DEPARTAMENTO
DE COMPUTACION
Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico 3

Cuadrados Mínimos Lineales

1 de diciembre de 2021 — Metodos Numericos

Grupo 05

Integrante	LU	Correo electrónico
Casco, Rocío Diana	512/20	rociodcasco@gmail.com
Dallegri, Pablo	445/20	dallegri.p@gmail.com
Totaro, Facundo Ariel	43/20	facutotaro@gmail.com
Vitali, Lucas Marcelo	278/20	lucasvitali001@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2610 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (++54 +11) 4576-3300
https://exactas.uba.ar

Índice

Keywords: CML, EDA, life expectancy

1. Introducción

En este trabajo se busca investigar qué características de los países afectan positiva o negativamente a la expectativa de vida. Para ello se tiene un conjunto de datos que será analizado y luego, aplicando el método de cuadrados mínimos lineales, utilizado para determinar qué características se relacionan más con la expectativa de vida.

El proceso consiste en dos etapas. La primera etapa es un análisis exploratorio de datos, donde se pueden determinar correlaciones de los datos, ver si hay outliers y si los datos, en general, tienen sentido. La segunda parte es el análisis de regresión propiamente dicho, donde se generan modelos que nos ayudan a entender la contribución de las características al resultado final.

2. Métodos utilizados

2.1. Cuadrados Mínimos Lineales (CML)

Cuadrados Mínimos Lineales es un método de análisis numérico. Dado un conjunto de datos ordenados y una familia de funciones, este método encuentra una combinación lineal de dicha familia que mejor aproxime los datos.

Esto se realiza intentantando minimizar la suma de los cuadrados de las diferencias entre los datos obtenidos y los datos predichos por la función.

Se tiene una matriz A de la siguiente forma:

$$\begin{pmatrix} f_1(p_1^1) & f_2(p_2^1) & f_3(p_3^1) & ... & f_m(p_m^1) & 1 \\ f_1(p_1^2) & f_2(p_2^2) & f_3(p_3^2) & ... & f_m(p_m^2) & 1 \\ ... & ... & ... & ... & ... & ... \\ f_1(p_1^n) & f_2(p_2^n) & f_3(p_3^n) & ... & f_m(p_m^n) & 1 \end{pmatrix}$$

Donde, en nuestro caso, p_i^j es el feature i del país j. f_i es una función que se le aplica a los features para intentar buscar el mejor resultado. En general lo tomaremos como que se tienen de forma lineal. La columna de 1's es para agregar una ordenada al origen a la hora de aplicar el método. Luego se tiene el vector y el cual en su posición i tiene la expectativa de vida del país i, queriendo buscar la solución x^* de cuadrados mínimos del sistema $Ax = y$ para encontrar al x que disminuye el error con este método. x^* seria de la forma:

$$\begin{pmatrix} c_1 \\ c_2 \\ c_3 \\ ... \\ c_m \\ c_{m+1} \end{pmatrix}$$

Luego, dados los features de un país (que sean los mismos que se estaban usando para predecir un modelo), la expectativa de vida que predeciría el modelo seria de la forma:

$$(\sum_{i=1}^{m} c_i * f_i(p_i)) + c_{m+1}$$

2.2. Métricas Utilizadas

Las métricas que se utilizaron para medir la calidad del modelo durante la experimentación son:

- $\mathbf{R^2}$: es la proporción de variación de la variable dependiente que es predecible a traves de las variables independientes. En otras palabras, nos dice cuanta variabilidad explica el modelo. La misma se calcula como:

$$R^2 = \frac{TSS - RSS}{TSS}$$

donde TSS (total sum of squares) es una métrica de la variabilidad intrínseca de los datos y se calcula como:

$$\sum (y_i - \bar{y})^2$$

y RSS (residual sum of squares) es:

$$\sum (y_i - \hat{y})^2$$

donde y_i son los datos a predecir, $\bar{y}$ es la media de los y_i e $\hat{y}$ es la predicción del dato según el modelo.

- $\mathbf{R^2}$ ajustado: mide lo mismo que $\mathbf{R^2}$ pero agrega penalización por cantidad de variables (p) y tiene en cuenta la cantidad de datos (N):

$$R^2_{ajustado} = 1 - \frac{(1 - R^2)(N - 1)}{N - p - 1}$$

- VIF (Variance Inflation Factor): es una medida de la cantidad de multicolinearidad de una variable, en particular, en un conjunto de variables en un modelo de regresión. Para calcular esto se hace un modelo con la variable a la que le queremos calcular el VIF (X_j) como target y se usan el resto de las variables para predecirlo, se calcula el R^2 y se aplica la siguiente formula:

$$VIF(X_j) = \frac{1}{1 - R^2}$$

- Distancia de cook: se utiliza para estimar la influencia de una observación en los resultados del modelo. Se calcula de la siguiente forma:

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{ps^2}$$

donde $\hat{y}_j$ es la prediccion del dato j con el modelo completo, $\hat{y}_{j(i)}$ es la predicción del dato j del modelo sin cosiderar el dato i, p es la cantidad de features y s^2 es el error cuadratico medio del modelo de regresion que se calcula como:

$$s^2 = \frac{\sum_{j=1}^{n} (y_i - \hat{y}_i)^2}{n - p}$$

3. EDA

En esta sección se realiza el análisis exploratorio de datos (EDA), para lograr obtener un entendimiento de la estructura de los mismos y encontrar algunas primeras relaciones entre ellos.

Para comenzar vamos a describir las definiciones de cada dato, remarcando lo que representan y con que tipo se representan:

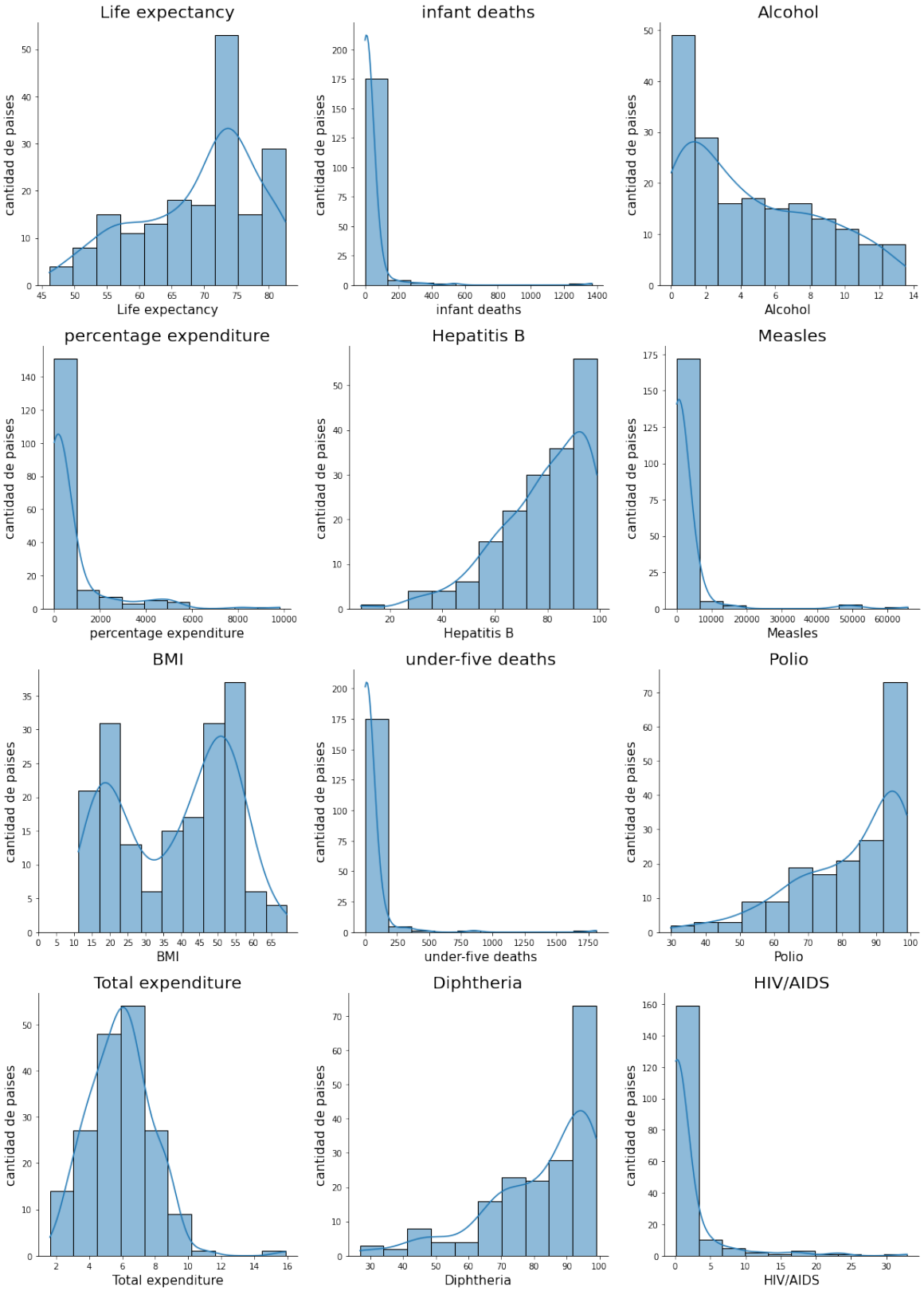
Feature	Fescripción	Tipo computacional	Tipo estadístico
Country	País	string	Identificador de datos
Life expectancy	Probabilidad de morir entre los 15 y 60 años cada 1000 habitantes	float	Porcentaje
Infants Deaths	Probabilidad de morir en el primer año de vida cada mil habitantes	float	Porcentaje
Alcohol	Consumo de alcohol mayores de 15 años en litros de alcohol puro	float	Numérico
Percentage expenditure	Gasto en salud relativo al GDP	float	Porcentaje
Hepatitis b	Inmunización entre los niños de un año	float	Porcentaje
Measles	Casos de sarampión reportados	float	Cantidad cada 1000
BMI	Promedio de índice de masa corporal	float	Numérico
Under-five deaths	Número de muertes de menores de 5 años	float	Cantidad cada 1000
Polio	Inmunización entre los niños de un año	float	Porcentaje
Total expenditure	Gasto en salud del gobierno sobre los gastos totales del gobierno	float	Numérico
Diphtheria	Inmunización entre los niños de un año	float	Porcentaje
HIV/AIDS	Muertes por HIV/AIDS en personas de 0 a 4 años	float	Numérico
GDP	(Producto Bruto Interno) valor monetario de la producción de bienes y servicios de demanda final	float	Numérico
Population	Población	float	Numérico
Thinness 1-19 years	Personas entre 1 y 19 años con IMC menor a 2 desviaciones estandar de la media	float	Porcentaje
Thinness 5-9 years	Personas entre 5 y 9 años con IMC menor a 2 desviaciones estandar de la media	float	Porcentaje
Income composition of resources	Como se utilizan los recursos productivos	float	Numérico
Schooling	Años de educación escolar en promedio	float	Numérico
Status	Calificación del país según la ONU (desarrollado o en vias de desarrollo)	string	Categórico
Poisoning *(*)*	Casos de envenenamiento no intencional	float	Cantidad cada 100 000

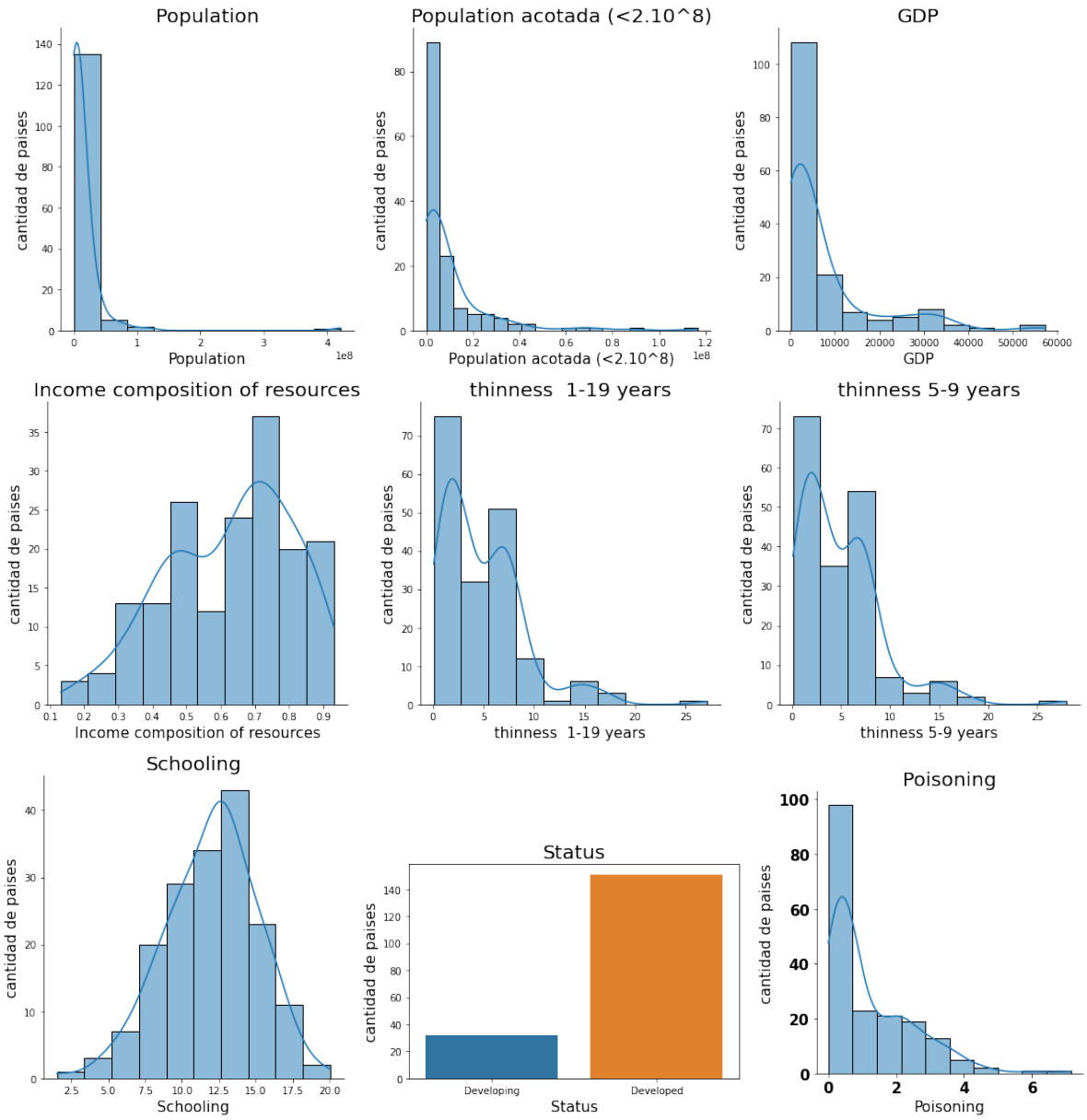
(*) El dataset se tomó de la misma fuente que el resto de los datos: LINK

El dataset consta de 183 paises con 20 features cada uno. Cada feature tiene la siguiente cantidad de datos faltantes:

- Alcohol: 1
- Hepatitis B: 9
- BMI: 2
- Total expenditure: 2
- GDP: 25
- Population: 40
- Thinness 1-19 years: 2
- Thinness 5-9 years: 2
- Income composition of resources: 10
- schooling: 10

Para obtener un mayor entendimiento de los datos, graficamos sus distribuciones:





Con esta primera parte del EDA se pueden realizar las siguientes observaciones:

- A lo sumo hay 5 datos faltantes por país.
- En el Índice de Masa Corporal (IMC/BMI) se observa una distribución bimodal, hay un pico en un índice que se interpreta como sano y un pico mayor en un índice que se interpreta como obesidad.

3.1. Análisis de correlaciones

Analizamos las correlaciones de los datos para ver que tan relacionadas están las features entre sí, luego se puede agrupar a aquellas cuya correlación sea alta, de esta manera es posible deshacerse de features que estén muy relacionadas con otras o ver qué features se correlacionan y hacer una sola feature que contenga a todas.

Mirando la distribución de datos creemos que existen los siguientes grupos de correlaciones:

- Hepatitis B - Polio - Dihtheria
- Thinness 1-19 years - Thinness 5-9 years - BMI
- Under-five deaths - Infant Deaths - HIV/AIDS
- Measeales - Percentage Expenditure
- Adult Mortality - Alcohol - Schooling
- GDP - Population - Income composition of resources - schooling
- Total Expenditure - Percentage Expenditure

Luego, analizamos las correlaciones entre las features. Consideramos que las features estaban relacionadas cuando el coeficiente de correlación era mayor a 0.7. A partir de esto obtuvimos los siguientes "conjuntos de correlación", donde cada elemento de un conjunto correlaciona de manera relevante con al menos otro de los elementos del conjunto.

1. Infant Deaths - Under-five deaths - Population
2. GDP - Percentage Expenditure
3. Hepatitis B - Polio - Diphtheria
4. BMI - Thinness 5-9 years - Thinness 1-19 years - Schooling - Income Composition of Resources

Los coeficientes de correlación entre estas features son los siguientes:

1. Conjunto 1:
   - Infant Deaths - Under-five deaths: 0.9969586952904269
   - Infant Deaths - Population: 0.9060955430328702
   - Under-five deaths - Population: 0.8910447036998426
2. Conjunto 2:
   - Percentage Expenditure - GDP : 0.9423748013157646
3. Conjunto 3:
   - Hepatitis B - Polio : 0.7908827871783201
   - Hepatitis B - Diphtheria : 0.7929991582444063
   - Polio - Diphtheria : 0.9456565342137309
4. Conjunto 4:
   - BMI - Thinness 1-19 years : -0.7146331627666422
   - BMI - Thinness 5-9 years : -0.7173653301255215
   - BMI - Schooling : 0.72126251838276
   - Thinness 1-19 years - Thinness 5-9 years: 0.9850978271410602
   - Income composition of resources - Schooling : 0.8640546067156053

Luego de armar estos conjuntos se observa que algunas features no estan incluidas en ninguno de ellos. Analizando las correlaciones más fuertes de estas features obtuvimos los siguientes resultados:

- Adult Mortality - HIV/AIDS: 0.696998872374393
- Alcohol - Schooling : 0.6234719694124654
- Total Expenditure - Alcohol: 0.4279160655448252
- Poisoning - Adult Mortality: 0.3859788264534555

Como Adult Mortality y HIV/AIDS tienen correlación muy cercana a 0.7, los vamos a tomar en cuenta como el quinto conjunto de correlación.

Finalmente, para agregarle interpretacion a este análisis, podemos decir que:

- El conjunto 1 representa a la cantidad de muertes infantiles de un país (las cuales parecen estar correlacionadas con la población)
- El conjunto 2 representa la riqueza de un país
- El conjunto 3 representa los niveles de vacunación de un país
- El conjunto 4 representa los niveles de nutrición de un país (tanto si existe desnutrición como obesidad) (lo cual parece correlacionar con schooling e income composition of resources)

3.2. Análisis de Outliers

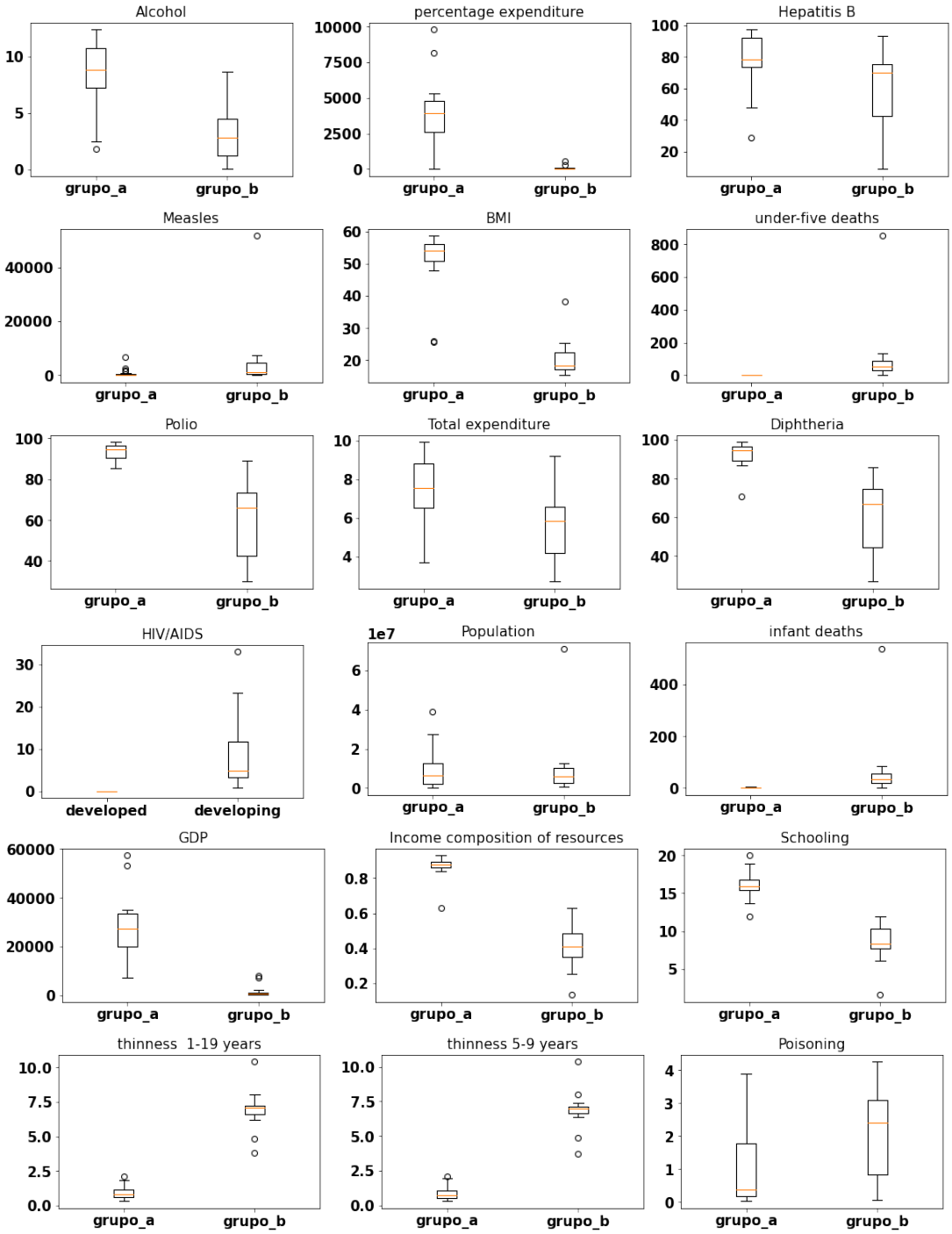
Analizamos los outliers para tenerlos en cuenta en el modelo y, de esta manera, evitar el efecto de apalancamiento. Realizando box plots de cada uno de los features se puede observar que la gran mayoría tiene outliers. Los únicos features que no tienen outliers son Alcohol, BMI e Income composition of Resources. Como los boxplots no agregan mayor información no lo agregamos en este informe pero se los puede ver en el jupyter notebook de EDA.

Con este mismo analisis pudimos detectar outliers que eran datos erroneos, es decir no estaban dentro de los rangos de las features:

- India tiene una mortalidad infantil (Infant Death) de 1366, pero esta metrica se mide cada 1000 habitantes.
- India tiene 1812.5 muertes en niños menores de 5 años (under-five deaths), pero esta métrica, además de que tiene que ser un número entero, se mide cada 1000 habitantes.
- En Measles se encuentran varios datos con valor mayor a 1000, cuando es un feature que se mide cada 1000 habitantes

3.3. Analisis por expectativa de vida

Para analizar las caracteristicas de los paises segun su expectativa de vida, tomamos en cuenta a los paises que estan por encima del 90 percentil (grupo A) y los que estan por debajo del 10 percentil (grupo B). El grupo A tiene casi un 80 % de países desarrollados mientras que en el grupo B están todos en vias de desarrollo. Luego graficamos los boxplots de las features, donde el eje y son los valores que toma cada feature:



Lo que en general se puede notar en los graficos anteriores es que los paises de grupo A tienen valores mas "favorables" de las features que el grupo B. Por ejemplo, tienen más alta tasa de vacunación (hepatitis B, Polio, etc), tienen GDP mas alto, y menos enfermedades (measles más bajo).

Como se mencionó antes, la mayor diferencia entre estos grupos es su Status (Developed/Developing) y será analizado en la siguiente sección en mas detalle.

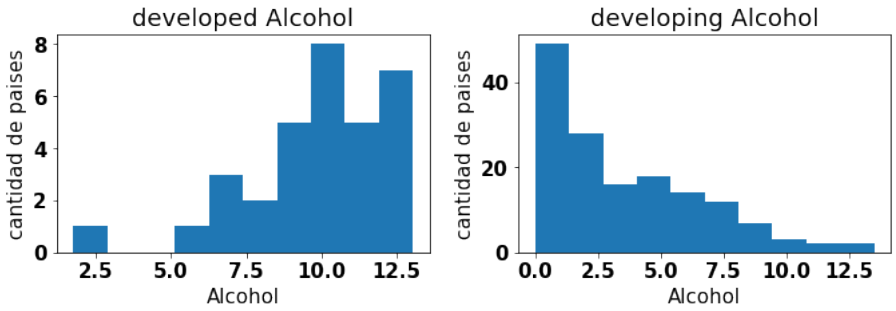
3.4. Análisis por Status

Realizamos un análisis de los países dados según su status para obtener una mejor intuición de como esta variable categórica puede ayudarnos a mejorar el modelo.

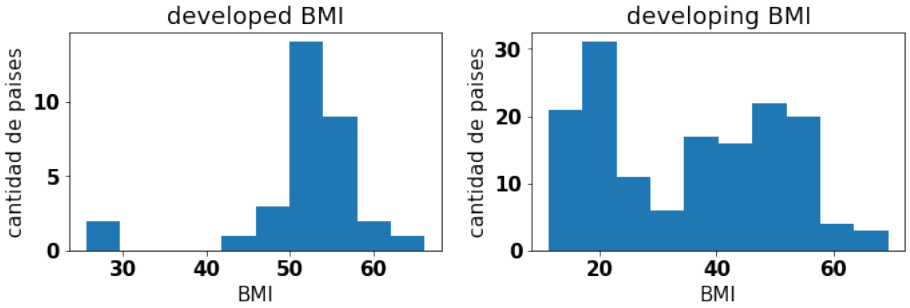
En primer lugar realizamos box plots para ver la distribución de las features según su status. Para esto dividimos los países en dos grupos "developed" y "developing" y luego comparamos los boxplot de sus features. Como resultado obtuvimos gráficos muy similares a los de la sección anterior (grupo a vs grupo b) por lo que no suma información agregalos a este informe.

En segundo lugar realizamos histogramas de cada feature diviendo nuevamente entre los grupos developed y developing. Comparando los histogramas del principio del EDA (cada feature teniendo en cuenta todos los países) pudimos observar que:

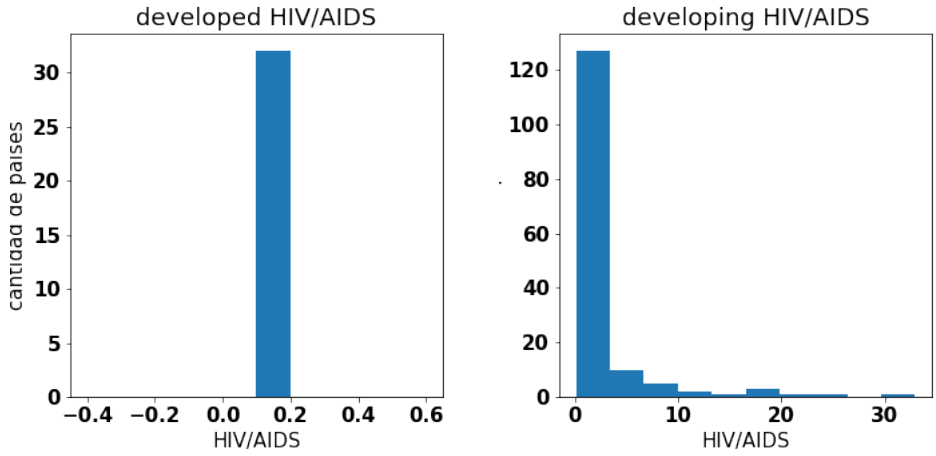
- La distribución del Alcohol de los países desarrollados es bastante diferente a la distribución que se obtiene con todos los países en su conjunto. A partir del gráfico se puede deducir que los países desarrollados tienden a un mayor consumo de alcohol que los que están en vías de desarrollo, habiendo una gran cantidad de estos últimos con valores cercanos a 0. Creemos que esto se debe a que los países desarrollados cuentan con los recursos sufientes para cubrir sus necesidades básicas, con exceso suficiente para permitirse consumir más.



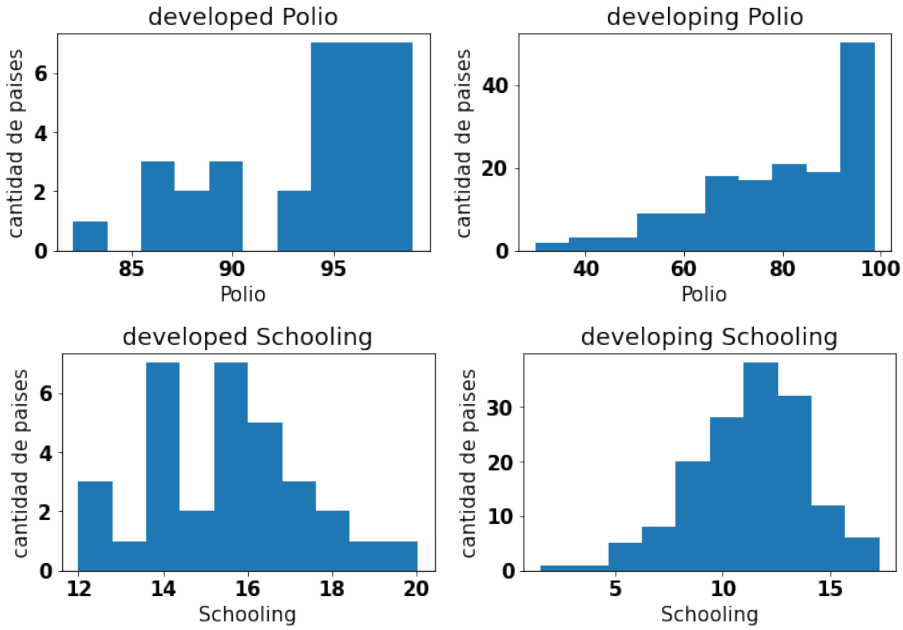
- La distribución del BMI en países en vias de desarrollo tiene la misma distribución que si tomamos todos los países, en el caso de países desarrollados tiene una tendencia a valores más altos (un poco de obesidad). Con los países desarrollados se toma una parte del "pico alto" de la bimodal. El BMI es un caso parecido al del alcohol, donde puede pasar que en los países desarrolados pueden permitirse consumir más gracias a su gran cantidad de recursos, como veremos más adelante.



- La distribucion de HIV-AIDS de los países desarrollados indica que los mismos tienen una muy baja mortalidad en personas entre 0 y 4 años por esta enfermedad. Más precisamente, al ver cada país en particular vimos que todos tenían valores extremedamente cercanos al 0.



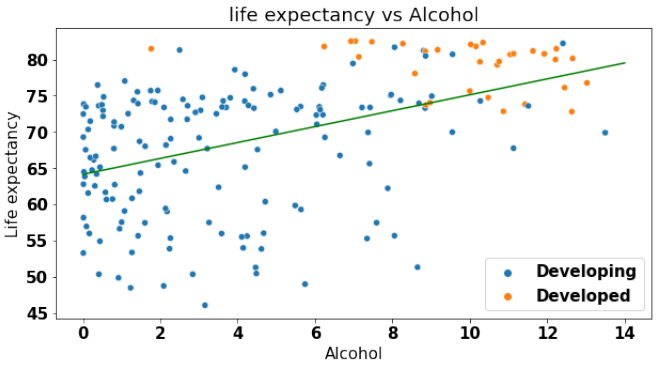
Luego de esto, analizamos los rangos que tomaban las variables en cada uno de los grupos. Acá se puede ver como en los paises desarrollados los rangos de variables son más "favorables" que en los países no desarrollados. Por ejemplo: en Polio, los paises desarrollados tienen un minimo de 80 % mientras que los que están en vías de desarrollo tienen un mínimo de 30 %; en Schooling, los paises desarrollados tienen un mínimo de 12 años, mientras que en los que están en vías de desarrollo tienen un mínimo de 2 años.



Considerando lo visto anteriormente, incluir el *status* del país en el modelo, podría ser beneficioso para la predicción de su expectativa de vida.

3.5. Análisis de Alcohol como feature confounding

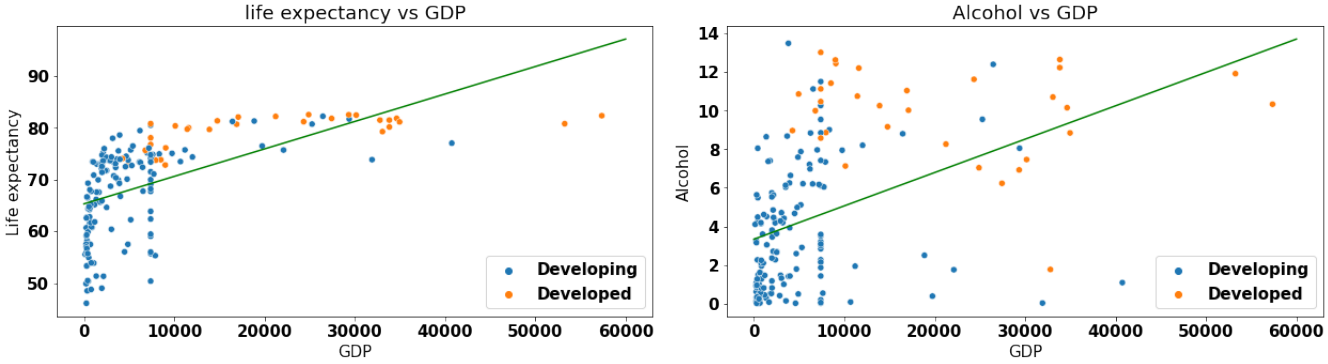
Analizando la relación entre las diferentes features y life expectancy encontramos que un incremento en el consumo de alcohol implicaría un incremento en la expectativa de vida.



Esta relación nos resultó sospechosa ya que generalmente el consumo de alcohol está relacionado con diferentes problemas y falta de salud. Partiendo de esta sospecha analizamos la posibilidad de que se tratase una feature confounding.

Estas son features que, más allá que podrían mejorar el R^2 ajustado, no aportan información valiosa al modelo ya que no afectan directamente la expectativa de vida y sus valores son efecto secundarios de valores de otras features.

Nuestra hipótesis es que en general los países con más riquezas pueden obtener una mayor expectativa de vida y a su vez tener recursos de sobra para utilizar en bienes no esenciales como el alcohol. Para corroborar dicha hipotesis analizamos la relacion entre GDP con life expectancy, y GDP con Alcohol, tomando a GDP como el indicador de la riqueza de un país.



Como se puede apreciar en los gráficos, nuestra hipótesis pareciera ser correcta, ya que un incremento en el GDP causa un incremento tanto en la expectativa de vida como en el consumo de alcohol, por lo tanto alcohol sería una variable confounding al momento de querer predecir la expectativa de vida.

4. Experimentación

4.1. Herramientas de experimentación

Implementamos en C++ un módulo que resuelve el problema de CML. Luego dicho módulo lo importamos desde Python para realizar la experimentación.

En Python usamos Jupyter Notebook para la ejecución del código. La mayor parte de la experimentación se basa en la observación de los resultados que se obtienen de la función análisis(), la cual recibe un data set y se le especifica qué parámetro se quiere predecir, en nuestro caso expectativa de vida, y qué regresores se toman en cuenta para hacerlo. Luego muestra:

- los coeficientes obtenidos para la regresión
- los coeficientes obtenidos para la regresión sobre los features normalizados (para ver qué tanto influye cada uno sobre el modelo)

- un gráfico de residuos estandarizado
- el VIF de cada predictor
- el R^2 y el R^2 ajustado del modelo
- los países con mayor distancia de Cook (y su correspondiente distancia)
- un grafico con las distancias de Cook

En base a estos resultados se fueron tomando las decisiones para encontrar el predictor que mejor se adapte a lo que estamos buscando.

4.2. Experimentación

4.2.1. Sanitización de datos

El primer problema con el que nos encontramos fue que al dataset le faltaban datos. Probamos:

- Completar con el promedio de los datos que se tienen. Esta opción no es muy buena ya que, por ejemplo, imaginando que tenemos un grupo reducido de países que tienen alto GDP, en relación a los demás. Al completar con la media, puede haber casos en donde los valores promedios estén muy lejos de ser el valor real del país que tiene el dato faltante, ya que el promedio es sensible a los outliers.
- Completar con la mediana de los datos que se tienen. Parece ser una mejor opción que el promedio, ya que no es tan sensible a los outliers.
- Completar con la mediana, pero teniendo en cuenta el *status* del país. La categoría de *status* de un país indica si este es desarrollado o está en vías de desarrollo. Al tener esto en cuenta, podemos lograr mejores estimaciones, ya que es más acertado pensar que un país va a tener valores similares a los países de su mismo *status*.

También se buscó inferir valores faltantes con otros datos que, en teoría, ya se tenían que estaban muy correlacionados con los faltantes. Es decir, por ejemplo, usando los datos del "percentage expenditure", calcular el GDP de los países que faltaban. El problema es que estos datos, no brindaban información, ya que los países con dato faltante en GDP, eran los mismos países con percentage expenditure completado con 0. Esto impedía la inferencia.

Entre las tres opciones que se mencionaron al principio, se buscó ver si alguna era mejor que los otras. Para esto se realizaron experimentos, variando la forma de completado y quitando algunos regresores que tenían correlación con otros. Lo que se observó fue que el R^2 ajustado era muy similar para todos. Decidimos quedarnos con la opción de la mediana teniendo en cuenta el status, ya que entendemos que es la opción que mejor podía modelar.

4.2.2. Armado de modelo (backward selection)

Reducción del máximo VIF

Se buscó ver que se podía hacer con los datos correlacionados, ya que había que tratarlos de alguna forma por lo mencionado en el EDA.

Se tomo la decisión de remover *adult mortality* ya que no es directamente accionable debido a qué engloba demasiados factores.

Por lo visto durante las clases, un VIF mayor a 5 en alguno de los predictores era algo por lo que nos teníamos que preocupar. Para tratar los regresores correlacionados, lo que se hizo fue un backward selection de estos según el VIF, es decir, se inició considerando todos los features y, de forma iterativa, se fue excluyendo el de mayor VIF, hasta quedarnos solamente con regresores de VIF menor a 5. Los VIFs de los valores exluidos en cada iteración fueron:

Iteración	R^2 adj.	Máximo VIF	feature problemático
1	0.8608	257.7700	infant deaths
2	0.8598	38.2407	thinness 5-9 years
3	0.8607	11.9683	under-five deaths
4	0.8598	10.6441	Diphtheria
5	0.8560	7.8218	GDP
6	0.8568	5.3826	Schooling

Llegando así al modelo siguiente:
R^2 adj: 0.8537

Feature	VIF	coef. con estandarización
HIV/AIDS	1.2529	-2.9912
Polio	3.9623	+2.8197
Income composition of resources	2.9241	+2.7556
BMI	3.2615	+1.8367
percentage expenditure	1.5944	+1.3136
Poisoning	1.1937	-0.7909
Hepatitis B	3.1091	-0.5510
Total expenditure	1.3658	+0.3487
thinness 1-19 years	2.9363	+0.2479
Measles	1.456	-0.0481
Population	1.6886	-0.0262
Alcohol	1.9172	+0.0106

Aplicación de conocimiento de dominio

Cabe considerar que *Population* no es directamente accionable (o eso esperamos), por lo que decidimos retirarla del modelo.

Luego, buscamos eliminar las posibles features confounding. Entre las features que quedan en el modelo, *Alcohol* es confounding [1], además de aportar poco al modelo, el consumo elevado de alcohol en países con mejor expectativa de vida se puede dar a que las poblaciones de dichos países tienen mayor poder económico, lo que las lleva a consumir más.

Eliminar datos redundantes o insignificantes

Luego de la exclusión de *Population* y *Alcohol* se buscó eliminar features que además de aportar poco al modelo (tenían un coeficiente muy bajo), estaban altamente correlacionadas con otras variables que sí aportaban:

Como sabíamos de antemano que *Polio* y *Hepatitis B* estaban correlacionadas, optamos por sacar la de menor coeficiente con datos estandarizados (*Polio*: 2.81, *Hepatitis B*: 0.05). Y con este nuevo modelo, el coeficiente de Measles se volvía insignificante (0.00059), por lo que se decidió removerla.

[1] Analizado en la sección 3.5

Modelo alcanzado
Hasta esta etapa, este es el modelo alcanzado. (También resulta ser el modelo final.)

R^2: 0.8621
R^2 ajustado: 0.8558

Feature	VIF	coef. con estandarización	coef. sin estandarización
Constante	-	-	42.6851
HIV/AIDS	+1.2312	-2.9605	-0.6674
Income composition of resources	+2.6524	+2.7797	+15.3923
Polio	+1.5895	+2.3443	+0.1469
BMI	+2.9737	+1.9088	+0.1209
percentage expenditure	+1.5088	+1.3782	+0.0009
Poisoning	+1.1507	-0.8015	-0.6326
Total expenditure	+1.2627	+0.3346	+0.1754
thinness 1-19 years	+2.1414	+0.2594	+0.0638

Reincorporación de datos previamente excluidos
Casi terminando con los experimentos, buscamos ver si era apropiado tener en cuenta de alguna otra manera a los regresores que se había descartado y estaban relacionados con alguno que se mantuvo. Guiándonos por lo visto en el EDA, optamos por tomar los grupos de regresores correlacionados que además compartían la misma unidad, y sumarlos. Este era el caso de los features que se refieren al porcentaje de la población vacunada ante ciertas enfermedades. Al experimentar con esto, vimos que el R^2 bajaba a 0.8127. Por lo que decidimos quedarnos solamente con la de *Polio*.

A los regresores que no compartían unidad, se decidió no reincorporarlos.

Obesidad vs desnutrición: BMI y thinness
El *BMI* es un indicador difícil de interpretar como factor lineal, ya que una persona con mucho *BMI* tiene sobrepeso o una persona con poco *BMI* tiene desnutrición, y sabemos que ambos casos son malos. Y entonces, el signo del coeficiente que acompañe este predictor, no sería lo suficientemente sugerente. Por esto, se decidió probar con una nueva variable que tome el módulo de la diferencia entre el *BMI* que se tiene y el ideal (22.5). Ahora, en parte, el *thinness 1-19 years* quedaría correlacionado, porque es sinónimo de mucha distancia al ideal del *BMI*. Removiendo thinness y usando este nuevo indicador, disminuye el R^2 ajustado en el orden de las centésimas.

Además, notamos que el coeficiente que acompaña a este nuevo predictor resulta positivo, contradiciendo lo que habíamos planteado. Esto puede deberse a que al tener un promedio del *BMI* total, quizás no tienen grandes módulos, debido a que se promedia y no se toma la cantidad de población obesa o desnutrida, y entonces no se vería reflejado de manera negativa en la esperanza de vida. Por lo que esta distancia no resulta un buen predictor y decidimos descartarla, volviendo a quedarnos con el modelo que incluye thinnes y BMI.

Nivel de desarrollo
Antes de comenzar con esta sección de experimentación teníamos un R^2 ajustado de **0.8558**. Cuando hablemos de que baja el R^2 ajustado, es en referencia a este. A continuación, buscamos formas de incorporar al modelo el nivel de desarrollo (*status*) de un país.

Un primer acercamiento se basó en tener dos modelos separados, uno para cada clase, pero esto solo produjo bajas en el R^2 ajustado, por lo que descartamos esta altenativa.

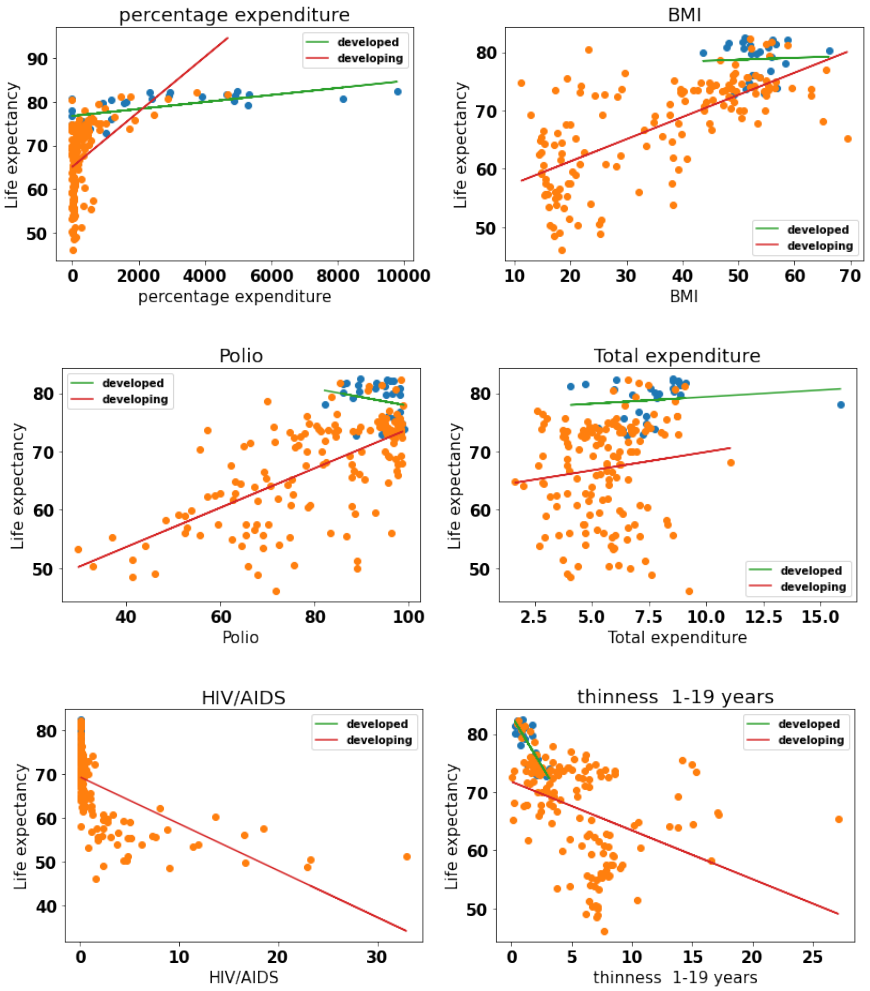
R^2adj desarrollados: 0.7440 | R^2adj en desarrollo: 0.8231

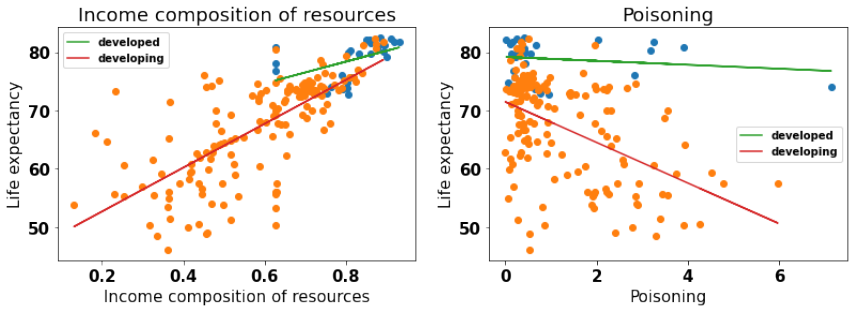
Otro intento de considerar el *status* de un país, fue incorporarlo como un predictor con valores binarios. A fines prácticos, esto es equivalente a incorporar un término constante solo para los países de una categoría.

La mejora en las predicciones del modelo no fue sufienciente para justificar el agregado de este nuevo predictor, disminuyendo el R^2-adj a 0.8551.

Fallados los intentos previos, intentamos incorporar el *status* al modelo por medio de una variable categórica (binaria), es decir, agregar un predictor que preserve los valores de otro ya existente, pero solo para los países de una categoría, y se anule para los demás. A fines prácticos, esto equivale a tener coeficientes distintos para un mismo predictor, según la categoría del país.

Analizando, para cada predictor del modelo, las pendientes de las dos regresiones lineales sobre la expectativa de vida al separar por categoría, concluimos que los mejores candidatos para variables categóricas (por la diferencia entre las pendientes) eran *percentage expenditure* y *Polio*





Sin embargo, al igual que en el experimento anterior, en ninguna de estas alternativas la mejora fue sufienciente para justificar el agregado de este nuevo predictor, dado que en ambas el R^2-adj disminuyó (0.8555 y 0.8438 respectivamente). Aunque en el caso del percentage expenditure disminuyó muy poco, y la diferencia puede ser ruido, sigue sin notarse una diferencia significativa que apoye la idea de dividir por *status*. Si tuvieramos más datos, de países desarrollados, quizás podríamos notar una mayor diferencia. En este caso, puede ser que no influya tanto esta diferencia porque tenemos menos de un 20 % de países desarrollados, por lo que es poco lo que estamos corrigiendo para un solo feature. Es decir, para el R^2 ajustado importa más la penalización de agregar el predictor, que el beneficio de la corrección que este genera.

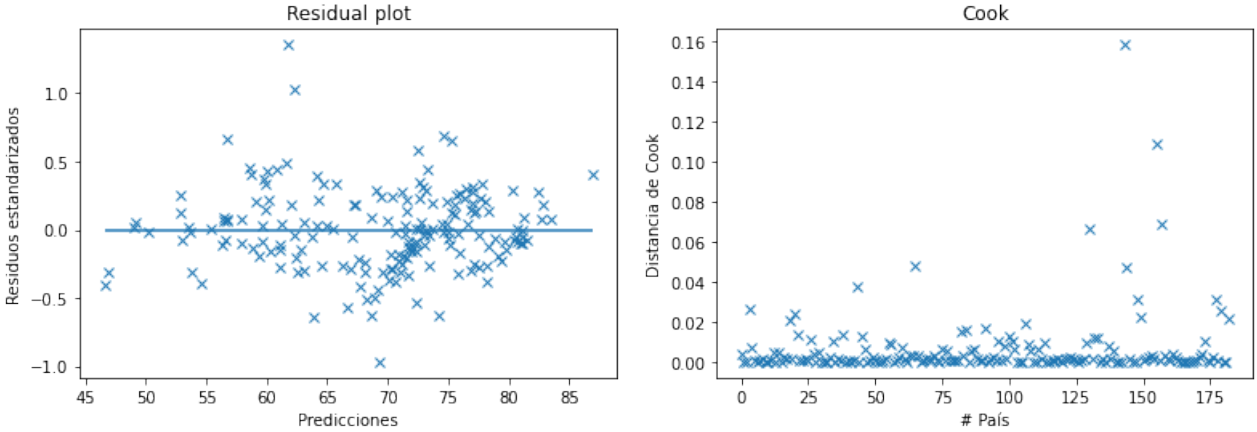
Para evitar ser víctimas de algo que ignoramos, tambíen probamos con HIV y BMI, pero funcionó. Obtuvimos un R^2-adj de 0.8440 y 0.8401 respectivamente.

Modelo final

Finalmente, romando en cuenta el análisis realizado, concluimos establecernos con el modelo descrito en la tabla anterior. (R^2 ajustado: 0.8558)

4.2.3. Outliers en el modelo

Analizando si había países o regresores que eran o generaban outliers, lo que se observó fue que, antes de remover adult mortality como regresor, India era un claro outlier. En el modelo final, ningún país parece ser un outlier, debido a que la distancia de Cook para todos los países es menor a 0,5 y la desviación estandar de los residuos normalizados es menor a 1,5 (es decir, no hay ninguno muy grosero) y no se observa ninguna correlación entre estos. Por lo que se optó por no remover ninguno del dataset.



5. Conclusión

Tras realizar un análisis exploratorio de los datos, corregir los necesarios, e iterar sobre distintas versiones, logramos obtener un modelo lineal capaz de predecir la expectativa de vida a partir

de pocas variables directamente accionables, pudiendo explicar aproximadamente un 85 % de la variabilidad de los datos.

La fórmula predictiva es la siguiente:

$$life\_expectancy = 9{,}33\times10^{-4}pe + 1{,}21\times10^{-1}BMI + 1{,}47\times10^{-1}polio + 1{,}75\times10^{-1}te - 0{,}667\times10^{-1}HIV/AIDS$$
$$+ 6.39\times10^{-2}t119 + 1{,}54\times10^{1}icr - 6{,}33\times10^{-1}poisoning + 42{,}68517297$$

donde:

- pe: *percentage expenditure*
- te: *total expenditure*
- t119: *thinness 1-19 years*
- icr: *Income composition of resources*

(notar que estos son los coeficientes para aplicar sobre los datos sin estandarizar)

A partir del mejor predictor que pudimos conseguir, podemos concluir que los factores (accionables) que más afectan la expectativa de vida de manera positiva son: *percentage expenditure*, *BMI*, *polio*, *total expenditure*, *thinness 1-19 years* e *Income composition of resources*, ya que todos estos tiene un coeficiente positivo en la formula. De manera contraria, los factores que afectan de manera negativa la expectativa de vida son: *HIV/AIDS* y poisoning, ya que tienen coeficientes negativos en la formula.

Sin embargo, queremos remarcar que los resultados del modelo pueden ser, a lo sumo, tan confiables como los datos en los que este está basado. La decisión sobre la validez y veracidad de los datos de partida, queda a criterio de los potenciales usuarios del modelo.

Además, resulta necesario advertir a quienes corresponda tomar decisiones que puedan afectar la expectativa de vida de la población, que las variables planteadas por el modelo NO deberían ser tomadas como vías de accion aisladas, sin tener en cuenta las posibles repercuciones y los sesgos concomitantes. Si alguna métrica reemplaza al objetivo original, dejará de ser una buena métrica.

Por último, se desea alentar a que futuras investigaciones en el tema, se enfoquen no solo en la expectativa de vida, sino también el periodo de la misma en el cual se goza de buena salud.