Analisis de Expectativa de vida por país

2024-03-27

Por medio de este análisis se busca estimar la expectativa de vida en distintos países considerando el conjunto de variables proporcionadas por la Organización mundial de salud.

Fuente: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Entre ellas encontramos las que se mostrarán acontinuación. ### Variables del set de datos

```
lifeExp <- read.csv("~/Downloads/lifeExp.csv")
lifeExp <- na.omit(lifeExp)
str(lifeExp)
```

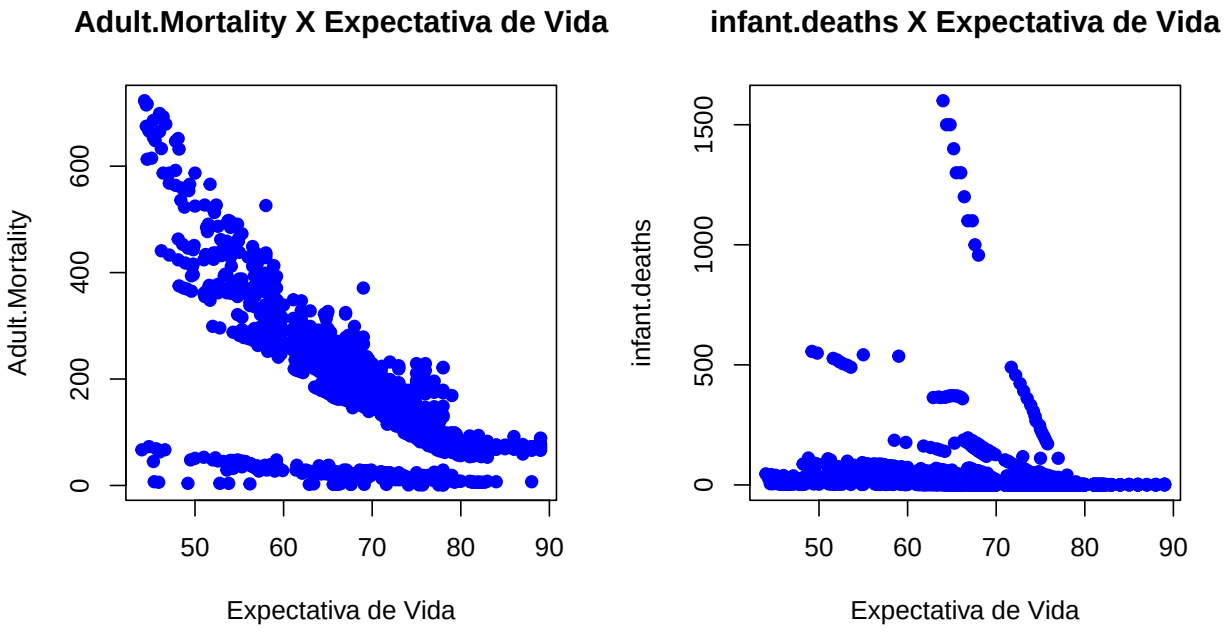
```
## 'data.frame':    1649 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1289] 33 45 46 47 48 49 58 59 60 61 ...
##   ..- attr(*, "names")= chr [1:1289] "33" "45" "46" "47" ...
```

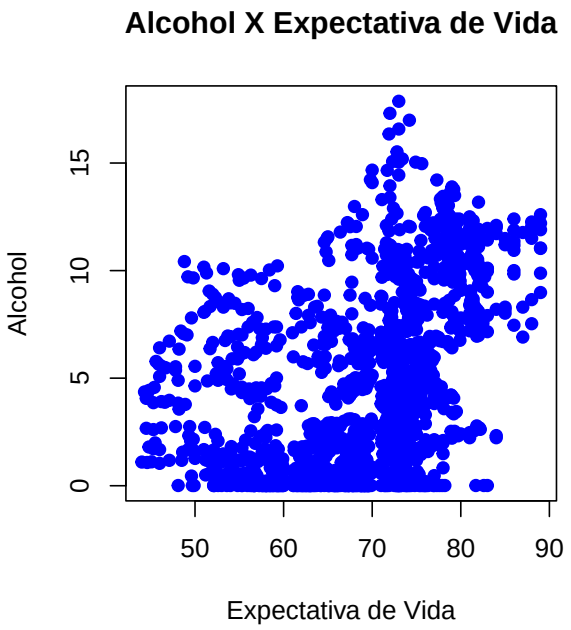
Gráficas de las variables selccionadas para el análisis

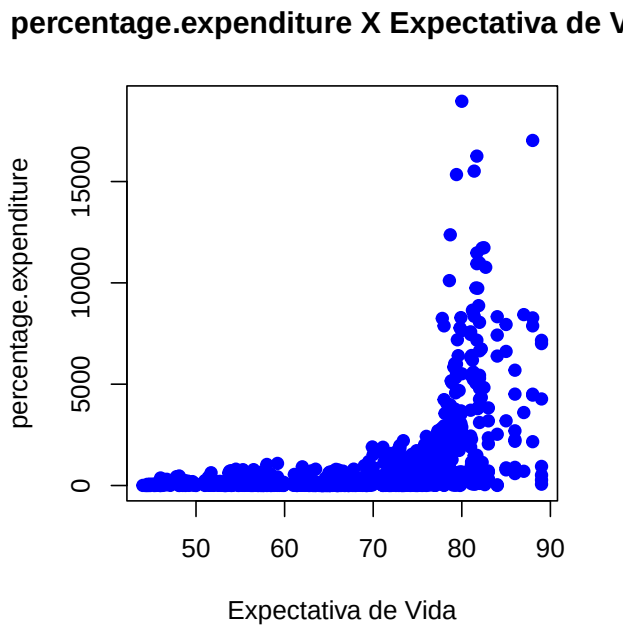
Tras realizar un análisis exploratorio de los datos, se encontró que ciertas variables no mostraban una correlación significativa con la expectativa de vida en todos los casos, ya que existían datos nulos debido a la ubicación de enfermedades y otros factores. Por lo tanto, estas variables fueron excluidas del análisis de regresión lineal. Las gráficas presentadas muestran las variables seleccionadas para el análisis, las cuales se consideraron más relevantes y mostraron una correlación más clara con la expectativa de vida. Las variables

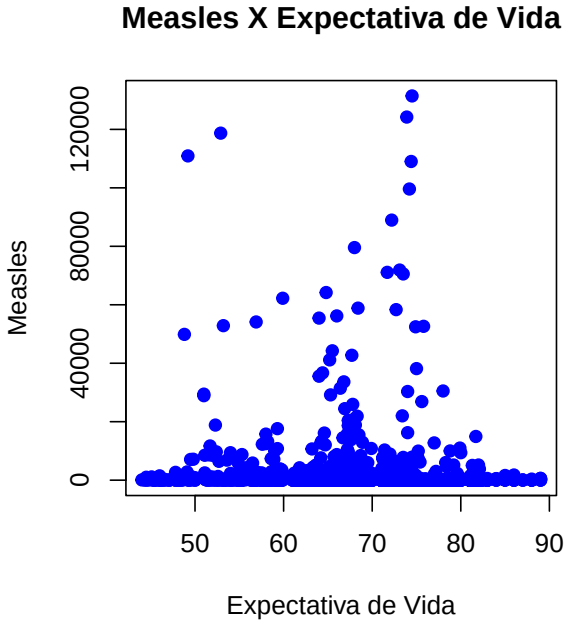
excluidas incluyen: “under.five.deaths”, “thinness..1.19.years”, “thinness.5.9.years”, “Polio”, “Diphtheria”, “Hepatitis.B”, “Total.expenditure” y “GDP”.

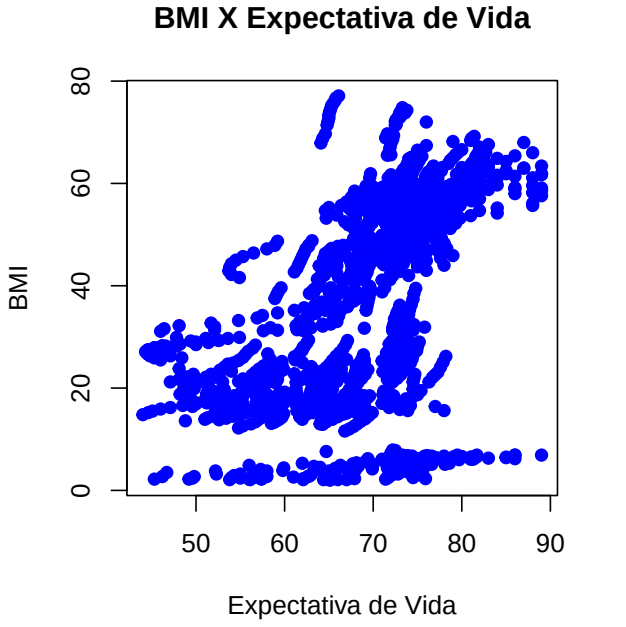
```
par(mfrow = c(1, 2))

columnas <- colnames(lifeExp)[!colnames(lifeExp) %in% c("Country", "Status", "Life.expectancy", "under.
num_columnas <- 2
num_filas <- ceiling(length(columnas) / num_columnas)

# Iterar sobre cada columna y hacer un gráfico
for (col in columnas) {
  plot(lifeExp$Life.expectancy, lifeExp[[col]],
       main = paste(col, "X Expectativa de Vida"),
       xlab = "Expectativa de Vida",
       ylab = col,
       pch = 19,
       col = "blue")
}
```

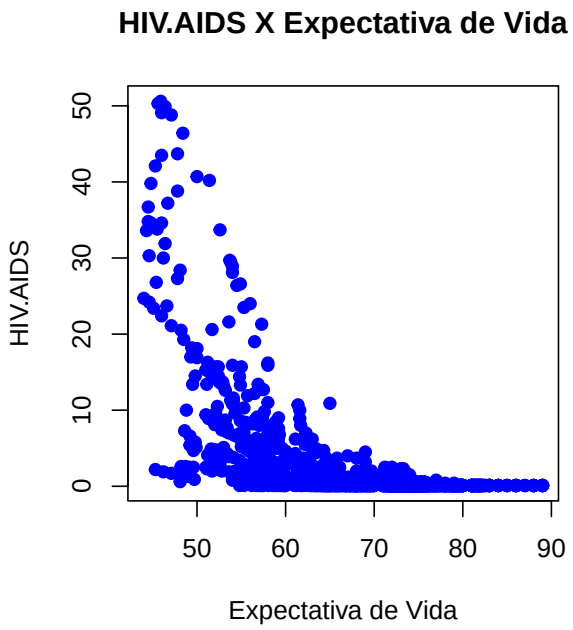


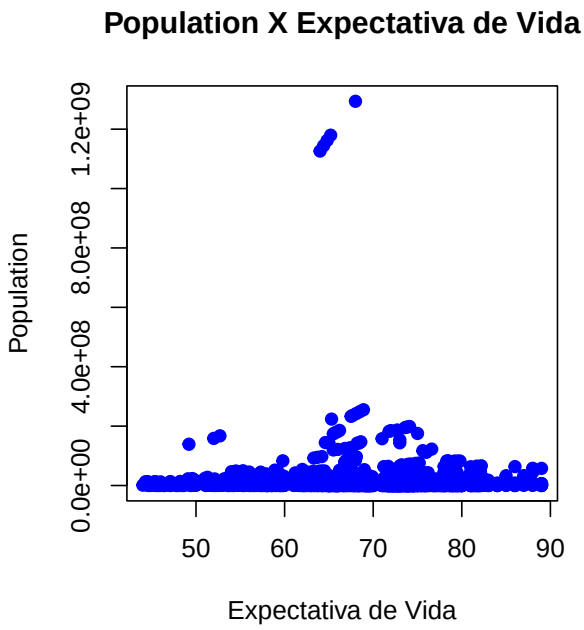

Alcohol X Expectativa de Vida
Alcohol
0
5
10
15
50
60
70
80
90
Expectativa de Vida

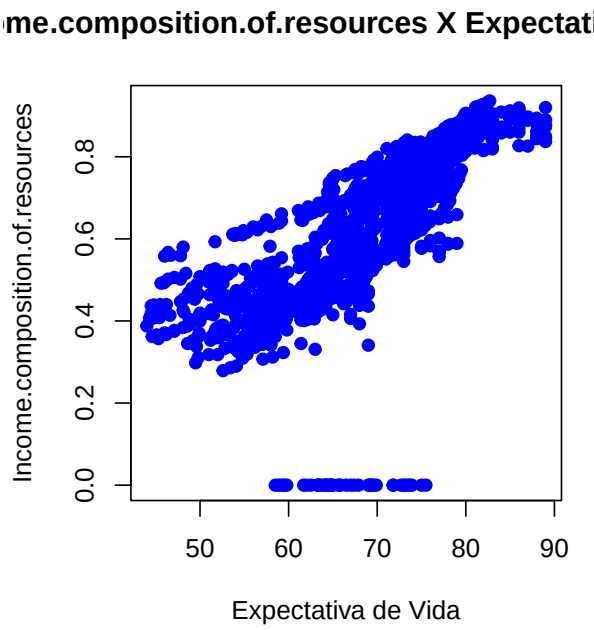

percentage.expenditure X Expectativa de V
percentage.expenditure
0
5000
10000
15000
50
60
70
80
90
Expectativa de Vida

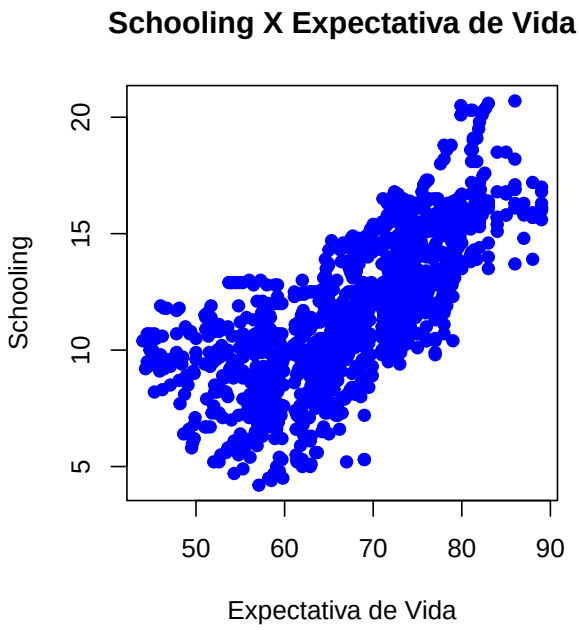

Measles X Expectativa de Vida
Measles
0
40000
80000
120000
50
60
70
80
90
Expectativa de Vida


BMI X Expectativa de Vida
BMI
0
20
40
60
80
50
60
70
80
90
Expectativa de Vida









1. **Mortalidad Adulta vs. Expectativa de Vida**: Relación negativa fuerte; a mayor mortalidad adulta, menor expectativa de vida.
2. **Muertes Infantiles vs. Expectativa de Vida**: Relación negativa; más muertes infantiles implican una menor expectativa de vida.
3. **Alcohol vs. Expectativa de Vida**: Relación débil o no aparente, sin un patrón claro.
4. **Gasto Porcentual en Salud vs. Expectativa de Vida**: Relación positiva; más gasto en salud podría asociarse con una mayor expectativa de vida, aunque con datos dispersos.
5. **Sarampión vs. Expectativa de Vida**: Existe una concentración de puntos en menor incidencia de sarampión con mayor expectativa de vida, pero sin una tendencia clara.

6. **IMC vs. Expectativa de Vida**: Dispersion amplia de puntos, indicando una relación compleja o no lineal entre el IMC y la expectativa de vida.
7. **VIH/SIDA vs. Expectativa de Vida**: Relación negativa fuerte; una mayor prevalencia de VIH/SIDA se correlaciona con una menor expectativa de vida.
8. **Población vs. Expectativa de Vida**: La relación no es clara debido a la amplia dispersión de los puntos de datos a través de los tamaños de población.
9. **Composición de Ingresos de los Recursos vs. Expectativa de Vida**: Relación positiva fuerte; una mayor composición de ingresos de los recursos tiende a correlacionarse con una mayor expectativa de vida.
10. **Escolaridad vs. Expectativa de Vida**: Indica una relación positiva fuerte; más años de escolaridad se correlacionan con una mayor expectativa de vida.

Modelo de regresión general (haciendo uso de todas las variables selccionadas)

```r
# Ajustamos un modelo de regresión lineal múltiple
modelo <- lm(lifeExp$Life.expectancy ~ ., data = lifeExp[, columnas])

summary(modelo)
```

```
##
## Call:
## lm(formula = lifeExp$Life.expectancy ~ ., data = lifeExp[, columnas])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.7520  -2.1316   0.0337   2.3975  12.2379
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5.342e+01  5.997e-01  89.080  < 2e-16 ***
## Adult.Mortality                -1.828e-02  9.679e-04 -18.880  < 2e-16 ***
## infant.deaths                  -4.407e-03  1.186e-03  -3.716 0.000209 ***
## Alcohol                        -8.633e-02  3.047e-02  -2.833 0.004662 **
## percentage.expenditure          4.553e-04  5.936e-05   7.670 2.94e-14 ***
## Measles                         1.714e-05  1.077e-05   1.592 0.111561
## BMI                             3.738e-02  5.769e-03   6.480 1.21e-10 ***
## HIV.AIDS                       -4.393e-01  1.830e-02 -24.003  < 2e-16 ***
## Population                      2.638e-09  1.777e-09   1.484 0.138006
## Income.composition.of.resources  1.093e+01  8.484e-01  12.878  < 2e-16 ***
## Schooling                       9.606e-01  6.015e-02  15.971  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.716 on 1638 degrees of freedom
## Multiple R-squared:  0.8227, Adjusted R-squared:  0.8216
## F-statistic: 759.9 on 10 and 1638 DF,  p-value: < 2.2e-16
```

Se puede evidencia que existen variables que no contribuyen significativamente a la explicación de la variable "Expectativa de vida", entre ellas esta Measles (La tasa de afectacción del Sarampión en los países observados por cada 1000 habitantes) y el tamaño de la población.

```r
if (!require("olsrr")) install.packages("olsrr")
```

```
## Loading required package: olsrr
```

```
##
## Attaching package: 'olsrr'
```

```
## The following object is masked from 'package:datasets':
##
##     rivers
```

```r
if (!require("leaps")) install.packages("leaps")
```

```
## Loading required package: leaps
```

```r
library(olsrr)
library(leaps)
mejor_subconjunto <-  ols_step_best_subset(modelo)
mejor_subconjunto
```

```
##                                                                 Best Subsets Regression
## ---------------------------------------------------------------------------------------------------------------
## Model Index    Predictors
## ---------------------------------------------------------------------------------------------------------------
##      1         Schooling
##      2         HIV.AIDS Schooling
##      3         Adult.Mortality HIV.AIDS Schooling
##      4         Adult.Mortality HIV.AIDS Income.composition.of.resources Schooling
##      5         Adult.Mortality percentage.expenditure HIV.AIDS Income.composition.of.resources Schooli
##      6         Adult.Mortality percentage.expenditure BMI HIV.AIDS Income.composition.of.resources Sc
##      7         Adult.Mortality infant.deaths percentage.expenditure BMI HIV.AIDS Income.composition.
##      8         Adult.Mortality infant.deaths Alcohol percentage.expenditure BMI HIV.AIDS Income.compo
##      9         Adult.Mortality infant.deaths Alcohol percentage.expenditure Measles BMI HIV.AIDS Inco
##     10         Adult.Mortality infant.deaths Alcohol percentage.expenditure Measles BMI HIV.AIDS Popu
## ---------------------------------------------------------------------------------------------------------------
##
##                                                 Subsets Regression Summary
## ---------------------------------------------------------------------------------------------------------------
##                        Adj.        Pred
## Model    R-Square    R-Square    R-Square      C(p)          AIC          SBIC          SBC          6
## ---------------------------------------------------------------------------------------------------------------
##   1        0.5294      0.5292      0.5283    2701.7340    10612.7157    5929.8035    10628.9394    6
##   2        0.7304      0.7301      0.7286     847.5114     9696.3271    5014.4121     9717.9588    3
##   3        0.7871      0.7867      0.7853     325.7004     9308.9473    4627.9435     9335.9869    2
##   4        0.8092      0.8087      0.8071     123.7365     9130.3986    4450.0858     9162.8462    2
##   5        0.8147      0.8141      0.8125      74.5914     9083.8457    4403.7449     9121.7012    2
##   6        0.8201      0.8194      0.8177      27.2448     9037.6049    4357.8339     9080.8683    2
##   7        0.8213      0.8205      0.8188      17.7163     9028.1284    4348.4525     9076.7997    2
##   8        0.8222      0.8213      0.8194      11.4903     9021.8904    4342.3024     9075.9696    2
##   9        0.8224      0.8215      0.8192      11.2022     9021.5916    4342.0396     9081.0788    2
##  10        0.8227      0.8216      0.8193      11.0000     9021.3761    4341.8645     9086.2712    2
```

```
## ---------------------------------------------------------------------------------------------------
## AIC: Akaike Information Criteria
##  SBIC: Sawa's Bayesian Information Criteria
##  SBC: Schwarz Bayesian Criteria
##  MSEP: Estimated error of prediction, assuming multivariate normality
##  FPE: Final Prediction Error
##  HSP: Hocking's Sp
##  APC: Amemiya Prediction Criteria
```

Mejor Subconjunto de variables.

Este método presentado nos permite identidicar los modelos que con menor cantidad de variables predictoras logran tener un buen rendimiento.

- **R2 ajustado**: Buscamos máximinzar la función pues alor más alto indica que el modelo explica una mayor proporción de la variabilidad de la respuesta.
- **AIC**: . Un menor valor de AIC es mejor, en este caso el modelo de 11 variables también es el mejor.
- **SBIC**: Asi como el AIC se busca minimazar, el resutado más favorable considerando esta premisa es el modelo que considera 9 variables.
- **C(p)**: Como el BIC, este también nos dice que el mejor modelo es aquel que cuenta con 9 variables.

Selección de variables utilizando el método Forward.

```
modelo_forward <- ols_step_forward_p(modelo,  penter=0.08)
modelo_forward
```

```
##
##
##                                        Stepwise Summary
## ---------------------------------------------------------------------------------------------------
## Step    Variable                             AIC          SBC         SBIC        R2       Adj. R2
## ---------------------------------------------------------------------------------------------------
##  0      Base Model                         11853.803    11864.619    7171.152    0.00000    0.00000
##  1      Schooling                          10612.716    10628.939    5929.803    0.52945    0.52916
##  2      HIV.AIDS                            9696.327     9717.959    5014.412    0.73039    0.73006
##  3      Adult.Mortality                     9308.947     9335.987    4627.944    0.78710    0.78671
##  4      Income.composition.of.resources     9130.399     9162.846    4450.086    0.80918    0.80871
##  5      percentage.expenditure              9083.846     9121.701    4403.745    0.81471    0.81415
##  6      BMI                                 9037.605     9080.868    4357.834    0.82005    0.81940
##  7      infant.deaths                       9028.128     9076.800    4348.452    0.82130    0.82054
##  8      Alcohol                             9021.890     9075.970    4342.302    0.82219    0.82133
##  9      Measles                             9021.592     9081.079    4342.040    0.82244    0.82147
##  10     Population                          9021.376     9086.271    4341.865    0.82268    0.82160
## ---------------------------------------------------------------------------------------------------
##
## Final Model Output
## ------------------
##
##                            Model Summary
## ---------------------------------------------------------------
## R                       0.907       RMSE                 3.703
```

```
## R-Squared               0.823       MSE                 13.806
## Adj. R-Squared          0.822       Coef. Var            5.361
## Pred R-Squared          0.819       AIC                9021.376
## MAE                     2.830       SBC                9086.271
## ----------------------------------------------------------------
##  RMSE: Root Mean Square Error
##  MSE: Mean Square Error
##  MAE: Mean Absolute Error
##  AIC: Akaike Information Criteria
##  SBC: Schwarz Bayesian Criteria
##
##                                 ANOVA
## --------------------------------------------------------------------------
##                   Sum of
##                  Squares          DF    Mean Square       F         Sig.
## --------------------------------------------------------------------------
## Regression    104915.654          10      10491.565    759.947    0.0000
## Residual       22613.658        1638         13.806
## Total         127529.311        1648
## --------------------------------------------------------------------------
##
##                                          Parameter Estimates
## ------------------------------------------------------------------------------------------------------
##                           model      Beta    Std. Error    Std. Beta       t        Sig      lower
## ------------------------------------------------------------------------------------------------------
##                     (Intercept)    53.420         0.600                89.080    0.000    52.244
##                       Schooling     0.961         0.060        0.305   15.971    0.000     0.843
##                        HIV.AIDS    -0.439         0.018       -0.301  -24.003    0.000    -0.475
##                  Adult.Mortality    -0.018         0.001       -0.260  -18.880    0.000    -0.020
## Income.composition.of.resources    10.926         0.848        0.227   12.878    0.000     9.262
##          percentage.expenditure     0.000         0.000        0.091    7.670    0.000     0.000
##                             BMI     0.037         0.006        0.084    6.480    0.000     0.026
##                   infant.deaths    -0.004         0.001       -0.061   -3.716    0.000    -0.007
##                         Alcohol    -0.086         0.030       -0.040   -2.833    0.005    -0.146
##                        Measles     0.000         0.000        0.020    1.592    0.112     0.000
##                      Population     0.000         0.000        0.021    1.484    0.138     0.000
## ------------------------------------------------------------------------------------------------------
```

Al agregar progresivamente variables al modelo resulta que el modelo completo, considerando la mayoría de los indicadores es el que nos ofrecer una mejor predictibilidad.

Selección de variables utilizando el método Backward.

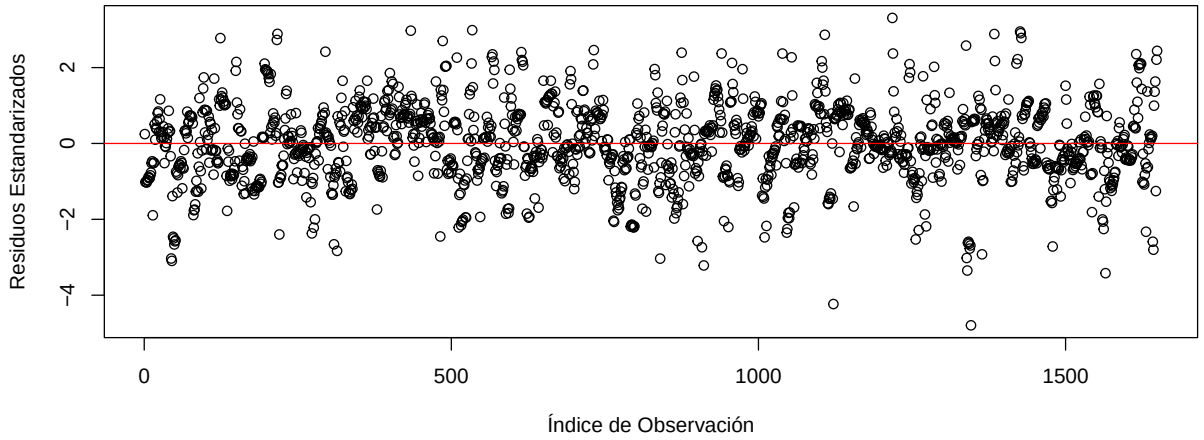
```
backward <- ols_step_backward_p(modelo)
backward
```

```
## [1] "No variables have been removed from the model."
```

Al remover variables del modelo, ninguno de los modelos resultantes termina por tener alguno de los indicadores en un estado más favorable, esto es consistente con los resultados en las metodologias anteriores.

Análisis de diagnostico através residuales estandarizados

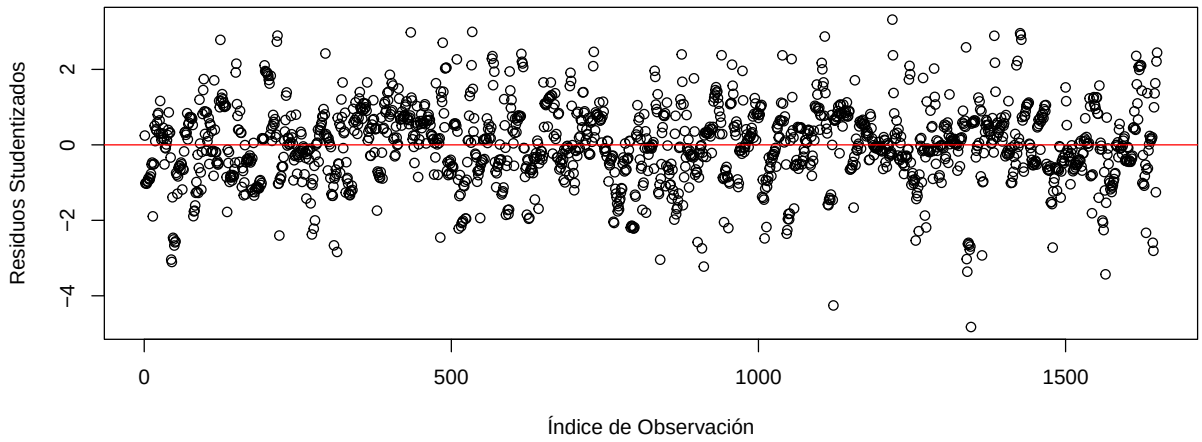
```
residuos_estandarizados <- rstandard(modelo)

# Graficar los residuos estandarizados
plot(residuos_estandarizados, ylab = "Residuos Estandarizados", xlab = "Índice de Observación")
abline(h = 0, col = "red")
```



Como se puede ver en la imagen los residuos estandarizados no muestran ningún tipo de patrón, lo que es un buen indicio considerando que la bondad de nuestro modelo, es decir la precisión y confiabilidad de un modelo estadístico. Sin embargo, es evidente la existencia de valores atípicos, pues existen varios puntos que se encuentran bien sea en un rango entre -4:-2 o entre 2:3 aproximadamente, lo que nos puede sugerir indagar más sobre el contexto y papel de los datos en este contexto.

Análisis de diagnostico através residuales studentizados.

```
# Calcular los residuos studentizados
residuos_studentizados <- rstudent(modelo)

# Graficar los residuos studentizados
plot(residuos_studentizados, ylab = "Residuos Studentizados", xlab = "Índice de Observación")
abline(h = 0, col = "red")
```



Por otro lado, tenemos los residuales studentizados que parece tener resultados consistentes con los estandarizados, parecen no existir patrones en los errores lo que nos sugiere que el modelo está capturando adecuadamente la variabilidad de los datos y que las suposiciones subyacentes del modelo. Además de que también se visualizan outliers que se encuentran en rangos similares a los ya mencionados.