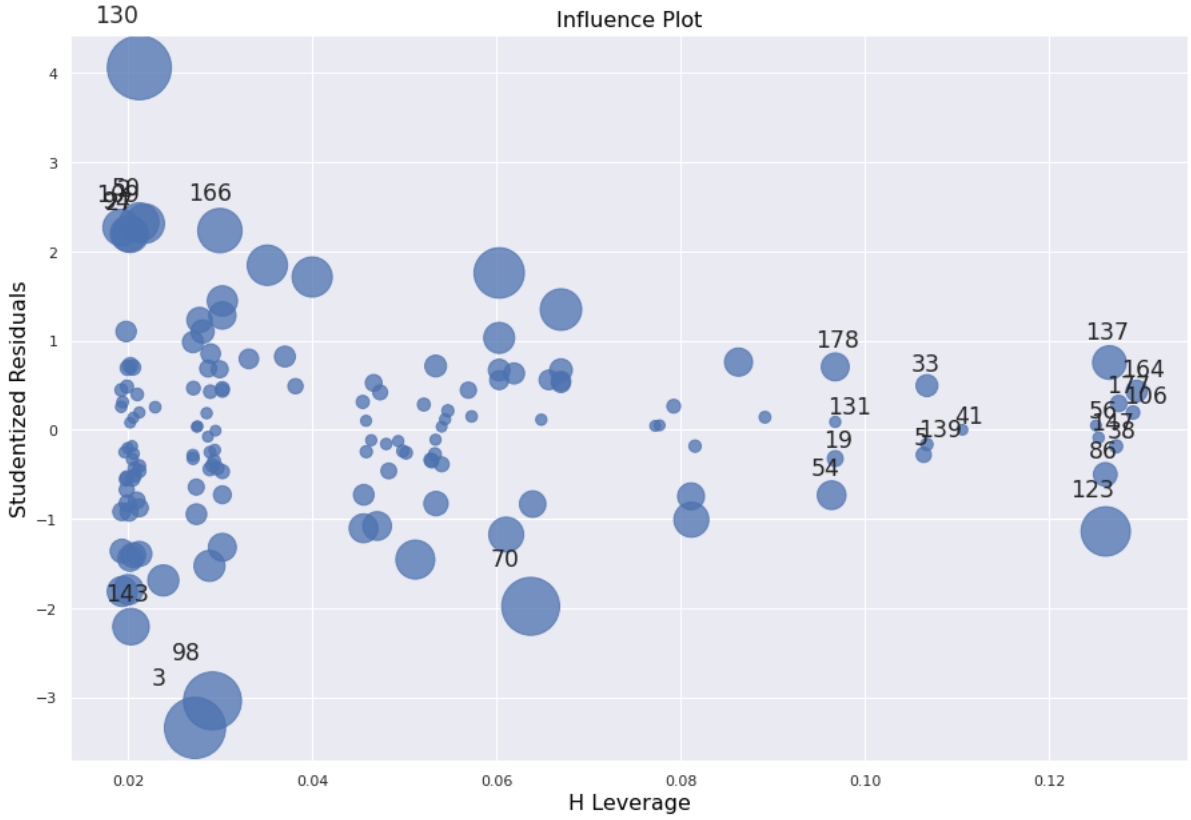


Figura 11: Influence Plot. El tamaño de los círculos representa la distancia de Cook

Debido a que sacamos varios datos outliers, esta vez obtuvimos mejores resultados alcanzando lso valores R^2=0.7504, R^2 ajustado=0.7360 y rmse= 4.096

Aclaración: Para reproducir este experimento correr el notebook Exp-Locacion.ipynb

3.2. Relación entre expectativa de vida e indicadores sobre enfermedades virales

En la Figura 12 mostramos un scatter plot de la expectativa de vida frente al índice que definimos para resumir la información relacionada sobre la prevalencia de enfermedades virales. Podemos apreciar tanto una correlación negativa entre los datos como la presencia de outliers.

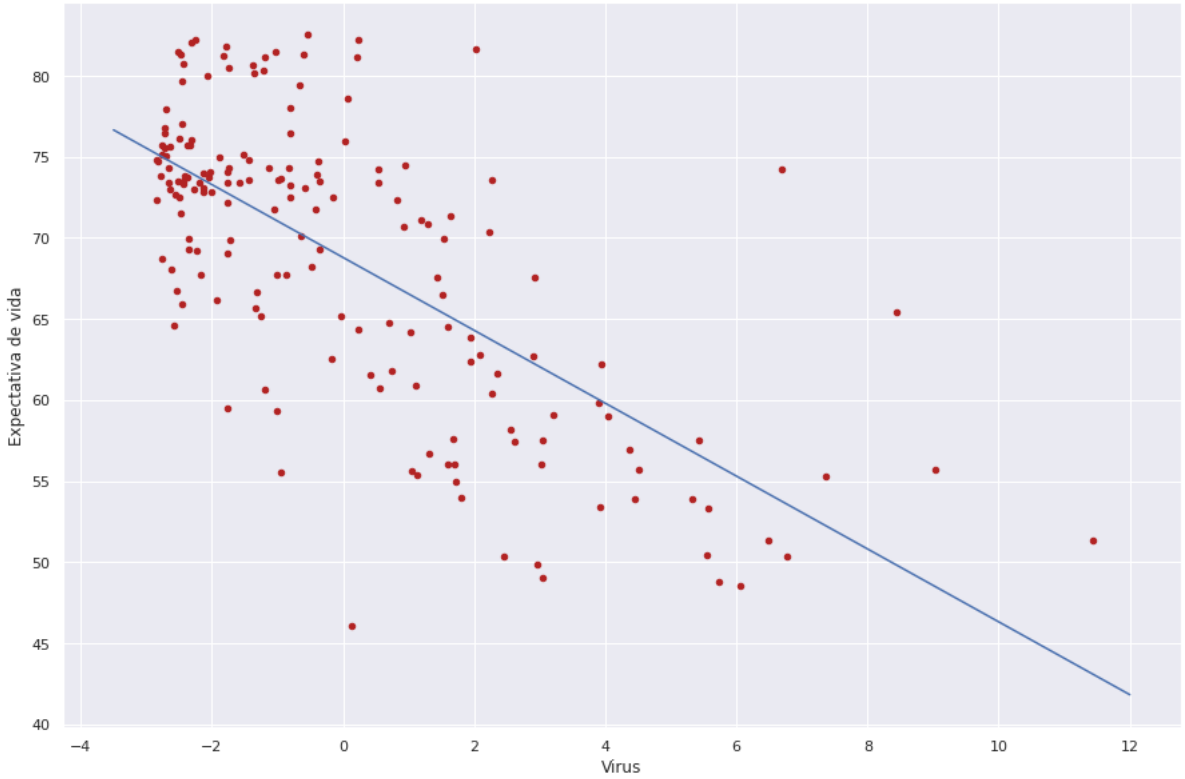


Figura 12: Scatter plot y regresión lineal incluyendo outliers para los features Expectativa de vida y Virus.

Podemos identificar a los outliers fácilmente a partir del Influence Plot de la Figura 13, que resume los 3 estadísticos más importantes que tenemos para esto: Residuos studentizados, Leverage y distancia de Cook. Este último se representa en el gráfico con el tamaño de cada círculo.

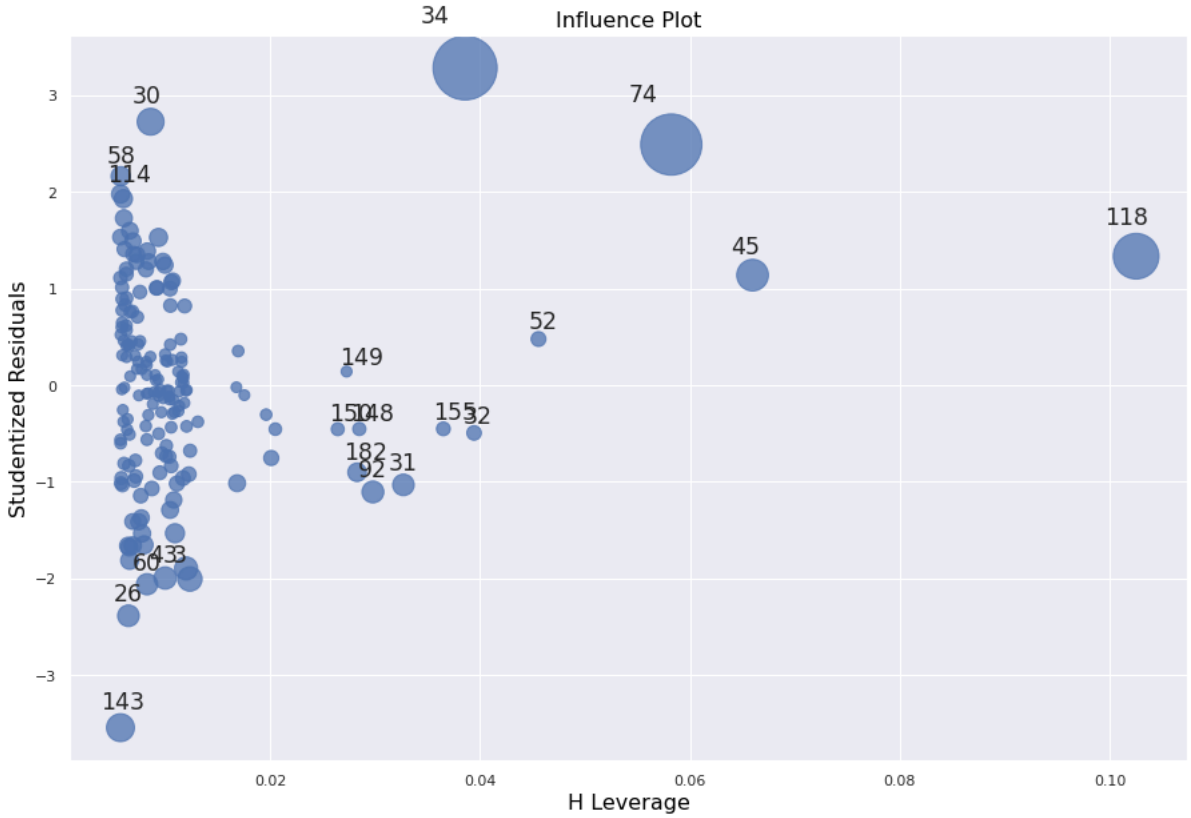


Figura 13: Influence Plot. El tamaño de los círculos representa la distancia de Cook.

Los países con mayor distancia de Cook son:

- 34: China
- 74: India
- 118: Nigeria

Mayor Leverage:

- 118: Nigeria
- 45: Costa de Marfil
- 74: India
- 52: Guinea Ecuatorial

Mayores residuos:

- 143: Sierra Leona
- 30: Canadá
- 26: Burundi
- 60: Gambia
- 58: Francia

Excluimos a estos 10 países de nuestro dataset y volvemos a realizar el ajuste. En la Figura 14 se muestra el scatter plot sin los outliers y la nueva recta de regresión. Incluimos también en el gráfico la recta calculada anteriormente para poder analizar el cambio. Podemos ver que la nueva recta tiene una pendiente más negativa (correlaciona con más fuerza) y que el cambio es considerable si tenemos en cuenta que sólo eliminamos 10 puntos.

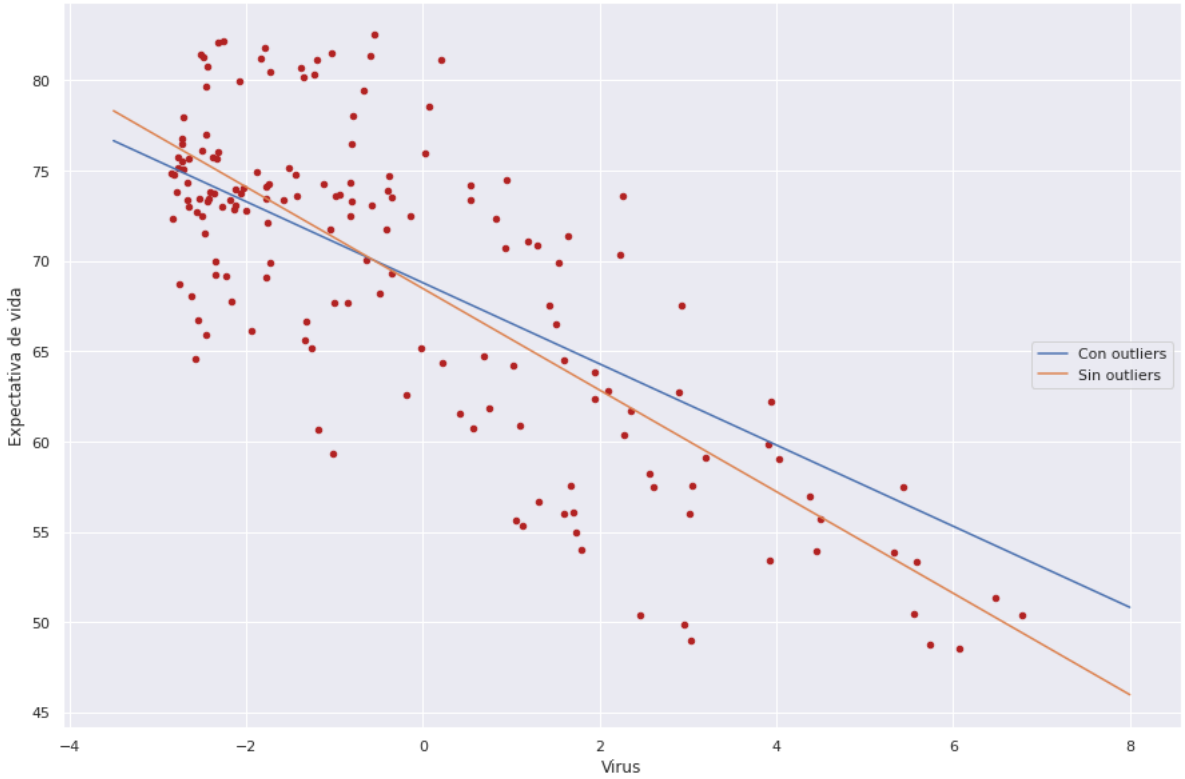


Figura 14: Scatter plot sin outliers, junto con las rectas de regresión lineal calculadas con y sin outliers.

En el Cuadro 5 mostramos los valores de los estadísticos R^2 y R^2 ajustado correspondientes a las dos regresiones anteriores, y a un tercer ajuste en el que incluimos además como predictor el feature Status. Podemos ver que logramos una mejora significativa al eliminar outliers, y que la inclusión de Status también contribuye a explicar la expectativa de vida de los países.

Aclaración: Para reproducir este experimento correr el notebook Exp-virus.ipynb

Análisis de regresión	R^2	R^2 ajustado
Con outliers	0.477	0.474
Sin outliers	0.604	0.602
Sin outliers y con *Status*	0.675	0.671

Cuadro 5: Estimadores R^2 y R^2 ajustado para los distintos análisis de regresión

3.3. Relación entre expectativa de vida y tasa de suicidios

3.3.1. Conociendo la data nueva

Primero hay que chequear con qué data nos encontramos. Realizamos varios histogramas para conocer el comportamiento de los datos con nuestro dataset original. Parece bastante aceptable la escala. Aunque vemos que hay un caso en particular que se va muy por encima de la media. Este país, caracterizado con el índice 92, es *Lesotho*. Podríamos decir que ni la data oficial se escapa de casos súper atípicos. Durante las experimentaciones previas nos habían entrado dudas sobre la veracidad de ciertos datos extremos, pero al menos la data de suicidios nos muestras que es normal que existan. Hay que tratarlos con cuidados. Ya el hacer estos histogramas nos muestra nuestro primer outlier (habrá más), el país mencionado anteriormente. Al ser tan extremo, lo quitamos. Este primer vistazo de los datos nos dio dos revelaciones. Una a favor de nuestras hipótesis y otra en contra.

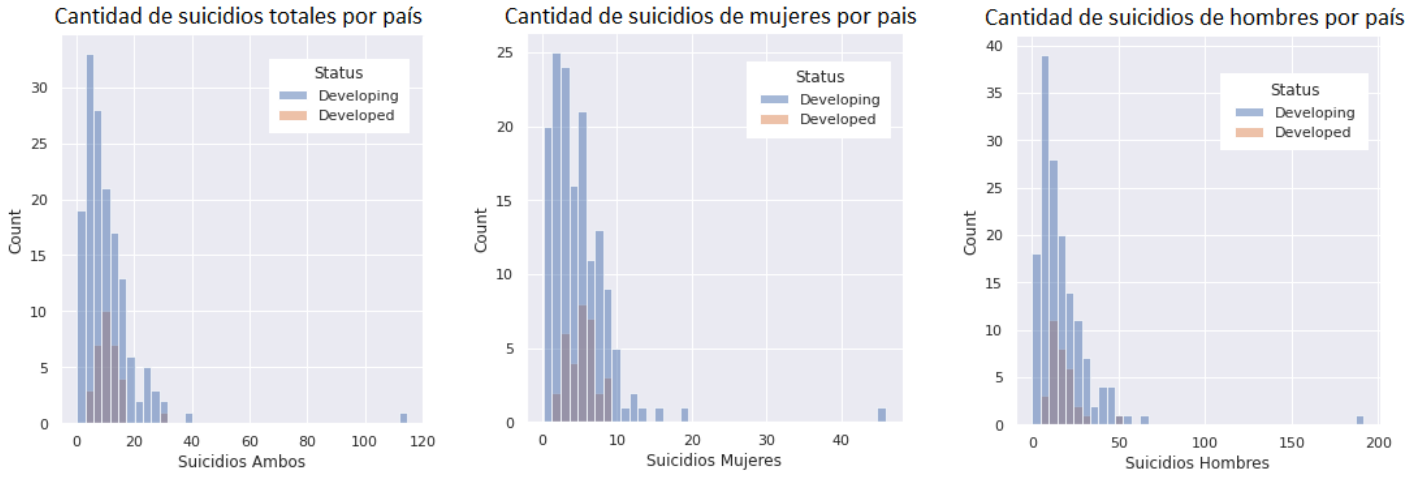


Figura 15: Histogramas de la cantidad de países y la cantidad de suicidios cada cien mil habitantes entre ambos sexos, hombres y mujeres.

Como primer hipótesis, creímos que la muestra tendría valores mucho mas grandes para los países desarrollados. Se nos había hecho común escuchar historias sobre que había grandes tasas de suicidios en países como Suecia o Japón por dar algún ejemplo. Pero viendo los datos se nos demuestra que esto no es así. No vemos una gran diferencia entre la tasa entre países desarrollados o sub-desarrollados. De hecho, entre los valores más altos, se encuentran países con el tag de "*Developing*".

Nuestra segunda hipótesis trataba sobre esta categoría trata sobre la idea de que entre los hombres se suelen cometer más suicidios que entre las mujeres. Los datos nos mostraron que esto es cierto. En el histograma de abajo claramente se puede notar una diferencia generosa entre los índices de suicidio de ambos géneros. Esto puede ser provocado por la responsabilidad que la sociedad y los hombres mismos se ponen en la cabeza sobre "proveer a la familia, ser exitoso, trabajar largas horas, etc". Además a esto se le suma el hecho de que a los hombres puede costarles más expresar sus sentimientos debido a décadas y décadas de reprimirlos (no está bien visto que lloren por ejemplo). Las mujeres suelen ser más abiertas en ese sentido, por lo tanto estos pueden ser posibles causas de los datos categóricos mostrados en el histograma.

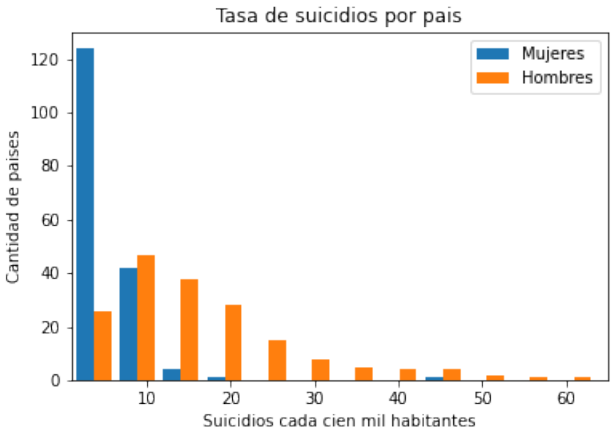


Figura 16: Histograma entre la cantidad de suicidios cada cien mil habitantes entre hombres y mujeres por país.

3.3.2. Análisis del ajuste con outliers

Normalizamos los datos de suicidios para los tres casos (ambos géneros, hombres y mujeres) y realizamos el ajuste de regresión. Los valores arrojados de ajuste son los siguientes:

- $R^2 = 0.089$ | R^2 ajustado $= 0.084$ | RMSE $= 8.645$ (para el caso de ambos sexos).
- $R^2 = 0.087$ | R^2 ajustado $= 0.082$ | RMSE $= 8.655$ (para el caso de los hombres).
- $R^2 = 0.116$ | R^2 ajustado $= 0.111$ | RMSE $= 8.515$ (para el caso de las mujeres).

Notamos valores pobrísimos en la regresión de las tres muestras. Algo inesperado, no pensábamos que diera tan poco. Esto demuestra que la expectativa de vida de un país no se relaciona para nada con la cantidad de suicidios de su población. Por lo pronto, aunque mínima, vemos una ventaja en el caso del modelo de suicidios en mujeres. Veamos qué sucede si se quitan outliers de la muestra. ¿Realmente puede haber una mejoría notable? Creemos que no.

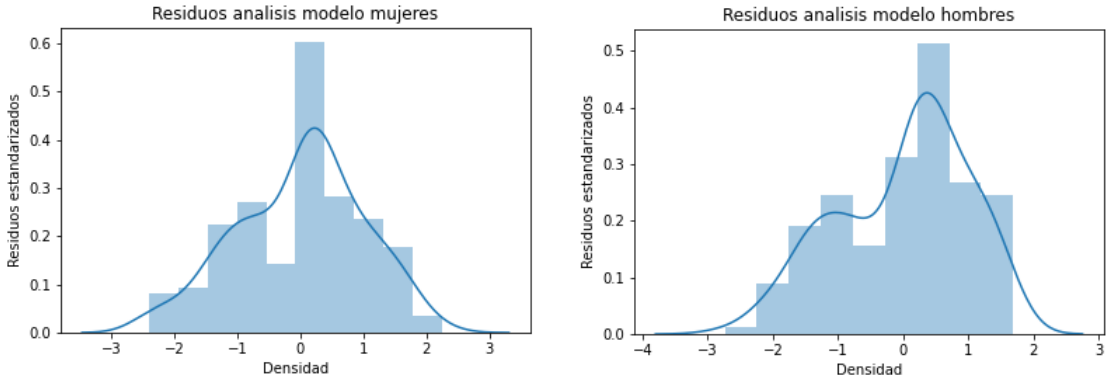


Figura 17: Histograma de los residuos (y su distribución) de cada país sobre la linea de regresión de los modelos de hombres y mujeres.

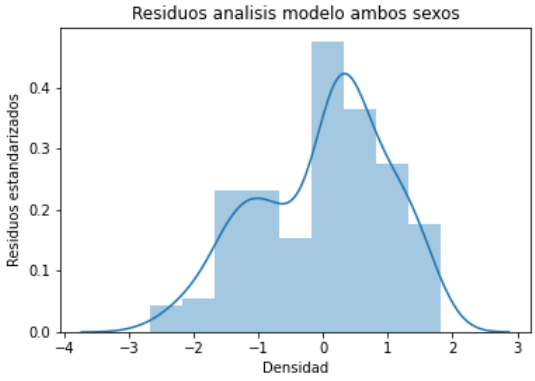


Figura 18: Histograma de los residuos (y su distribución) de cada país sobre la linea de regresión del modelo de ambos géneros.

3.3.3. Análisis del ajuste sin outliers

Sacamos outliers en los tres modelos, usando el mismo criterio de cálculo para los tres.

Primero nos centramos en la quita de residuos que tengan una desviación estándar mayor a 2 de la linea de regresión. Este suele ser el procedimiento que se deshace de más puntos atípicos.

Los países con mayor desviación en cada modelo fueron los siguientes:

- Sierra Leone con una desviación de -2.670 (para el caso de ambos sexos).
- Otra vez Sierra Leone con una desviación de -2.740 (para el caso de los hombres).
- Devuelta Sierra Leone con una desviación de -2.400 (para el caso de las mujeres).

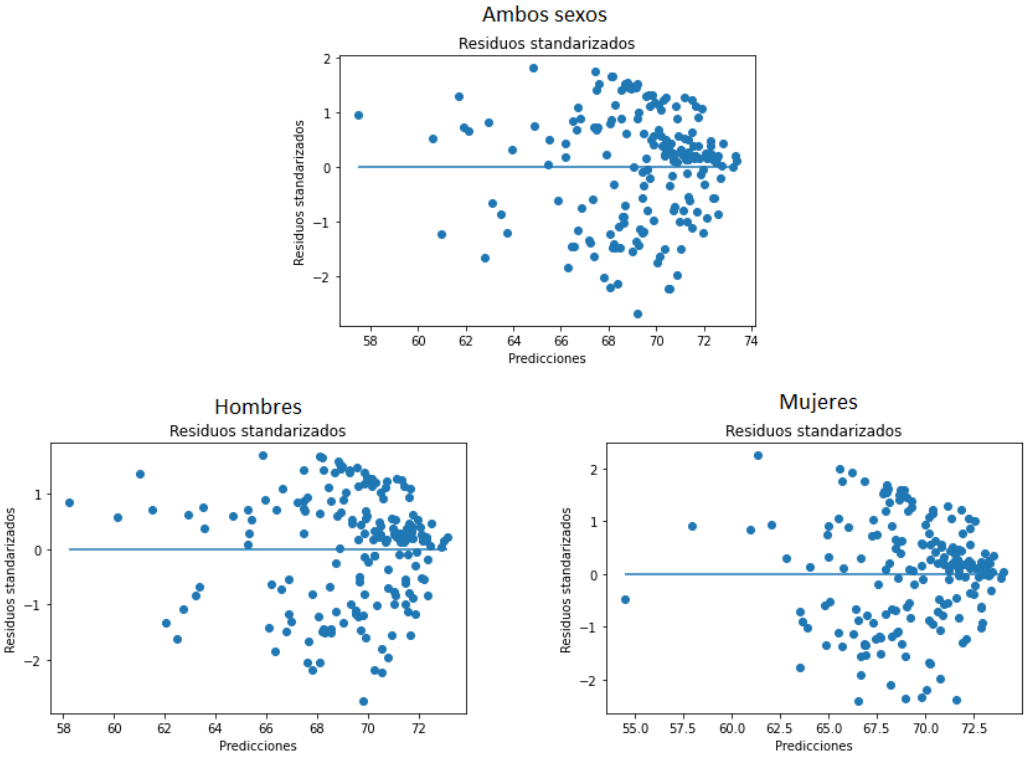


Figura 19: Gráficos de residuos de cada modelo.

Luego nos fijamos en todos aquellos países que tengan una Distancia de Cook mayor a 0.1 Pero ningún país se distanciaba tanto, por lo que no quitamos a ninguno por esta vía.

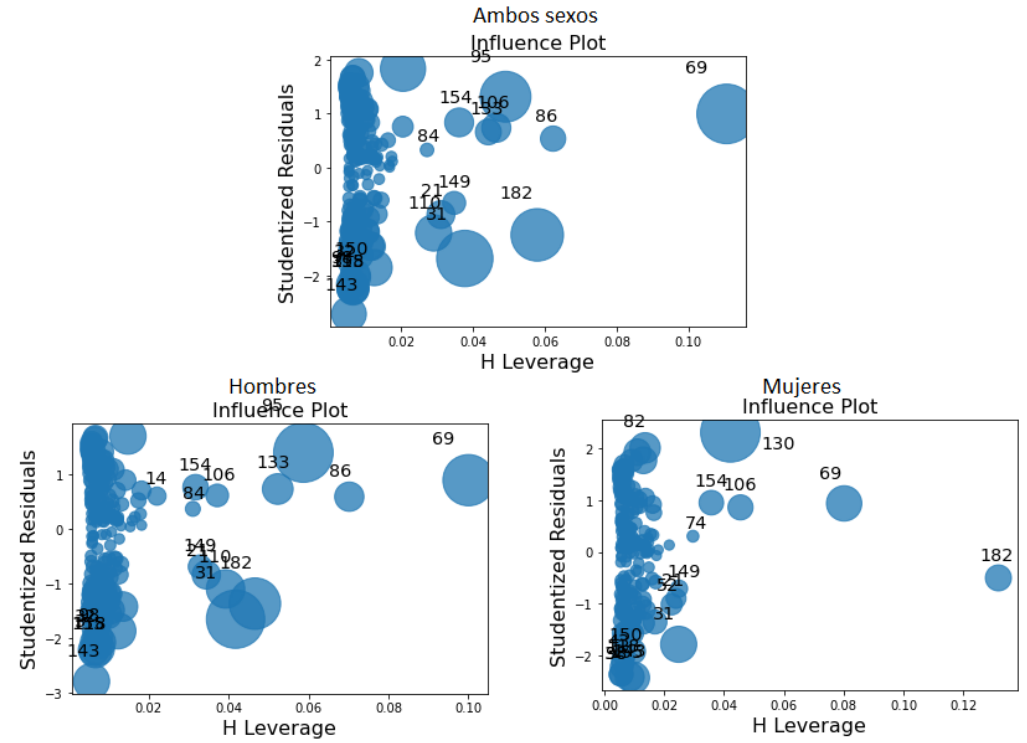


Figura 20: Influence Plot. El tamaño de los círculos representa la distancia de Cook

Por último, y para terminar así la quita de outliers, descartamos aquellas naciones que muestren un Leverage mayor a 0.06

Los países que presentaron mayor Leverage en cada modelo fueron los siguientes:

- Guyana con un apalancamiento de 0.112 (para el caso de ambos sexos).
- Guyana nuevamente, con un apalancamiento de 0.101 (para el caso de los hombres).
- Zimbabwe con un apalancamiento de 0.139 (para el caso de las mujeres).

Y así es como nos quedamos con una muestra con menos casos atípicos para cada modelo. Ahora procedemos a realizar el ajuste nuevamente. Obteniendo los siguientes resultados:

- $R^2 = 0.105$ | R^2 ajustado $= 0.100$ | RMSE $= 8.012$ (para el caso de ambos sexos).
- $R^2 = 0.102$ | R^2 ajustado $= 0.096$ | RMSE $= 8.027$ (para el caso de los hombres).
- $R^2 = 0.135$ | R^2 ajustado $= 0.130$ | RMSE $= 7.750$ (para el caso de las mujeres).

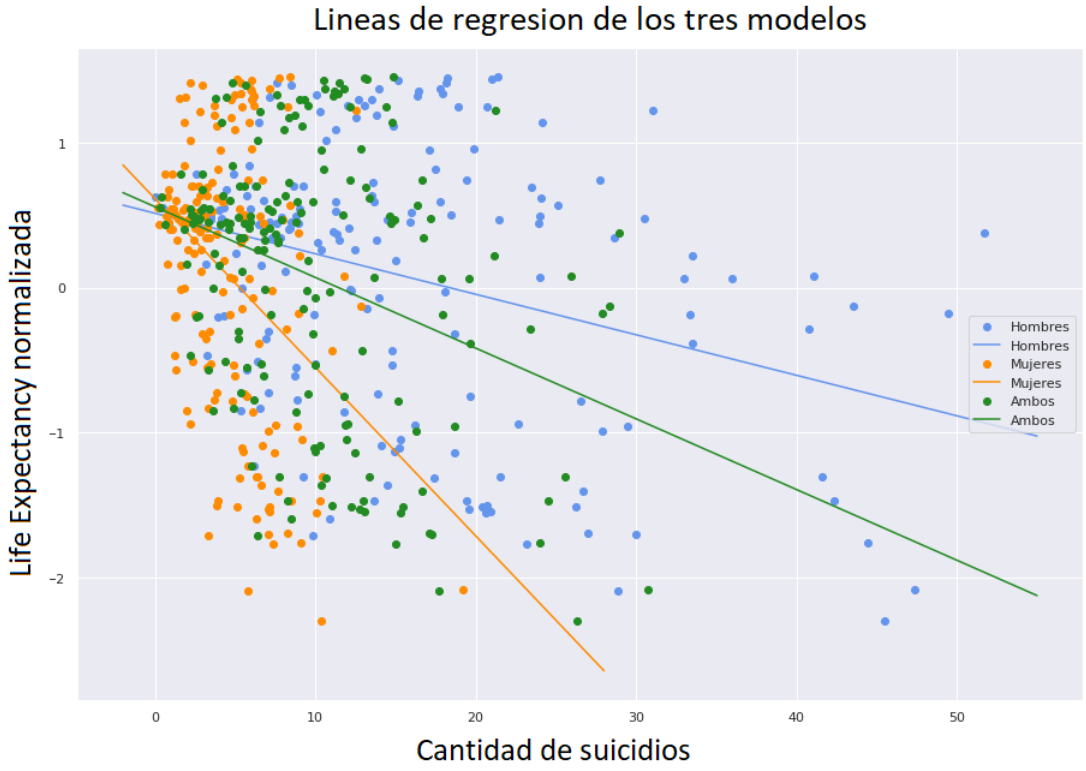


Figura 21: Gráfico de estilo plot. Las líneas son la predicción de la regresión de cada modelo (ambos géneros, mujeres y hombres).

Dan mejor pero ni por asomo obtienen el ajuste esperado. Perdimos toda esperanza de que solamente la tasa de suicidios explique la expectativa de vida de un país. Por lo tanto, al menos veamos que sucede al agregar más variables al análisis. Supusimos que al tener más contexto, el ajuste mejoraría.

3.3.4. Análisis del ajuste con nueva variable

Incluimos el estatus de un país al análisis del ajuste para ver si eso da el suficiente contexto como para que el ajuste se acerque a un valor mas aceptable (mayor a 0.5). Volvimos a realizar el análisis de regresión con nuestros tres modelos anteriores, sumándoles esta nueva variable. Los resultados que arrojaron fueron bastante mejores, pero no suficientes. Aquí están los resultados:

- $R^2 = 0.397$ | R^2 ajustado $= 0.390$ | RMSE $= 6.575$ (para el caso de ambos sexos).
- $R^2 = 0.388$ | R^2 ajustado $= 0.381$ | RMSE $= 6.623$ (para el caso de los hombres).
- $R^2 = 0.449$ | R^2 ajustado $= 0.442$ | RMSE $= 6.187$ (para el caso de las mujeres).

El ajuste va en aumento. De hecho en el caso de las mujeres se acerca considerablemente al 0.5. Tener el contexto socio-económico del país (como lo es su estatus), es de gran ayuda para que al combinarlo con la tasa de suicidios de su población, nos pueda explicar un poco más como es su expectativa de vida. Hay más contexto, pero no el suficiente.

Estos resultados nos toman por sorpresa ya que en las tres corridas de análisis la variable que trata sobre el suicidio de las mujeres fue la que dio mejores resultados en cada una de ellas. Por lo tanto, esto nos muestra que, por un mínimo margen, la tasa de las mujeres ayuda a explicar más la expectativa de vida de un país que la tasa de ambos géneros. Una verdadera revelación.

Aclaración: Para reproducir este experimento correr el notebook Exp-Suicidios.ipynb

4. Conclusiones

En este trabajo estudiamos un dataset complejo con mucha información sobre distintos países buscando distintas formas de explicar la esperanza de vida de los países, obteniendo valores muy distintos para cada una.

Vimos que es importante comprender los datos con los que estamos trabajando antes de formular conclusiones basadas en cálculos de correlaciones, porque por ejemplo, vimos que había una correlación positiva entre el consumo de alcohol y la expectativa de vida, pero sabemos que no es razonable establecer una relación causal entre las dos cosas y sólo pudimos explicar esa correlación al estudiar mejor las características de los países que suministraban esos datos (muchos países con muy baja esperanza de vida tienen consumo cercano a 0 porque la religión predominante lo prohíbe y, por otra parte, los países europeos consumen mucho).

Vimos además que las medidas de resumen como la media y la mediana no siempre son útiles para comprender mejor lo que estamos estudiando, ya que en este caso nos encontramos con dos grupos bien diferenciados de países con expectativas de vida muy diferentes.

Para generar nuestro primer modelo, utilizamos datos geográficos y económicos obteniendo de esta manera los mejores resultados posibles utilizando como métricas los estadísticos R^2, R^2 ajustado y RMSE. Vimos que utilizar la longitud y latitud de los países obtiene resultados significativamente mucho peores que utilizando el continente al que pertenecen. Observamos que en conjunto a los continentes, el tener en cuenta el status de un país genera mayor información que saber el porcentaje que emplean en salud.

Después, se utilizo información sobre distintas enfermedades que sufre la población para crear una nueva feature que llamamos *Virus* y vimos como la variación en esta nueva variable cambiaba la estimación en la esperanza de vida.

Por ultimo, estudiamos como la tasa de suicidios en una población podría explicar la esperanza de vida de un país. No obtuvimos ajustes con buen valor, por lo tanto quedó demostrado que la tasa de suicidios de un país no nos sirve para explicar la esperanza de vida. Al menos no como única variable. Otra conclusión que podemos sacar de esta experimentación es que la tasa de suicidio de mujeres es la variable mas representativa entre los tres modelos elegidos. Pudiendo explicar mejor la expectativa de vida que , por ejemplo, la tasa de suicidios de toda la población. Se percibió una mejora en el ajuste al agregar otra variable, el estatus de cada país. Por lo que la tasa de suicidios funciona mejor con más contexto de cada país.

En todos los casos vimos que la eliminación de datos atípicos generó mejoras en nuestros análisis,

5. Referencias

[1] https://support.minitab.com/es-mx/minitab/18/help-and-how-to/modeling-statistics/regression/supporting-topics/basics/what-are-categorical-discrete-and-continuous-variables/

[2] https://worldpopulationreview.com/country-rankings/developed-countries

[3] https://www.nhs.uk/common-health-questions/lifestyle/what-is-the-body-mass-index-bmi/

[4] https://www.who.int/hiv/HIVCP_SWZ.pdf

[5] https://www.usaid.gov/south-sudan/education

[6] https://github.com/andresmbar/tp3-metodos/blob/master/notebooks/country-and-continent-codes-list.csv
https://github.com/andresmbar/tp3-metodos/blob/master/notebooks/average-latitude-longitude-countries.csv

[7] https://dataunodc.un.org/content/firearms

[8] https://www.who.int/data/gho/data/themes/mental-health/suicide-rates

[9] https://stats.idre.ucla.edu/other/mult-pkg/faq/general/faqwhat-is-dummy-coding/: :text=Dummy