PEC 2 – APRENDIZAJE Y ANÁLISIS ESTADÍSTICO

El objetivo de este documento es el de trabajar con un problema de libre elección en el ámbito de la inferencia estadística prestando especial atención a los aspectos formales del análisis. El documento se compone de cuatro puntos que comprenden el planteamiento del problema, la hipótesis de partida sobre el problema planteado, el diseño experimental y metodología seguida, modelo elaborado, así como sus resultado y conclusiones.

1. Planteamiento del problema e Hipótesis

Uno de los aspectos primordiales en la vida de los seres humanos es la salud, y la esperanza de vida es la métrica clave para evaluar la salud de la población, ya que captura la mortalidad a lo largo de todo el ciclo de vida, pero ¿qué significa exactamente la esperanza de vida?

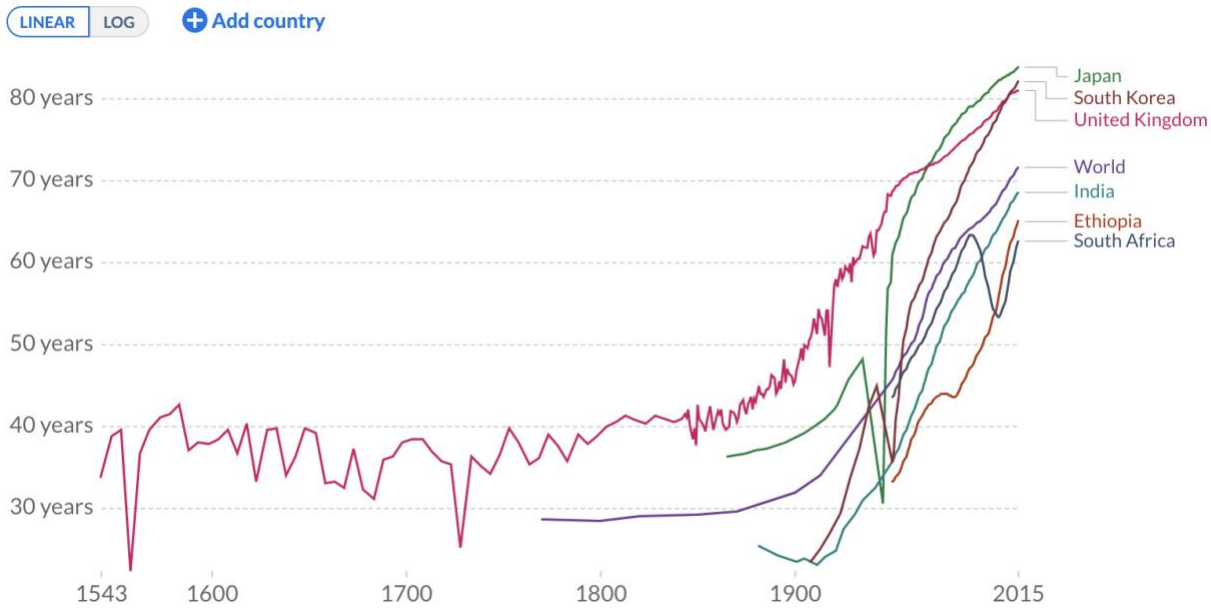
A pesar de la importancia y prominencia en la investigación y en la política, a veces es sorprendentemente difícil encontrar una descripción simple pero detallada de lo que realmente significa. El término esperanza de vida se refiere a la cantidad de años que una persona puede esperar vivir. Por definición, la esperanza de vida se basa en una estimación de la edad promedio que tendrán los miembros de un grupo de población particular cuando mueran.

El dato que se calcula habitualmente y al que se refiere popularmente como esperanza de vida consiste en estimar la duración media de vida, desde el nacimiento hasta la muerte, que está expuesta a las tasas de mortalidad observadas en un período determinado, normalmente un año. Este enfoque conduce a lo que se conoce como "período de esperanza de vida" que es la métrica de esperanza de vida más utilizada, y es la definición de la que hacen uso la mayoría de las organizaciones internacionales.

Un punto importante a tener en cuenta al interpretar las estimaciones de esperanza de vida es que muy pocas personas morirán exactamente a la edad indicada por esta, incluso si los patrones de mortalidad se mantienen constantes.

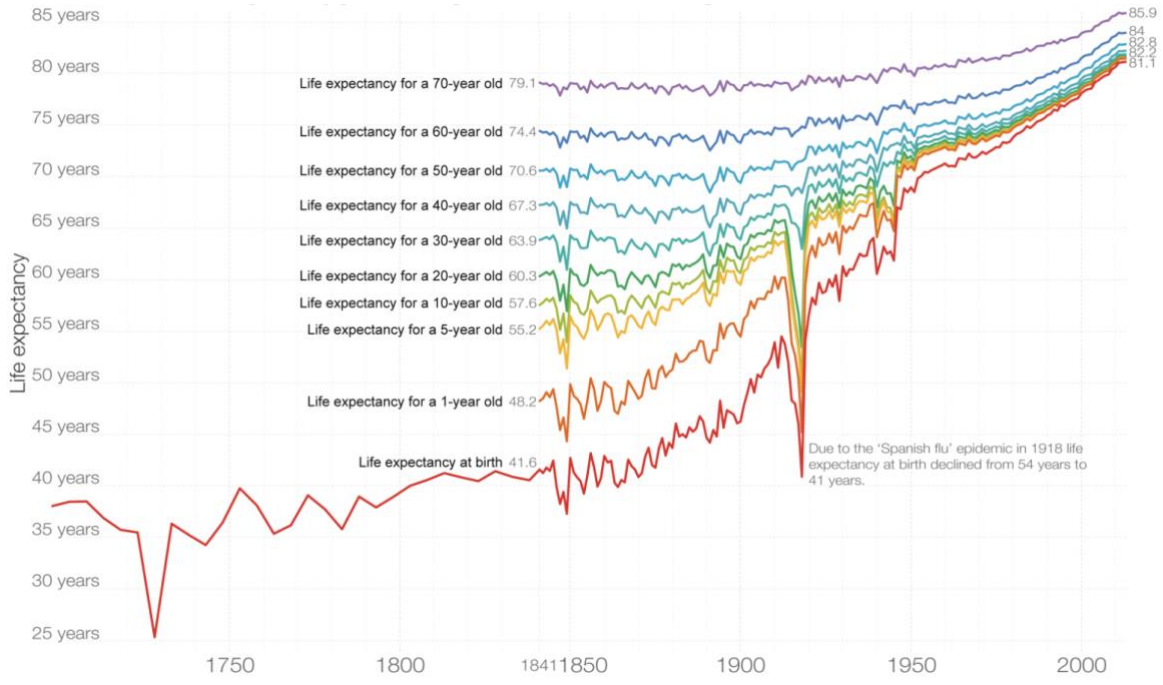
Expuesto el concepto, si se observan los datos, se demuestra que la esperanza de vida ha aumentado drásticamente desde el siglo XIX en los primeros países industrializados mientras que se mantuvo baja en el resto del mundo. Esto condujo a una desigualdad muy alta en la distribución de la salud en el mundo. Buena salud en los países ricos y mala salud en los países que seguían siendo pobres. A pesar de que sigue existiendo una amplia desigualdad en cuanto a la salud en el mundo actual, esta es menor que en el siglo XIX.

Desde 1900, la esperanza de vida media mundial ha aumentado drásticamente. El siguiente gráfico muestra la evolución de la esperanza de vida en diferentes países y a nivel mundial.



Fuente: www.ourworldindata.org

El gráfico nos muestra como en cuestión de dos siglos la esperanza de vida ha sufrido un incremento de más del doble de su valor. Estas mejoras son una clara señal histórica de progreso, de mejoras sostenidas en la salud de poblaciones enteras después de milenios de estancamiento en terribles condiciones de salud que conducían a una esperanza de vida muy baja. Es una evidencia que en los últimos se han producido grandes mejoras en el ámbito de la salud y tecnológico que permiten que la esperanza de vida mundial haya aumentado a todas las edades:



Fuente: www.ourworldindata.org

Por último, otro punto a tener en cuenta es que las estimaciones de la esperanza de vida son medidas estadísticas y no tienen en cuenta ningún factor específico de la persona, como las opciones de estilo de vida. En términos prácticos, estimar la esperanza de vida implica predecir la probabilidad de sobrevivir años sucesivos de vida, sobre la base de las tasas de mortalidad específicas por edad observadas.

Claramente, la duración de la vida de una persona promedio no es muy informativa sobre la duración de vida prevista para una persona que lleva un estilo de vida particularmente insalubre.

Estos comentarios conducen a las preguntas que se pretenden analizar en este documento:

- ¿Es posible predecir la esperanza de vida actual de un país en base a los valores de mortalidad incluyendo aspectos del ámbito de la salud?
- Si excluimos los valores de mortalidad, ¿es posible predecir la esperanza de vida actual?

En el primero de los casos, es esperable que sea posible predecir la esperanza de vida a partir de los valores de mortalidad junto con variables que tengan una mayor relación con la salud o el estilo de vida medio de un país ya que la esperanza de vida se calcula a partir de las tasas de mortalidad, pero en el segundo caso, al excluir los factores de mortalidad, la capacidad predictiva de un modelo debería disminuir complicando la explicabilidad de la variable dependiente a partir de las independientes. Por último, es necesario comentar que el interés de este documento es analizar la esperanza de vida actual, por lo que se intentará trabajar con los datos más actuales que sea posible.

2. Diseño Experimental y Metodología

En primer lugar, se ha de comentar la extracción de los datos y qué contienen los archivos de los que se va a hacer uso. Se ha hecho uso de dos fuentes datos con dos modelos de extracción diferentes:

- El primer csv del que se hace uso se obtiene de la plataforma Kaggle. El dataset contiene la esperanza de vida de una gran cantidad de países durante los años 2000-2015 junto con una serie de factores que se analizarán durante el transcurso del documento. Como se ha mencionado en el apartado anterior, se busca el análisis de la esperanza de vida desde un marco de actualidad, por lo que se limitará el uso a los datos del año 2015. El dataset se puede descargar directamente de la página web en formato csv.

- El segundo csv es extraído debido al interés de incluir los valores de población, densidad poblacional y edad media para observar si tienen significancia en el modelo. Para ello se utiliza la técnica webscrapping para descargar los datos de la página web de worldometers a través de Python con la ayuda de la librería BeautifulSoup en un jupyter notebook.

Tras la extracción de datos se realiza un preprocesamiento de los datos que se divide en dos partes. La primera consiste en la continuación del notebook en Python. Los datos obtenidos son del año 2020, pero los que se van a analizar son los del año 2015, por lo que es necesario realizar alguna modificación. Las variables interesantes para este proceso son la población en 2020, la superficie geométrica, la edad media y el cambio porcentual de la población entre los años 2019 y 2020. Para obtener los datos deseados, se considera que el cambio porcentual en la población es constante, por lo tanto, es el mismo que se produce cada año hasta 2015, y de esta forma se puede obtener la población en tal año con la siguiente fórmula:

$$Pob2015 = Pob2020 * \left(1 - \left(\frac{YearlyChange}{100}\right)\right)^5$$

Mediante la aplicación de la fórmula anterior en Python se obtiene la población del año 2015, que, dividida por la superficie del país, que se mantiene constante, da como resultado la densidad poblacional. Finalmente se considera que la edad media es constante, se eliminan el resto de las variables y se guarda el dataframe en formato csv para proseguir el análisis en RStudio.

Para continuar con el preprocesamiento, hay que comenzar con el proceso de limpieza de los datasets, recordando nuevamente que el primer csv recoge datos de cada país durante los años que transcurren entre 2000 y 2015. Como el análisis se va a realizar únicamente sobre el año 2015, se filtra por ese valor para obtener los datos de ese año.

Una vez abiertos los datasets, se unen ambos datasets por la variable país, y se empieza por comprobar que variables tienen valores nulos, y de estas cuáles resultan interesantes para el

modelo y necesitan de un análisis particular de la variable. Las variables que contienen valores nulos son:

```
> colnames(df)[colSums(is.na(df)) > 0]
 [1] "Alcohol"                         "Hepatitis B"
 [3] "BMI"                             "Total expenditure"
 [5] "GDP"                             "Population"
 [7] "thinness  1-19 years"            "thinness 5-9 years"
 [9] "Income composition of resources" "Schooling"
```

El dataset resultante de la unión contiene 176 registros, que es un número bajo. Por lo tanto, para el tratamiento de los valores nulos, si existe una cantidad superior al 5% de los registros, y no hay posibilidad de conseguir los datos de otros años porque la cantidad de valores nulos en esas variables es alto en el conjunto de todos los años, se eliminará la variable. Se han estudiado otras alternativas que se usan de forma frecuente en el tratamiento de valores nulos como asignar la media de los valores de la variable a los registros con valor nulo, pero no sería muy apropiado ya que la gran mayoría de valores nulos pertenecen a países pequeños y recónditos de los que ha sido difícil conseguir los datos, por lo que posiblemente la media no se ajuste a la situación real del país para esa variable, pero tampoco se le puede asignar un valor aleatorio o el de otro registro por simple suposición de similitud con otro país, por lo tanto, para evitar que afecte al modelo negativamente, se eliminará la variable.

Expuesta la metodología a seguir, se analiza cada variable de forma individual:

- Alcohol: en este caso, se muestra que es en particular el año 2015 es el que tiene una gran cantidad de valores nulos, por lo que se recogen los datos del año 2014 y se asume que se mantienen constantes en el siguiente año, consiguiendo un dataset con un único valor nulo.
- Hepatitis B: hay una gran cantidad de valores nulos en el dataset global y un porcentaje menor en el año 2015 en particular, pero se decide eliminar la variable por lo expuesto en la metodología a seguir con los valores nulos.
- BMI: únicamente hay 2 valores nulos por lo que se mantiene la variable.
- Total Expenditure: caso similar a la variable alcohol con una gran cantidad de valores nulos en el año 2015, pero no en el 2014 por lo que se asumen que son constantes y se obtiene que solo hay 1 valor nulo en el dataset final para esta variable
- GDP: situación similar a la variable Hepatitis B. Gran cantidad de valores nulos tanto en el año 2015 como en el dataset con todos los años.
- Population: se ha obtenido los valores de la población por otro dataset por lo que esta variable se puede eliminar.
- Thinness 1-19 years: únicamente hay 2 valores nulos por lo que se mantiene la variable.
- Thinness 5-9 years: únicamente hay 2 valores nulos por lo que se mantiene la variable.
- Income composition of resources: de nuevo es una situación simular a la de la variable Hepatitis B por lo que se decide prescindir de esta variable.
- Schooling: situación simular a la de la variable anterior por lo que se decide prescindir de esta variable.

Tras haber analizado las variables con valores nulos, se eliminan del dataset del año 2015 las variables de las que se ha decidido prescindir junto con aquellas variables que o bien, no aportan valor al dataset como Year, o no se ha encontrado la explicación de la variable como under-five deaths. Posteriormente se eliminan las filas con registros nulos de forma que se obtiene un dataset sin valores nulos habiendo perdido únicamente tres registros.

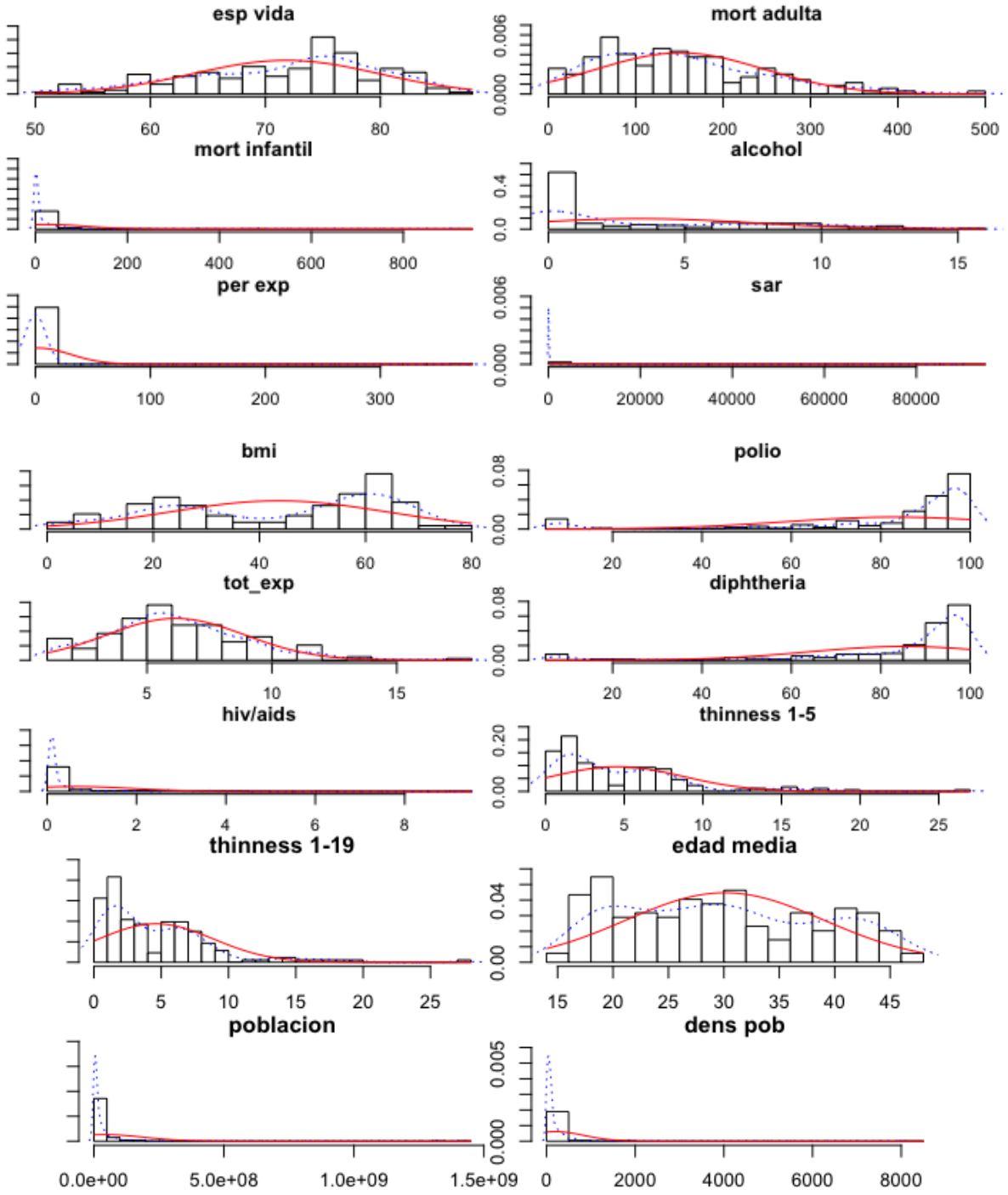
Una vez se obtiene el dataset sin valores nulos, se ha de modificar el tipo de la variable *Status*, ya que es una variable categórica porque el país puede estar desarrollado o en vías de desarrollo, y también se mapean sus valores de forma que en caso de que esté desarrollado se le otorga valor 1 y en caso de que esté en vías de desarrollo sería valor 0. También se modifica el tipo de la variable *Med. Age* convirtiendo de character a numeric. Finalmente, se realiza un cambio de nombres de las variables para ayudar a la visualización de futuros gráficos.

Una vez se han seleccionado las variables que se consideran interesantes, es recomendable llevar a cabo una tarea de análisis descriptivo de las variables. Para ello se genera un resumen de valores estadísticos de cada variable y se crean histogramas para cada una de ellas.

El resumen de estadísticas es:

```
> summary(df)
     pais           des          esp           mort_a        mort_i
 Length:173         0:141   Min.   :51.0   Min.   :  1   Min.   :  0.00
 Class :character   1: 32   1st Qu.:66.0   1st Qu.: 74   1st Qu.:  0.00
 Mode  :character           Median :74.0   Median :138   Median :  2.00
                            Mean   :71.8   Mean   :151   Mean   : 24.28
                            3rd Qu.:77.0   3rd Qu.:199   3rd Qu.: 17.00
                            Max.   :88.0   Max.   :484   Max.   :910.00
      alc            per_exp            sar             bmi             pol
 Min.   : 0.010   Min.   :  0.000   Min.   :    0   Min.   : 2.50   Min.   : 5.00
 1st Qu.: 0.010   1st Qu.:  0.000   1st Qu.:    0   1st Qu.:24.30   1st Qu.:83.00
 Median : 0.400   Median :  0.000   Median :   17   Median :51.10   Median :93.00
 Mean   : 3.336   Mean   :  2.522   Mean   : 1516   Mean   :43.62   Mean   :83.14
 3rd Qu.: 6.740   3rd Qu.:  0.000   3rd Qu.:  195   3rd Qu.:61.70   3rd Qu.:97.00
 Max.   :15.190   Max.   :364.975   Max.   :90387   Max.   :77.60   Max.   :99.00
    tot_exp           diph          hiv_aids          th5              th19
 Min.   : 1.210   Min.   : 6.00   Min.   :0.1000   Min.   : 0.100   Min.   : 0.100
 1st Qu.: 4.380   1st Qu.:84.00   1st Qu.:0.1000   1st Qu.: 1.500   1st Qu.: 1.400
 Median : 5.790   Median :93.00   Median :0.1000   Median : 3.300   Median : 3.400
 Mean   : 6.147   Mean   :85.18   Mean   :0.6283   Mean   : 4.569   Mean   : 4.617
 3rd Qu.: 7.560   3rd Qu.:97.00   3rd Qu.:0.4000   3rd Qu.: 6.500   3rd Qu.: 6.500
 Max.   :17.140   Max.   :99.00   Max.   :9.3000   Max.   :26.700   Max.   :27.300
     ed_med            pob                d_pob
 Min.   :15.00   Min.   :9.388e+04   Min.   :   2.0
 1st Qu.:22.00   1st Qu.:2.897e+06   1st Qu.:  33.0
 Median :30.00   Median :9.297e+06   Median :  79.0
 Mean   :30.09   Mean   :4.082e+07   Mean   : 194.8
 3rd Qu.:38.00   3rd Qu.:2.783e+07   3rd Qu.: 150.0
 Max.   :48.00   Max.   :1.411e+09   Max.   :8033.0
```

Para la creación de los histogramas por variables se generan en dos grupos de 6 y uno de 4 para ayudar a la visibilidad de los mismos, teniendo en cuenta que las variables país y status se obvian ya que el histograma no va a aportar ningún valor adicional a los resultados presentados en el resumen de estadísticas anterior.



Tras los datos expuestos, es necesario comentar las conclusiones que se puedan obtener de las variables, de nuevo obviando los valores de las variables país y el desarrollo del país:

- Esperanza de vida: es la variable a predecir y se observa como los datos están distribuidos de una forma que se podría considerar simétrica, con valores de mediana y media muy similares.
- Mortalidad adulta: los datos se encuentran bastante repartidos y se aprecia una gran diferencia entre el máximo y el mínimo.
- Mortalidad infantil: hay una gran diferencia entre la mayoría de países y unos pocos que tienen valores extremadamente altos que hacen subir la media mientras que el tercer cuartil se mantiene con un valor bajo.
- Alcohol: hay una gran concentración de países con niveles de alcoholemia bastante bajos en los que se aglutina más de la mitad.
- Percentage expenditure: parece que todos los valores se sitúan en el 0, o al menos el 75% de ellos, por lo que apenas aporta valor y se decide eliminar esta variable.
- Sarampión: hay una gran diferencia entre los valores lo que provoca mucha dispersión.
- BMI: existe una concentración mayor de valores en zonas de bmi altas, pero se encuentran distribuidos a lo largo de todo el eje.
- Polio: gran cantidad de países con valores altos de polio de forma que el 75% de los mismos se encuentra en valores superiores 82 siendo el máximo 99, pudiendo estar condicionado por la relación de países desarrollados y en vías de desarrollo ya que apenas el 18% se considera desarrollado.
- Total expenditure: los valores se encuentran bastante distribuidos sin que exista mucha diferencia entre el máximo y el mínimo.
- Diphteria: situación prácticamente idéntica a la expuesta con la variable polio.
- HIV/AIDS: en este caso, la gran mayoría de países se encuentra en valores cercanos a 0.
- Edad Media: valores distribuidos de forma simétrica, esperable ya que los valores de mediana y media son muy similares
- Delgadez de 1 a 5: gran concentración de valores bajos con unos pocos países en valores altos.
- Delgadez de 1 a 19: situación idéntica a la anterior lo que puede demostrar una gran correlación y la necesidad de eliminar una de las dos variables
- Población: existen dos países que aglutinan prácticamente el 35% de la población mundial, por lo que no está distribuida uniformemente y hace que el histograma no aporte mucho valor.
- Densidad Poblacional: situación similar a la anterior de forma que una gran mayoría de países se sitúan en valores bajos, existiendo unos pocos que hacen aumentar la media.

En ciertas variables se puede apreciar la existencia de outliers, pero en este caso no se van a eliminar debido a la falta de registros y a que se considera que pueden aportar valor adicional al modelo. También es necesario comprobar la correlación entre las variables de delgadez, obteniendo una correlación de 0.97 por lo que se decide eliminar la delgadez de 1 a 19 años

y también la variable percentage expenditure debido a que prácticamente la totalidad de sus valores son 0.

Finalmente, una vez se ha definido el conjunto de datos final y se han analizado sus variables, hay que comentar las técnicas que se van a utilizar para intentar resolver el problema inicial.

En primer lugar, se va a generar una regresión múltiple que es un modelo lineal que permite determinar el valor de una variable dependiente, que en este caso será la esperanza de vida, a partir de un conjunto de variables independientes denominadas predictores. La regresión lineal múltiple sigue la siguiente ecuación:

$$Y_i = \beta_0 + \beta_1 * X_{1i} + \beta_2 * X_{2i} + \cdots + \beta_n * X_{ni} + e_1 * Y_i$$

De forma que:

- β_0 es la ordenada en el origen, es decir, el valor de la variable dependiente si todas las variables independientes son nulas.
- β_i es el efecto que tiene el incremento en una unidad de una variable predictora manteniendo el resto constantes.
- e_i es el residuo o error, es decir, la diferencia entre el valor observado y el estimado por el modelo.

Una vez se genera un modelo inicial hay que observar que variables se consideran significativas y escoger los mejores predictores para generar el modelo. Para ello, se hará uso del método stepwise que emplea criterios matemáticos para decidir qué predictores contribuyen significativamente al modelo y en qué orden se introducen.

El método anterior ayudará en la selección de los mejores predictores, pero se necesita una validación adicional para comprobar que se cumplen todos los supuestos de una regresión múltiple para lo cual se analizará la relación lineal entre los predictores numéricos, que se puede realizar con una matriz de correlaciones. Otro aspecto a estudiar es la distribución de los residuos, para lo que se hará uso de ciertos gráficos que demuestren su distribución, y también es conveniente analizar el factor de inflación de varianza (VIF) que cuantifica la intensidad de la multicolinealidad en un análisis de regresión, y proporciona un índice que mide hasta qué punto la varianza de un coeficiente de regresión se incrementa a cause de la colinealidad.

3. Resultados Experimentales

Para comenzar este apartado, se realiza una regresión múltiple con el objetivo de comprobar como funciona el modelo si se genera con todas las variables excepto país, que evidentemente no tiene sentido incluirlo. EL resultado obtenido es el siguiente:

```
Call:
lm(formula = esp ~ des + mort_a + mort_i + alc + sar + bmi +
    pol + tot_exp + diph + hiv_aids + ed_med + th5 + pob + d_pob,
    data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9575 -1.8938 -0.0152  1.9056  9.2199

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.028e+01  2.071e+00  29.100  < 2e-16 ***
des1        -1.346e-01  9.123e-01  -0.148  0.88288
mort_a      -2.944e-02  3.863e-03  -7.619 2.21e-12 ***
mort_i       1.411e-03  5.699e-03   0.248  0.80473
alc         -1.415e-02  8.443e-02  -0.168  0.86714
sar         -3.213e-06  7.267e-05  -0.044  0.96479
bmi          1.794e-02  1.595e-02   1.125  0.26248
pol          6.940e-03  1.461e-02   0.475  0.63540
tot_exp      1.709e-01  1.004e-01   1.703  0.09061 .
diph         4.348e-02  1.642e-02   2.647  0.00894 **
hiv_aids    -7.202e-01  2.628e-01  -2.741  0.00683 **
ed_med       3.751e-01  5.548e-02   6.762 2.50e-10 ***
th5         -2.316e-01  8.592e-02  -2.695  0.00779 **
pob          3.927e-10  3.265e-09   0.120  0.90441
d_pob        5.037e-04  3.993e-04   1.261  0.20908
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.264 on 158 degrees of freedom
Multiple R-squared:  0.8487,    Adjusted R-squared:  0.8353
F-statistic: 63.29 on 14 and 158 DF,  p-value: < 2.2e-16
```

El modelo se genera a partir de la función lm y es interesante conocer si es explicativo o no. Si se acepta la hipótesis nula, el modelo no es explicativo, es decir, ninguna de las variables explicativas influye en la variable dependiente, mientras que, si se rechaza la hipótesis nula, el modelo es explicativo porque al menos una de las variables introducidas tiene relevancia en la variable dependiente.

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_n = 0$$
$$H_1: \beta_j \neq 0 \; para \; algún \; j = 1, \ldots, p$$

El p-value del modelo es significativo ya que es menor que 0.05, que se corresponde con el nivel de significancia que por defecto es del 5%, y más concretamente, en este caso es menor que 2.2e-16, por lo tanto, se rechaza la hipótesis nula y se pude concluir que el modelo no es puro azar ya que al menos alguno de los coeficientes parciales de regresión es distinto de 0 por lo que es explicativo, pero puede que otros predictores no sean significativos y no aporten valor adicional al modelo. También se observa que el valor de R^2 es alto, ya que es capaz de explicar el 83.53% de la variabilidad.

Tras este primer análisis, es interesante elegir qué predictores son los mejores para el modelo con el objetivo de no incluir variables que no aporten información. Para ello, se va a utilizar la metodología stepwise explicada en el apartado anterior:

```
Call:
lm(formula = esp ~ mort_a + tot_exp + diph + hiv_aids + ed_med +
    th5, data = df)

Coefficients:
(Intercept)       mort_a      tot_exp         diph     hiv_aids       ed_med
   60.99426     -0.03008      0.17583      0.04752     -0.76125      0.38993
        th5
   -0.22780
```

Se observa como las variables quedan claramente reducidas. El primer modelo consistía en 14 variables que se han visto disminuidas a 6, que son las que en el resumen del modelo aparecían con valores de p-value significativos, es decir, valores menos a 0.05 junto a total expenditure cuyo p-value era de 0.09.

Una vez se han definido los predictores, es interesante mostrar el intervalo de confianza para cada uno de los coeficientes parciales de regresión:

```
                   2.5 %      97.5 %
(Intercept) 57.459508813 64.52901606
mort_a      -0.037449002 -0.02270305
tot_exp     -0.009588945  0.36125448
diph         0.021512102  0.07353005
hiv_aids    -1.262135619 -0.26036809
ed_med       0.311315510  0.46854325
th5         -0.360601731 -0.09498971
```

Cada pendiente de los coeficientes parciales de regresión de los predictores se define de forma que, si el resto de las variables se mantienen constantes, por cada unidad que aumente el predictor, la variable dependiente varía en promedio tantas unidades como indica la pendiente.

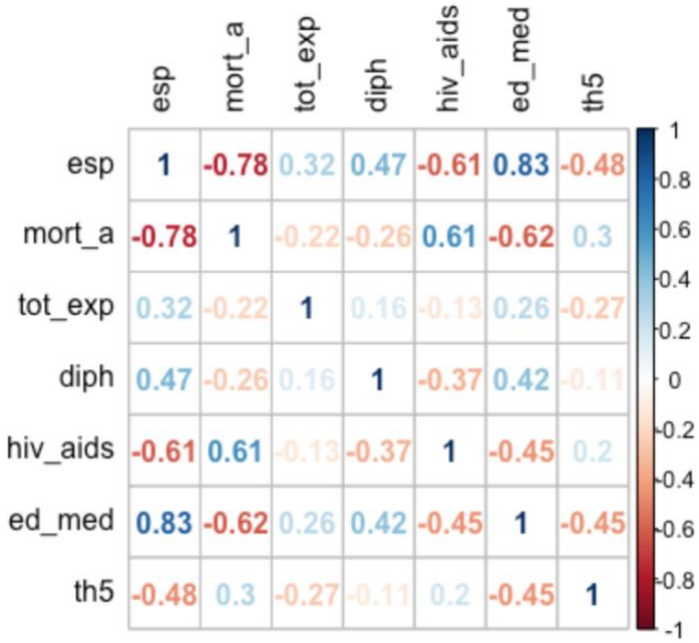
Para continuar con el análisis, es necesario generar el nuevo modelo con las variables indicadas:

```
Call:
lm(formula = esp ~ mort_a + tot_exp + diph + hiv_aids + ed_med +
    th5, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.2484  -1.9592   0.0721   1.9432   9.4313

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 60.994262   1.790331  34.069  < 2e-16 ***
mort_a      -0.030076   0.003734  -8.054 1.48e-13 ***
tot_exp      0.175833   0.093915   1.872 0.062930 .
diph         0.047521   0.013173   3.607 0.000409 ***
hiv_aids    -0.761252   0.253695  -3.001 0.003110 **
ed_med       0.389929   0.039817   9.793  < 2e-16 ***
th5         -0.227796   0.067265  -3.387 0.000884 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.217 on 166 degrees of freedom
Multiple R-squared:  0.8456,    Adjusted R-squared:   0.84
F-statistic: 151.5 on 6 and 166 DF,  p-value: < 2.2e-16
```

De nuevo, se puede comprobar que el modelo es explicativo siendo todas las variables significativas y explicando el 84% de la variabilidad, ligeramente superior al modelo anterior, pero con menos variables haciéndolo más eficiente.

Hasta este punto se ha llevado a cabo un preprocesamiento de datos, análisis de variables, elección de predictores y creación del modelo de regresión lineal múltiple, por lo que una vez efectuado todo ello, se debe examinar si los supuestos detrás de un modelo de regresión se cumplen. Estos son: no multicolinealidad, homocedasticidad, distribución normal de los errores, independencia y linealidad. El cumplimiento de estos supuestos implica que el modelo pueda aplicarse con precisión o no.

La independencia y linealidad se puede analizar comprobando la correlación que existe entre las variables dependientes entre sí y con la variable independiente. El paquete corrplot permite generar un gráfico que ilustra las correlaciones entre las variables:



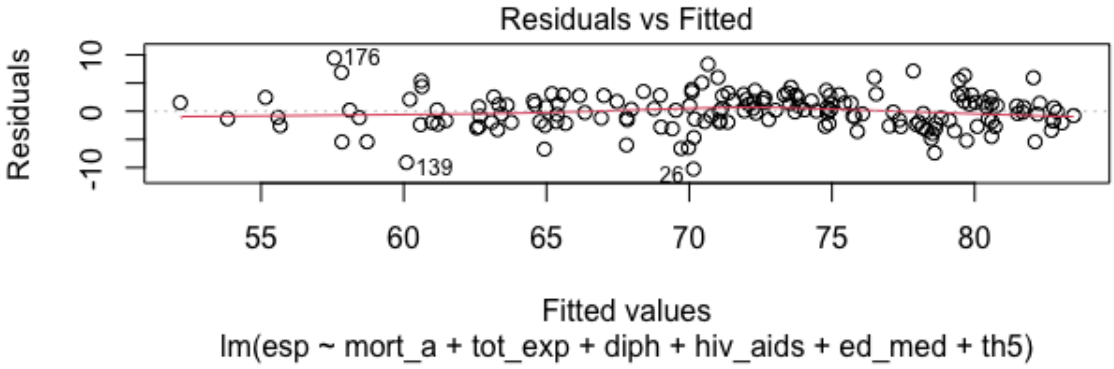
El objetivo es que no exista una correlación alta entre las variables independientes para que puedan considerarse como tal y que exista una cierta correlación, o una linealidad, entre la variable dependiente y las independientes. Teniendo en cuenta que a partir de un valor de ± 0.4 se considera que hay una correlación moderada entre las variables, y que a partir de un valor de ± 0.7 se considera que la correlación es alta, se puede determinar que todas las variables independientes están relacionadas en mayor o menor medida con la variable independiente incluso total expenditure, aunque en este caso con una correlación baja, que era esperable puesto que su p-value ya era menor que los del resto de variables significativas, por lo que se demuestra que es la que mantiene una correlación menor, y será la variable que menos información aporte al modelo. Por otro lado, ninguna pareja de variables independientes muestra una correlación alta.

Se puede continuar el análisis comprobando el Factor de Inflación de la Varianza (VIF) de las variables, que es un parámetro que permite conocer si existe colinealidad entre los predictores:

```
> vif(modelo_final)
  mort_a  tot_exp     diph hiv_aids   ed_med      th5
2.115778 1.119431 1.308912 1.729252 2.102131 1.310835
```

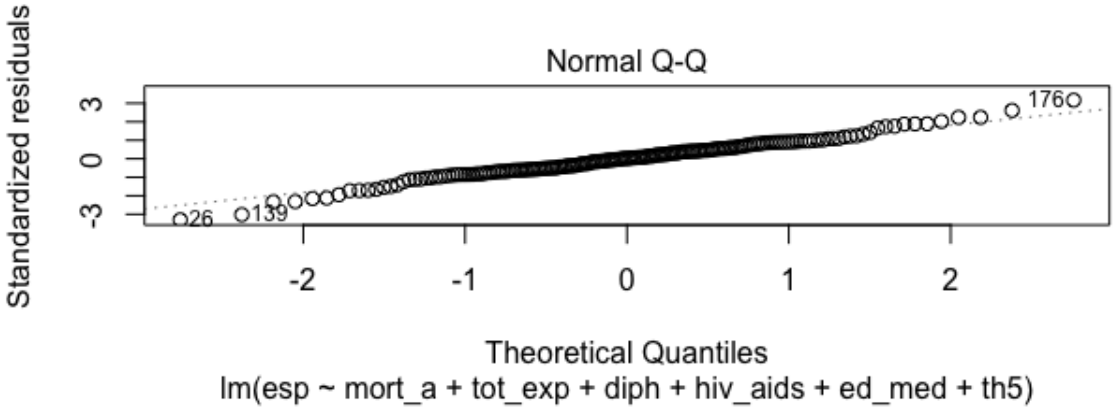
Un valor igual a 1 significaría la ausencia total de colinealidad, mientras que valores entre 1 y 5 muestran que el modelo puede verse afectado por ello y un valor entre 5 y 10 sería una señal de preocupación. Los resultados muestran que existe algo de colinealidad, pero no se considera significativa y no es necesario modificar el modelo.

Para finalizar con el análisis de los supuestos mencionados anteriormente, se debe analizar la homocedasticidad, que es la variabilidad constante de los residuos, y la independencia de los mismos. Para ello, se debe presentar en un gráfico los valores ajustados en el eje x y los residuos estandarizados en el eje y.



En el gráfico anterior se observa como la dispersión de puntos está distribuida uniformemente alrededor del cero, siendo esta una señal de que la varianza de los errores es constante y por lo tanto el supuesto de homocedasticidad se cumple.

También se puede comprobar si una variable tiene una distribución normal mediante el gráfico Q-Q, que incluye los valores acumulados de las variables contra la probabilidad acumulada de la distribución normal.



Si los errores tienen una distribución normal, el gráfico ha de mostrar una línea diagonal recta, como es el caso.

Por último, en el inicio del documento se menciona que la esperanza de vida se calcula en base a las tasas de mortalidad por lo que es lógico que este parámetro tenga una gran influencia en la variable dependientes. Por lo tanto, es interesante generar la regresión múltiple excluyendo está variable y comprobar la explicabilidad del modelo.

```
Call:
lm(formula = esp ~ tot_exp + diph + hiv_aids + ed_med + th5,
    data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-15.8547  -2.4536   0.2149   2.3959  10.9435

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.62502    1.80933  29.638  < 2e-16 ***
tot_exp      0.23470    0.11009   2.132   0.0345 *
diph         0.03520    0.01538   2.288   0.0234 *
hiv_aids    -1.75125    0.26093  -6.712 2.87e-10 ***
ed_med       0.52914    0.04217  12.546  < 2e-16 ***
th5         -0.23898    0.07907  -3.022   0.0029 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.782 on 167 degrees of freedom
Multiple R-squared:  0.7853,    Adjusted R-squared:  0.7788
F-statistic: 122.1 on 5 and 167 DF,  p-value: < 2.2e-16
```

Se puede comprobar que todos los coeficientes tienen un p-value inferior a 0.05 por lo que son significativos para el modelo y se puede explicar un 77.88% de la variabilidad, que aún siendo inferior al 84% del modelo anterior, se puede considerar que no es un valor bajo.

4. Conclusiones

A lo largo del documento se ha llevado a cabo un análisis de la esperanza de vida y de la capacidad de predecir su valor mediante otros factores a través de un modelo de regresión múltiple. Se han comprobado qué variables resultaban significativas para el modelo mediante el valor-p de las pruebas de hipótesis de que los coeficientes son 0, de forma que si este valor era menor a 0.05, que se corresponde con el nivel de significancia que por defecto es del 5%, se podía rechazar la hipótesis nula, y por lo tanto concluir que existe significancia en la regresión. También se ha utilizado la técnica stepwise, explicada durante el documento, para concluir qué variables había que escoger para generar el modelo.

Posteriormente se han revisado que todos los supuestos inherentes al modelo de regresión lineal múltiple se han cumplido, y tras ello, se puede concluir que el modelo previsto puede ser generalizable. De esta forma, la conclusión principal es que es posible predecir la esperanza de vida en función de una serie de factores relacionados con la mortalidad y la salud del país, que en este caso han sido variables como la delgadez hasta una edad de 5 años o enfermedades como la difteria. Este modelo permite explicar el 84% de la variabilidad, que se considera un porcentaje alto. El análisis realizado conduce a una dependencia de la variable esperanza de vida respecto a las demás que sigue la siguiente ecuación:

$$esp = 60.99 - 0.03morta + 0.18totexp + 0.05diph - 0.76hivaids + 0.39edmed - 0.23th5$$

También se comentó que la esperanza de vida se calcula en base a las tasas de mortalidad por lo que una dependencia con la mortalidad adulta era esperable, aunque la mortalidad infantil no ha afectado tanto al modelo, seguramente porque los valores recogidos son del año 2015, y gracias a los avances en medicina y tecnología se ha conseguido disminuir a nivel mundial la mortalidad infantil. Sin embargo, el parámetro que ha mostrado una correlación mayor con la esperanza de vida es la edad media de los habitantes del país, que también podría ser esperable. Finalmente, destacar que incluso eliminando la mortalidad como variable del modelo, se consigue explicar un 77.88% de la variabilidad, en gran parte debido a la edad media, de forma que los parámetros relacionados con la salud del país como pueden ser la delgadez o las enfermedades ayudan a predecir la esperanza de vida, ya que esta tiene una dependencia lineal de los mismos pero el porcentaje de variabilidad que se lograría explicar solo con aspectos relacionados con la salud de los habitantes no sería muy alto.