Esperanza de vida en países en vías de desarrollo

Métodos Estadísticos para Ingeniería de Datos

Diciembre, 2023

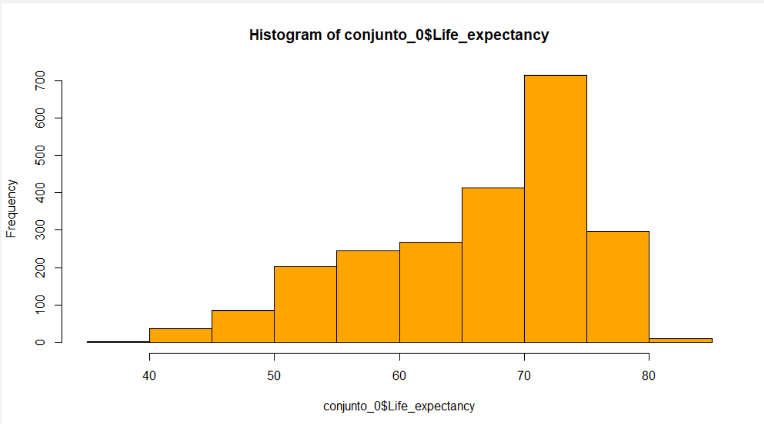
Óscar Marín Esteban
Álvaro Enol Alonso Ortega

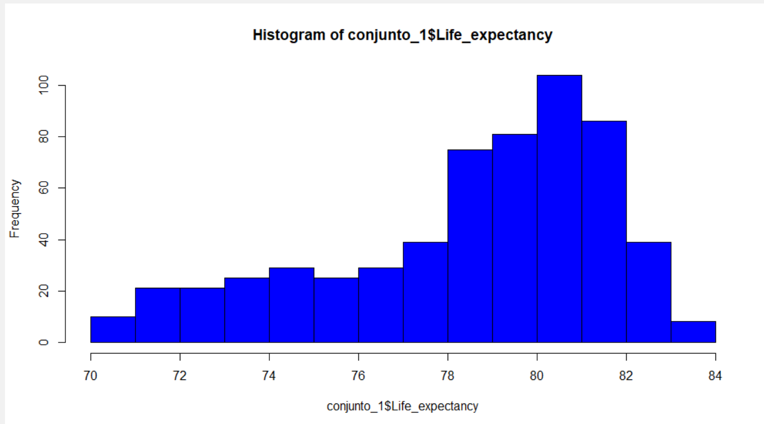
Análisis del conjunto

SEPARACIÓN DE LOS DATOS

Problema: nuestros datos no presentan una estructura lógica

- Hay una columna que separa nuestros datos entre países desarrollados y países en vías de desarrollo.
- Las condiciones entre estos tipos de países son muy diferentes. No podemos usar el mismo modelo para predecir los dos al mismo tiempo: cada uno de ellos sigue patrones distintos.
- Por esta razón, usaremos solo una parte de los datos: los países en vías de desarrollo.





Análisis del conjunto

PREVISUALIZACIÓN DE LOS DATOS

	Country	Region	Year	Infant_deaths	Under_five_deaths	Adult_mortality	Alcohol_consumption	Hepatitis_B	Measles	BMI	Polio	Diphtheria	Incidents_HIV	GDP_per_capita	Population_mln	Thinness_ten_nineteen_years	Thinness_five_nine_years	Schooling	Life_expectancy
1	Turkiye	Middle East	2015	11.1	13.0	105.8240	1.3200	97	65	27.8	97	97	0.08	11006	78.53	4.9	4.8	7.8	76.5
2	India	Asia	2007	51.5	67.9	201.0765	1.5700	60	35	21.2	67	64	0.13	1076	1183.21	27.1	28.0	5.0	65.4
3	Guyana	South America	2006	32.8	40.5	222.1965	5.6800	93	74	25.3	92	93	0.79	4146	0.75	5.7	5.5	7.9	67.0
4	Costa Rica	Central America and Caribbean	2006	9.8	11.2	95.2200	4.1900	88	86	26.4	89	89	0.16	9110	4.35	2.0	1.9	7.9	78.2
5	Russian Federation	Rest of Europe	2015	6.6	8.2	223.0000	8.0600	97	97	26.2	97	97	0.08	9313	144.10	2.3	2.3	12.0	71.2
6	Jordan	Middle East	2001	22.0	26.1	129.7640	0.5200	97	87	27.9	97	99	0.13	3708	5.22	4.0	3.9	9.6	71.9
7	Moldova	Rest of Europe	2008	15.3	17.8	217.8570	7.7200	97	92	26.5	96	90	0.43	2235	2.87	2.9	3.1	10.9	68.7
8	Brazil	South America	2012	15.4	17.2	150.2245	7.1200	96	70	26.1	96	95	0.24	9057	199.29	2.8	2.8	7.3	74.2
9	Bahamas, The	Central America and Caribbean	2011	13.0	15.2	165.5380	9.2300	95	83	27.6	97	98	0.46	32027	0.36	2.5	2.5	11.0	72.3
10	Ukraine	Rest of Europe	2002	14.3	16.6	261.6095	7.1300	48	98	25.8	99	99	0.55	1660	48.20	2.9	3.0	10.5	68.3
11	Comoros	Africa	2007	66.8	91.9	255.8815	0.1500	75	64	23.5	75	75	0.02	1166	0.64	7.3	7.2	3.5	60.7
12	Gabon	Africa	2012	39.1	57.1	256.8800	7.4700	82	64	24.9	80	82	1.57	7181	1.75	6.3	6.2	7.8	62.9
13	Ghana	Africa	2011	45.2	65.9	257.0865	1.6700	91	58	23.7	91	91	0.93	1580	25.39	6.9	6.8	6.8	61.4
14	Philippines	Asia	2001	28.2	36.9	214.2685	4.5300	45	19	22.3	76	79	0.01	1847	79.67	1.0	9.7	7.7	68.8
15	Congo, Rep.	Africa	2003	64.6	100.2	406.7020	0.9000	71	64	22.3	50	50	2.23	2072	3.41	9.1	8.8	5.7	53.8
16	Madagascar	Africa	2011	45.8	67.0	237.6755	0.8900	73	64	21.1	73	73	0.15	464	21.74	7.5	7.4	6.1	63.8

Análisis del conjunto

TIPOS DE LAS VARIABLES

<u>Cualitativas</u>

Nominales:
Country y *Region*

Ordinales:
Year

<u>Cuantitativas (continuas):</u>

Sanitarias:
Hepatitis_B
Measles
BMI
Incidents_HIV
Thinness_ten_nineteen_years
Thinness_five_nine_years

Económicas:
GPD_per_capita

Naturales:
Infant_deaths
Under_five_deaths
Adult_mortality
Life_Expectancy

Sociales:
Alcohol_consumption
Population_mln
Schooling

Depuración de los datos

VALORES NULOS Y SIMETRÍA

Na.omit()

Simetría:

```
[1] "Infant_deaths"
[1] 0.9196311
```

```
[1] "Year"
[1] 0
```

```
[1] "Under_five_deaths"
[1] 1.154998
```

```
[1] "Adult_mortality"
[1] 1.300519
```

```
[1] "Alcohol_consumption"
[1] 0.8650686
```

```
[1] "Hepatitis_B"
[1] -1.377481
```

```
[1] "Measles"
[1] -0.7653451
```

```
[1] "BMI"
[1] 0.1620363
```

```
[1] "Polio"
[1] -1.428882
```

```
[1] "Diphtheria"
[1] -1.52894
```

```
[1] "Incidents_HIV"
[1] 4.407009
```

```
[1] "GDP_per_capita"
[1] 3.649479
```

```
[1] "Population_mln"
[1] 7.54128
```

```
[1] "Thinness_ten_nineteen_years"
[1] 1.547902
```

```
[1] "Thinness_five_nine_years"
[1] 1.631239
```

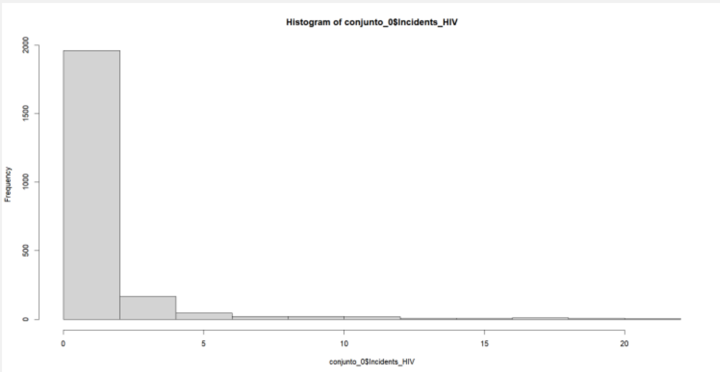
```
[1] "Schooling"
[1] -0.0516788
```

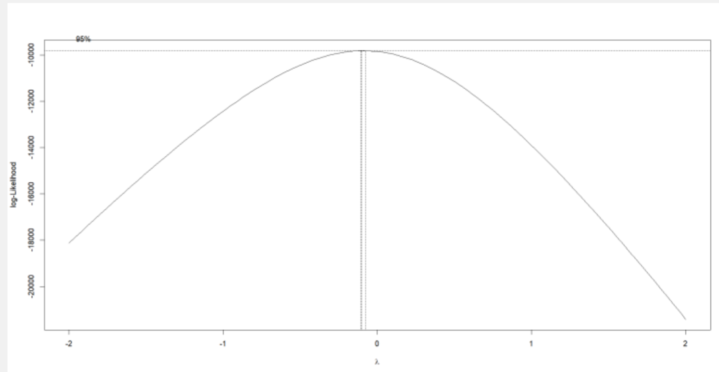
```
[1] "Life_expectancy"
[1] -0.7258087
```

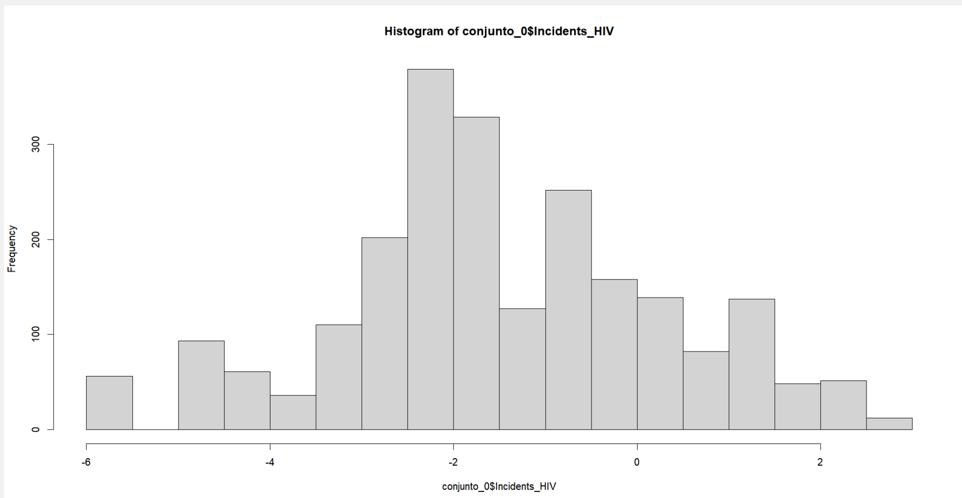
Depuración de los datos

TRANSFORMACIONES BOXCOX

Incidents_HIV:



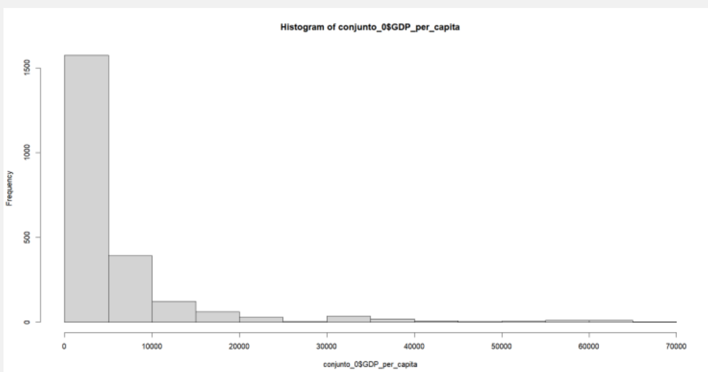


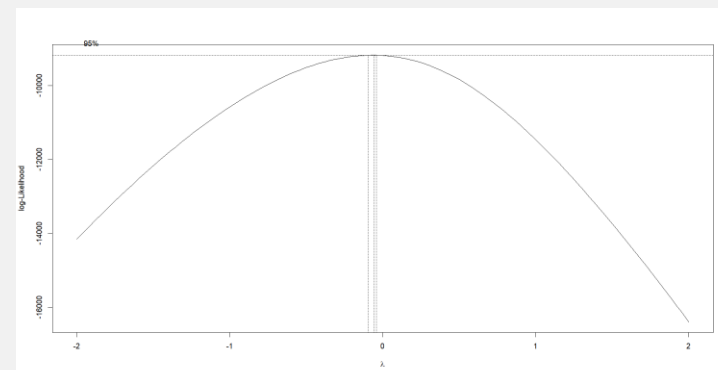


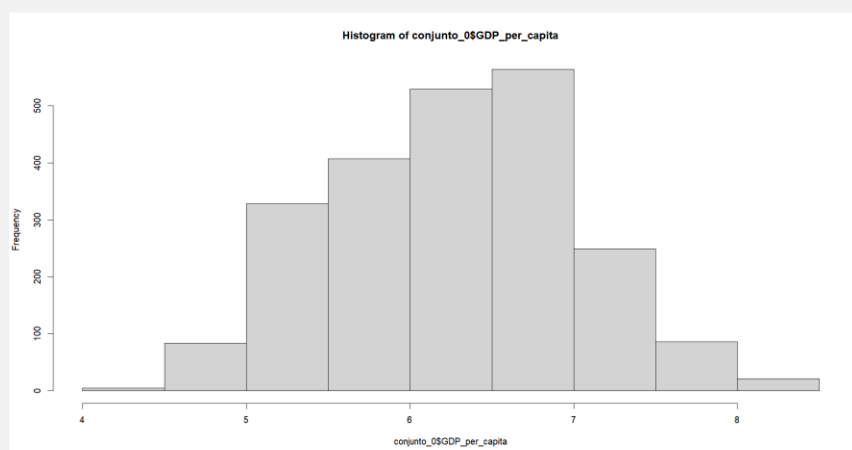
Depuración de los datos

TRANSFORMACIONES BOXCOX

GPD_per_capita



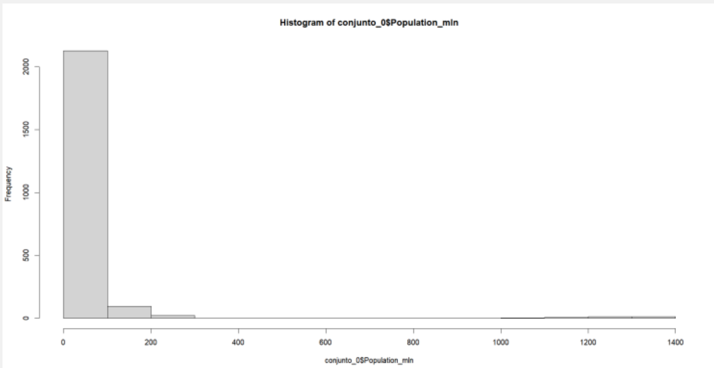


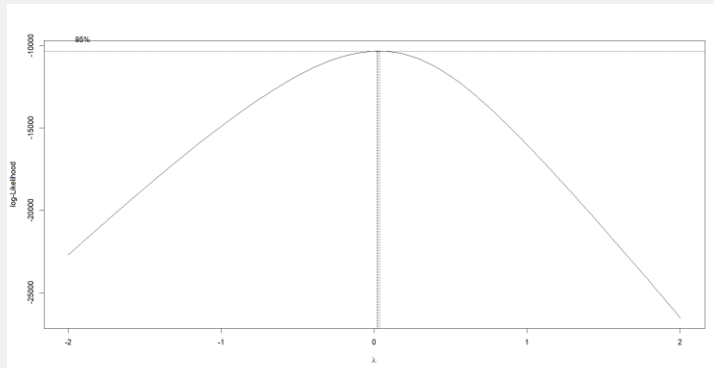


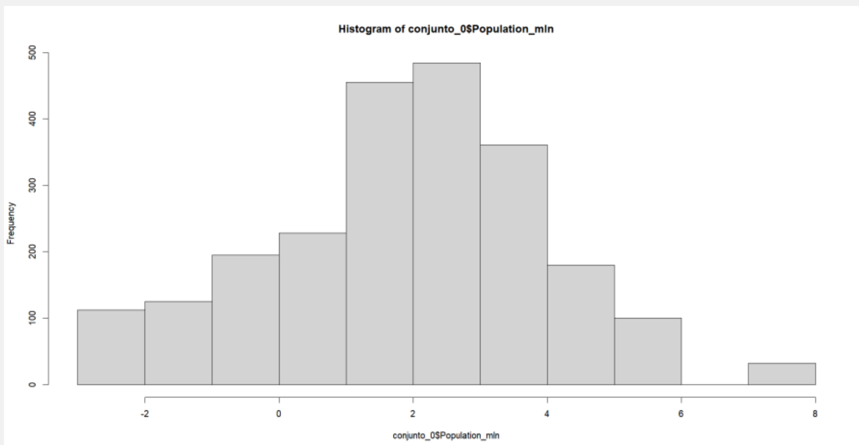
Depuración de los datos

TRANSFORMACIONES BOXCOX

Population_mln:

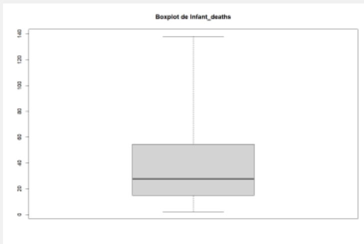


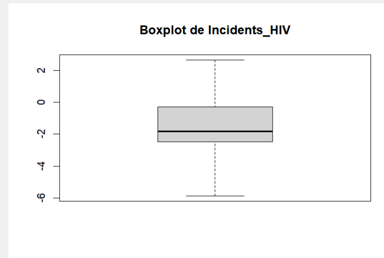


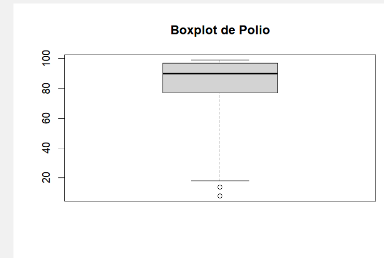


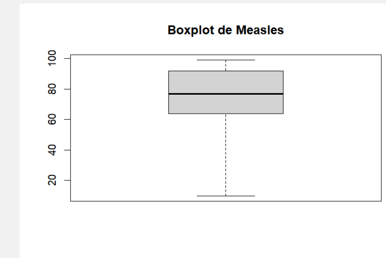
Depuración de los datos

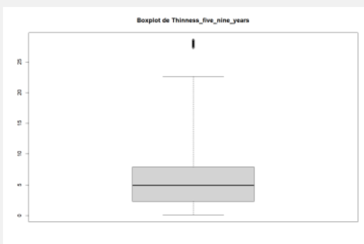
OUTLIERS

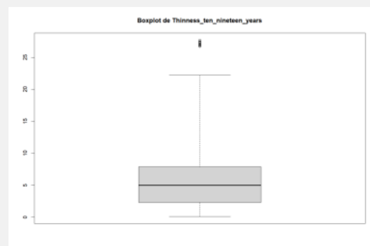


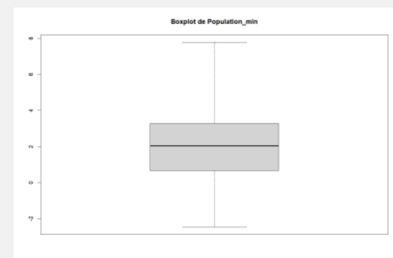


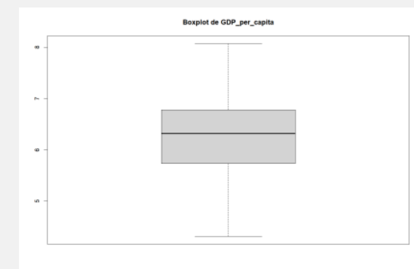


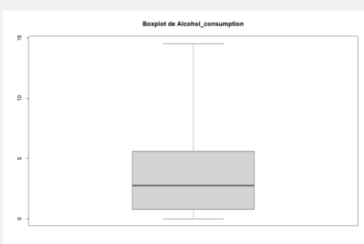


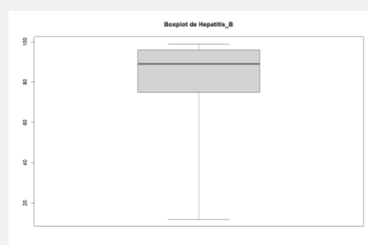


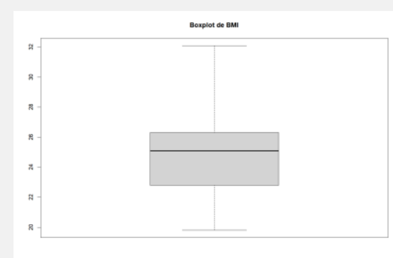


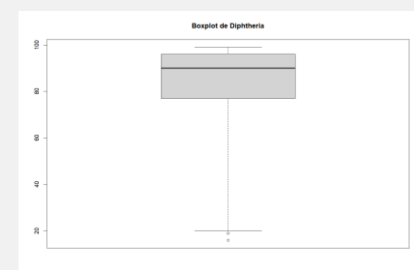


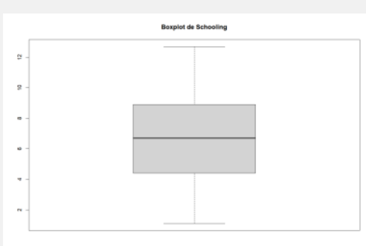


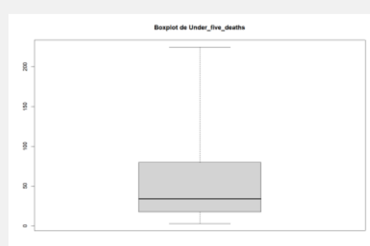


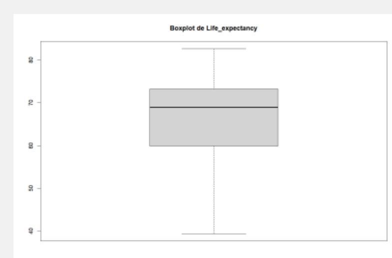


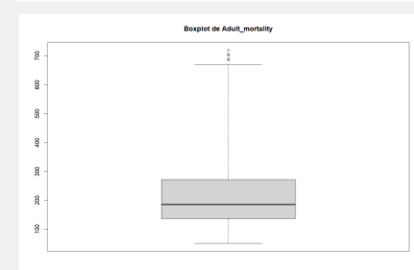






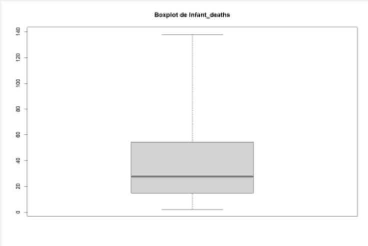


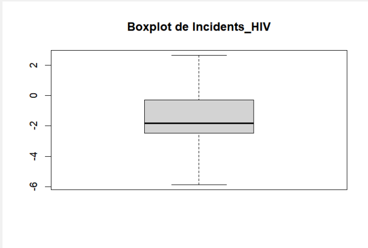


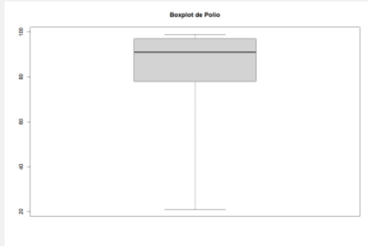


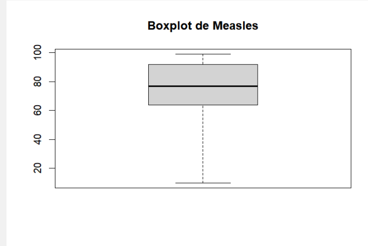
Depuración de los datos

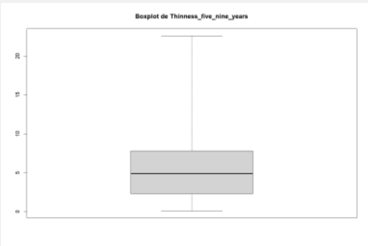
OUTLIERS

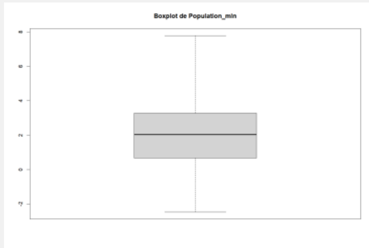


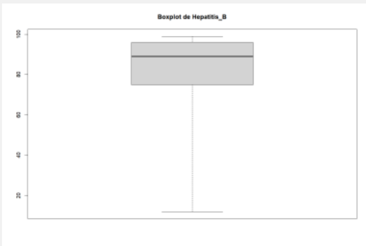


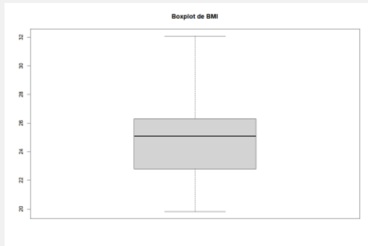


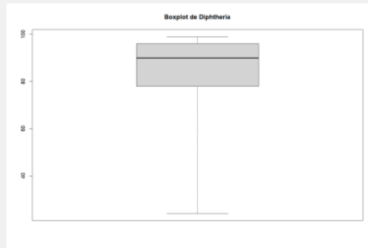


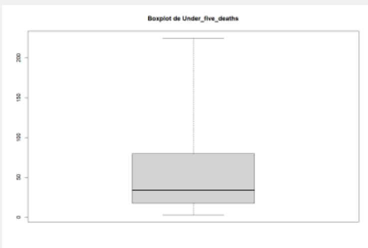


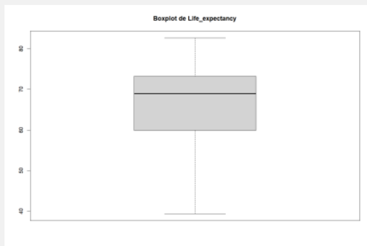


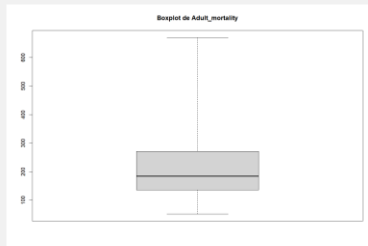












Distribución de las variables

NORMALIDAD DE LAS VARIABLES

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.0886, p-value < 2.2e-16

[1] "Year"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.14172, p-value < 2.2e-16

[1] "Infant_deaths"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.16982, p-value < 2.2e-16

[1] "Under_five_deaths"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.11513, p-value < 2.2e-16

[1] "Adult_mortality"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.12943, p-value < 2.2e-16

[1] "Alcohol_consumption"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.16772, p-value < 2.2e-16

[1] "Hepatitis_B"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.13501, p-value < 2.2e-16

[1] "Measles"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.069743, p-value < 2.2e-16

[1] "BMI"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.18106, p-value < 2.2e-16

[1] "Polio"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.18246, p-value < 2.2e-16

[1] "Diphtheria"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.33832, p-value < 2.2e-16

[1] "Incidents_HIV"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.26757, p-value < 2.2e-16

[1] "GDP_per_capita"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.39777, p-value < 2.2e-16

[1] "Population_mln"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.11553, p-value < 2.2e-16

[1] "Thinness_ten_nineteen_years"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.1158, p-value < 2.2e-16

[1] "Thinness_five_nine_years"
```

```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.051681, p-value = 2.001e-15

[1] "Schooling"
```

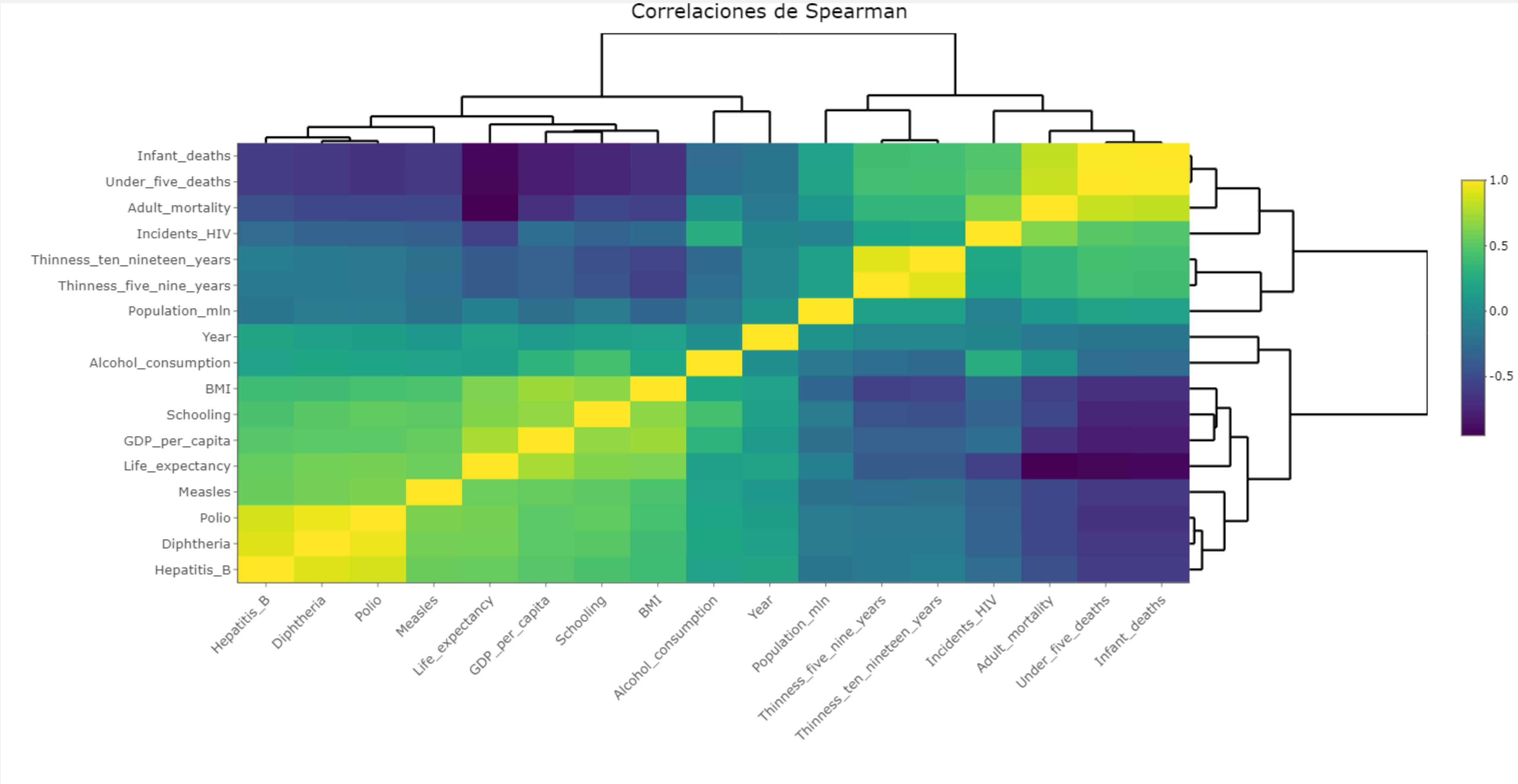
```
	Lilliefors (Kolmogorov-Smirnov) normality test

data:  columnas_numericas[, i]
D = 0.121, p-value < 2.2e-16

[1] "Life_expectancy"
```

Correlaciones

CORRELACIONES ENTRE VARIABLES



Correlaciones entre las variables

CORRELACIONES ENTRE VARIABLES

Utilizamos Spearman porque las variables no siguen distribuciones normales.

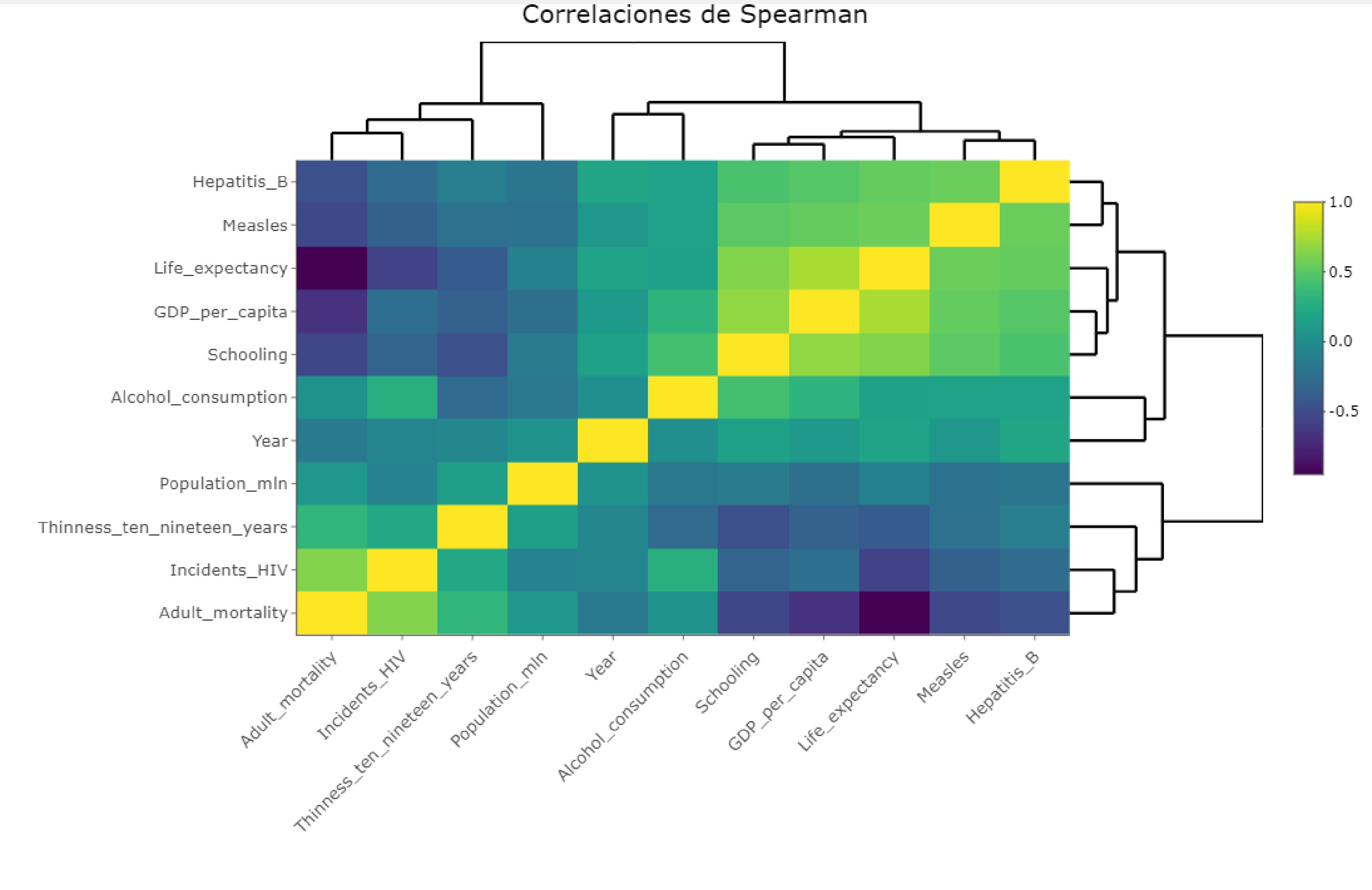
Tests:

- *Infant_deaths – Under_five_deaths* : p-value < 2.2e-16 (quitamos Infant_deaths)
- *Adult_mortality – Under_five_deaths*: p-value < 2.2e-16 (quitamos Under_five_deaths)
- *Polio – Diphteria*: p-value < 2.2e-16 (quitamos Diphtheria)
- *Hepatitis_B* - Polio: p-value < 2.2e-16 (quitamos Polio)
- *Thinness_five_nine_years – Thinness_ten_nineteen_years* : p-value < 2.2e-16 (quitamos Thinness_five_nine_years)
- *BMI – GPD_per_capita*: p-value < 2.2e-16 (quitamos BMI)

Todas las correlaciones son significativas, por lo que **podemos** tener en cuenta su correlación.

Correlaciones

CORRELACIONES ENTRE VARIABLES



REGRESIÓN LINEAL MÚLTIPLE (RLM)

¿ESTADÍSTICAMENTE SIGNIFICATIVAS?

```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)        61.2943251  0.6801629  90.117  < 2e-16 ***
Adult_mortality    -0.0585091  0.0006764 -86.500  < 2e-16 ***
Alcohol_consumption 0.1464876  0.0167562   8.742  < 2e-16 ***
Hepatitis_B         0.0490927  0.0031866  15.406  < 2e-16 ***
Incidents_HIV      -0.1183423  0.0359621  -3.291  0.00101 ** 
GDP_per_capita      1.5664005  0.0942647  16.617  < 2e-16 ***
Population_mln      0.1495752  0.0227576   6.573 6.14e-11 ***
Schooling           0.4315321  0.0235150  18.351  < 2e-16 ***
```

R^2 ajustado: 0.9779

Todas las variables que entran en el modelo son estadísticamente significativas.

REGRESIÓN LINEAL MÚLTIPLE (RLM)

DIVISIÓN ENTRE TRAIN Y TEST

TRAIN(70%)	TEST(30%)
1566	672

REGRESIÓN LINEAL MÚLTIPLE (RLM)

ENTRENAMIENTO DEL MODELO

Hacemos validación cruzada

Comprobamos hipótesis:

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  model.cv$residuals
D = 0.039311, p-value = 6.659e-06
```

```
        Durbin-Watson test

data:  model.cv
DW = 1.9189, p-value = 0.05418
```

```
        studentized Breusch-Pagan test

data:  model.cv
BP = 138.8, df = 7, p-value < 2.2e-16
```

REGRESIÓN LINEAL MÚLTIPLE (RLM)

ENTRENAMIENTO DEL MODELO

Quitamos outliers con alto leverage:

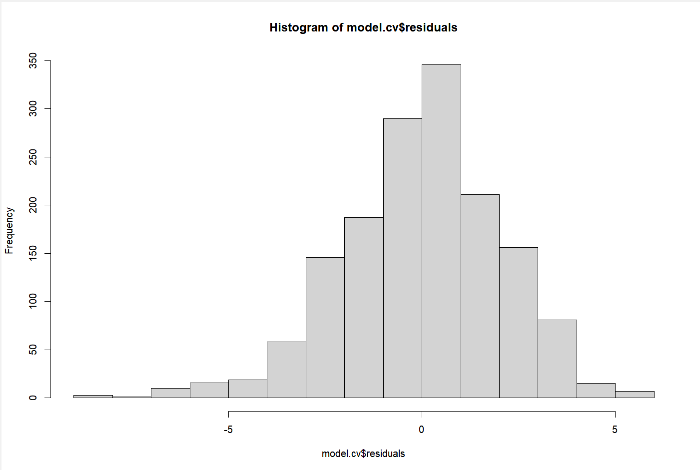
```
> (outliers=c(which(rest>3),which(rest<(-3))))
1564 2021 1790  131  735 2243  936 2023  308 2270  742 1558 1439
 181  373  463  512  582  672  870  901 1176 1401 1402 1469 1475
> rest[outliers] #obtenemos el valor del residuo estudentizado
     1564      2021      1790       131       735      2243       936      2023       308      2270
-3.289264 -3.741670 -3.271437 -4.086701 -3.121916 -3.348546 -4.064600 -3.110368 -3.394000 -3.231005
      742      1558      1439
-3.241067 -3.982211 -3.281908
```

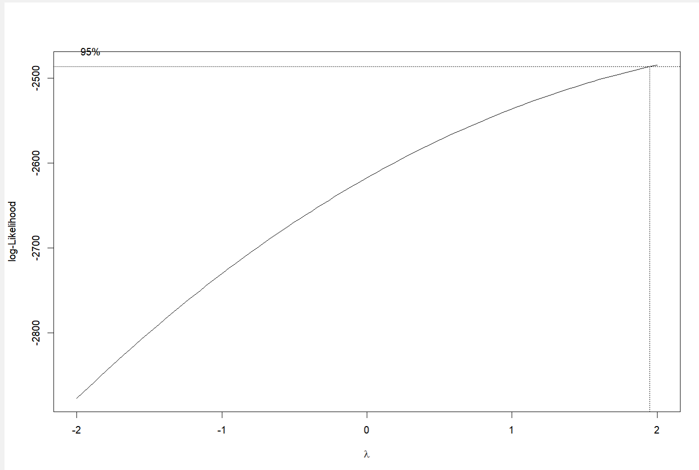
Quitamos 1 a 1 y vamos hacienda validación cruzada

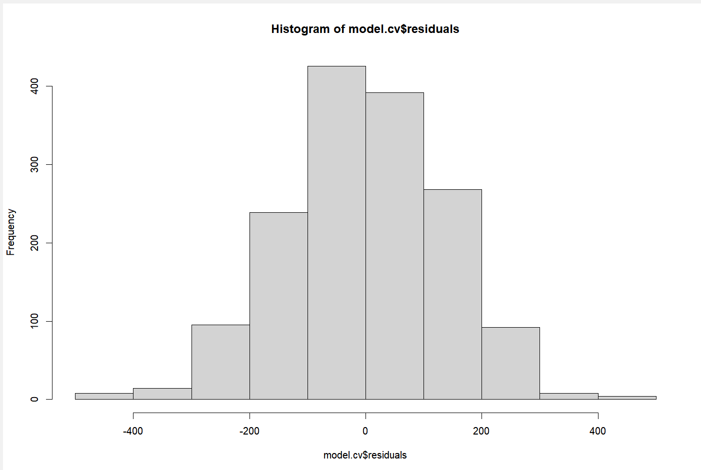
REGRESIÓN LINEAL MÚLTIPLE (RLM)

ENTRENAMIENTO DEL MODELO

Boxcox:







REGRESIÓN LINEAL MÚLTIPLE (RLM)

ENTRENAMIENTO DEL MODELO

Tras hacer boxcox comprobamos las variables:

```
        Lilliefors (Kolmogorov-Smirnov) normality test

data:  model.cv$residuals
D = 0.01606, p-value = 0.4331
```

```
        Durbin-Watson test

data:  model.cv
DW = 1.9183, p-value = 0.05385
```

```
        studentized Breusch-Pagan test

data:  model.cv
BP = 123.27, df = 7, p-value < 2.2e-16
```

REGRESIÓN LINEAL MÚLTIPLE (RLM)

ENTRENAMIENTO DEL MODELO

Quitamos outliers con alto leverage:

```
> (outliers=c(which(rest>3),which(rest<(-3))))
1564 2021 1790  131  735 2243  936 2023  308 2270  742 1558 1439
 181  373  463  512  582  672  870  901 1176 1401 1402 1469 1475
> rest[outliers] #obt
     1564      2021        Lilliefors (Kolmogorov-Smirnov) normality test          308      2270
-3.289264 -3.741670 -                                                         -3.394000 -3.231005
      742      1558   data:  model.cv$residuals
-3.241067 -3.982211 - D = 0.01606, p-value = 0.4331
```

Quitamos 1 a 1 y hacemos validación cruzada

REGRESIÓN LINEAL MÚLTIPLE (RLM)

ECM RSME

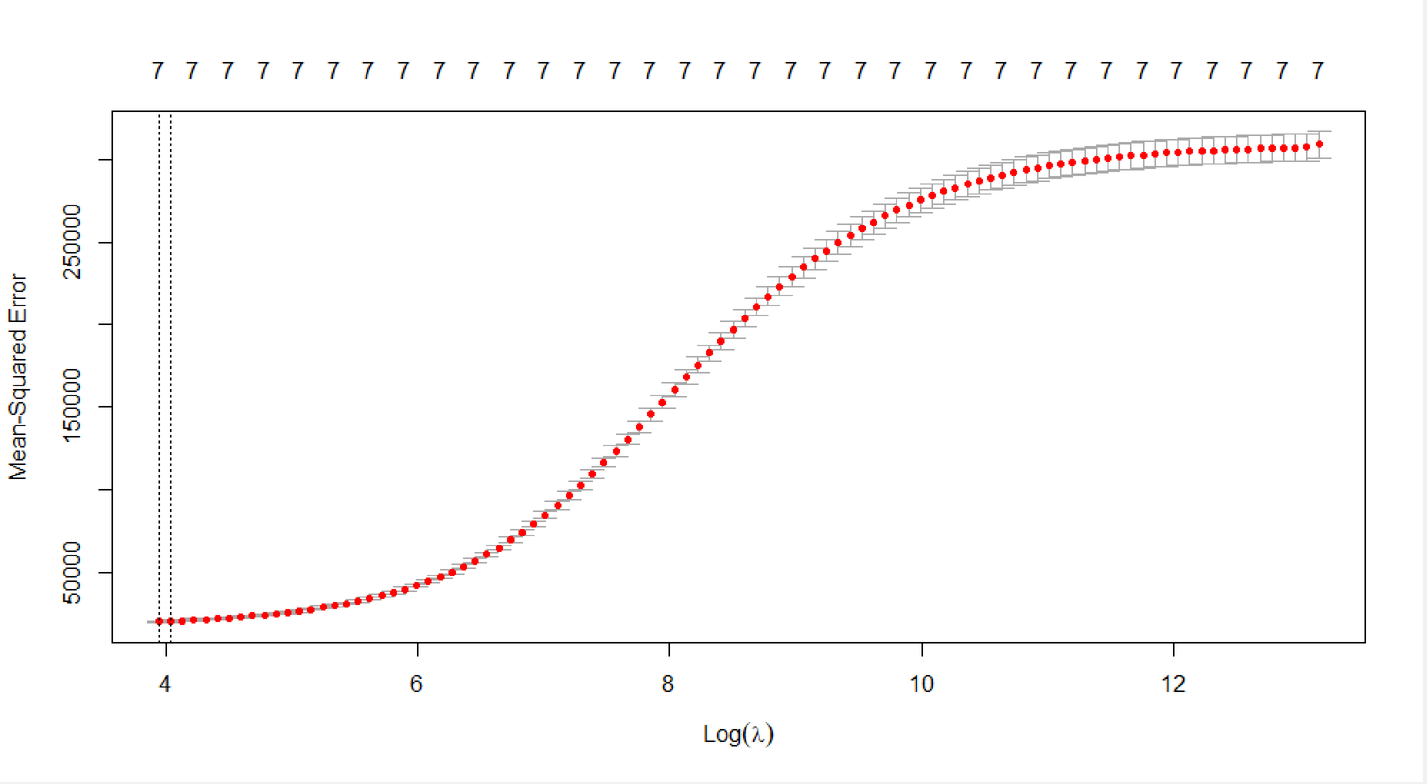
ECM: 1.306823

RMSE: 1.143164

MODELO	ADULT MORTALITY	ALCOHOL CONSUMPTION	HEPATITIS B	MEASLES	INCIDENTS HIV	GPD PER CAPITA	POPULATION MLN	THINNESS TEN NINETEEN YEARS	SCHOOLING
RLM	✓	✓	✓		✓	✓	✓		✓

MÉTODOS DE REGULARIZACIÓN

RIDGE



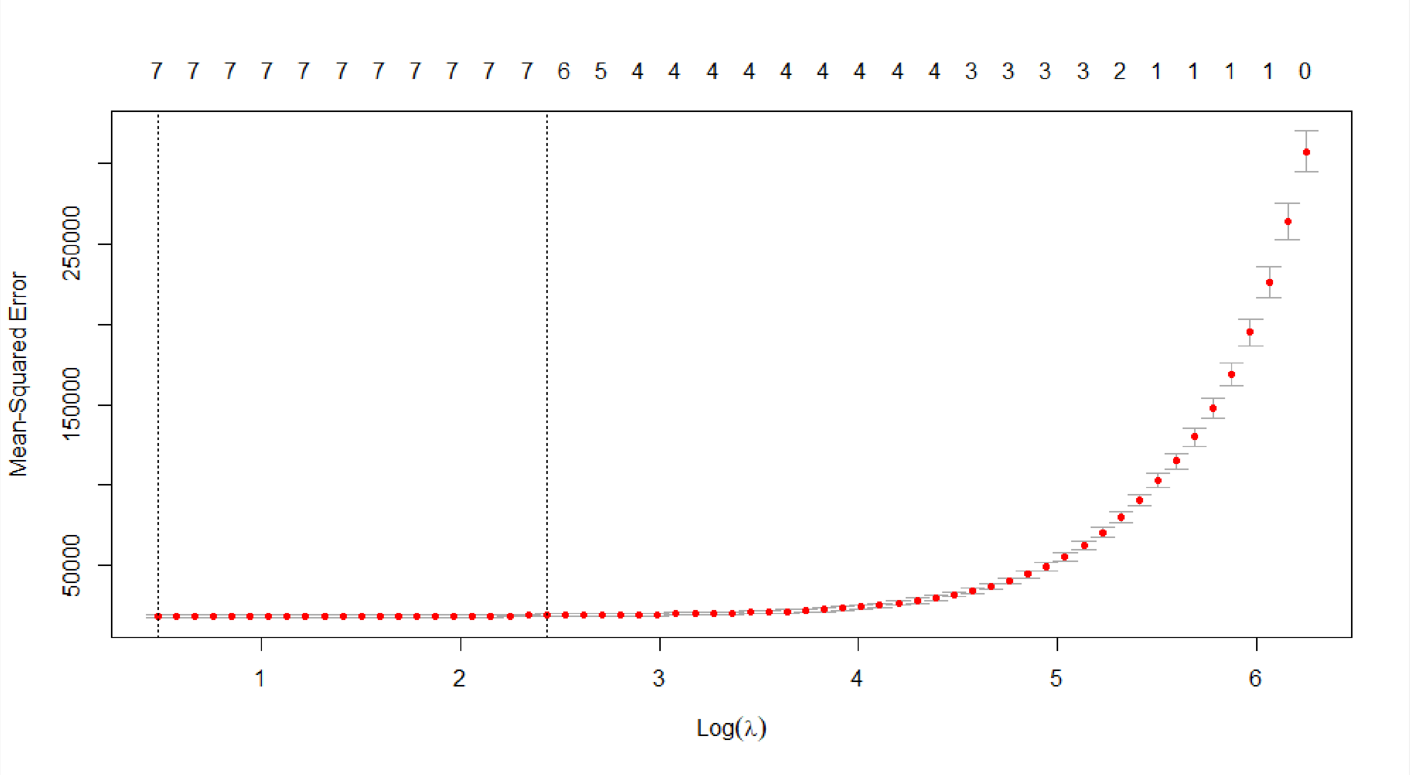
Lambda: 51.14625

ECM: 1.897826

RMSE: 1.377616

MÉTODOS DE REGULARIZACIÓN

LASSO



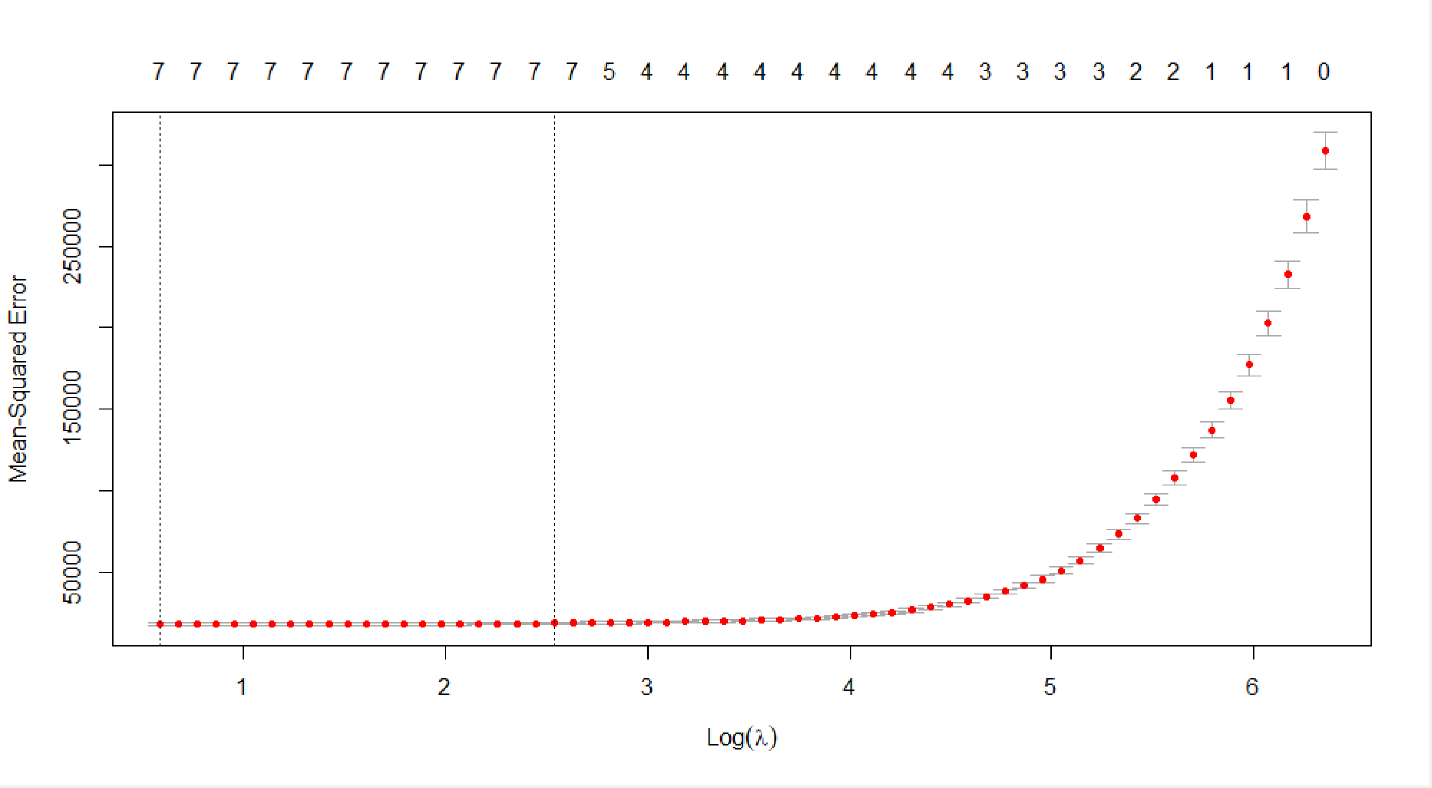
Lambda: 1.598686

ECM: 1.451612

RMSE: 1.204829

MÉTODOS DE REGULARIZACIÓN

ELASTIC NET



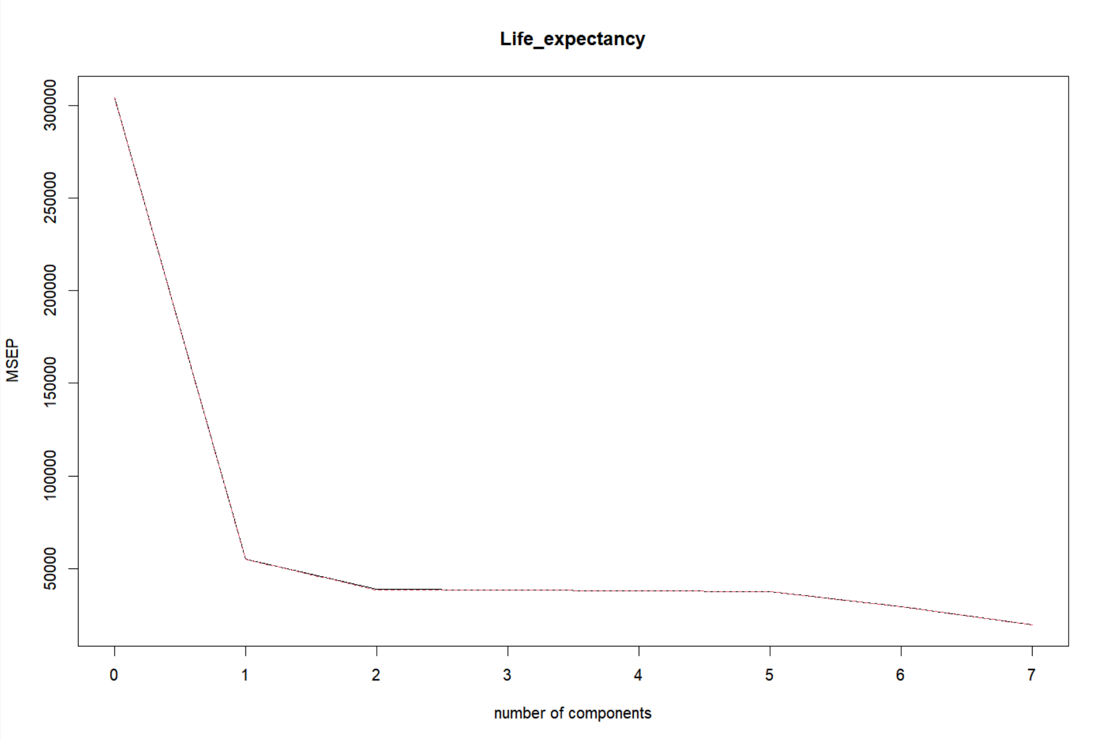
Lambda: 1.776318

ECM: 1.453092

RMSE: 1.205443

MÉTODOS DE REDUCCION DE DIMENSIONALIDADES

PCR



1 COMP: 3.16827

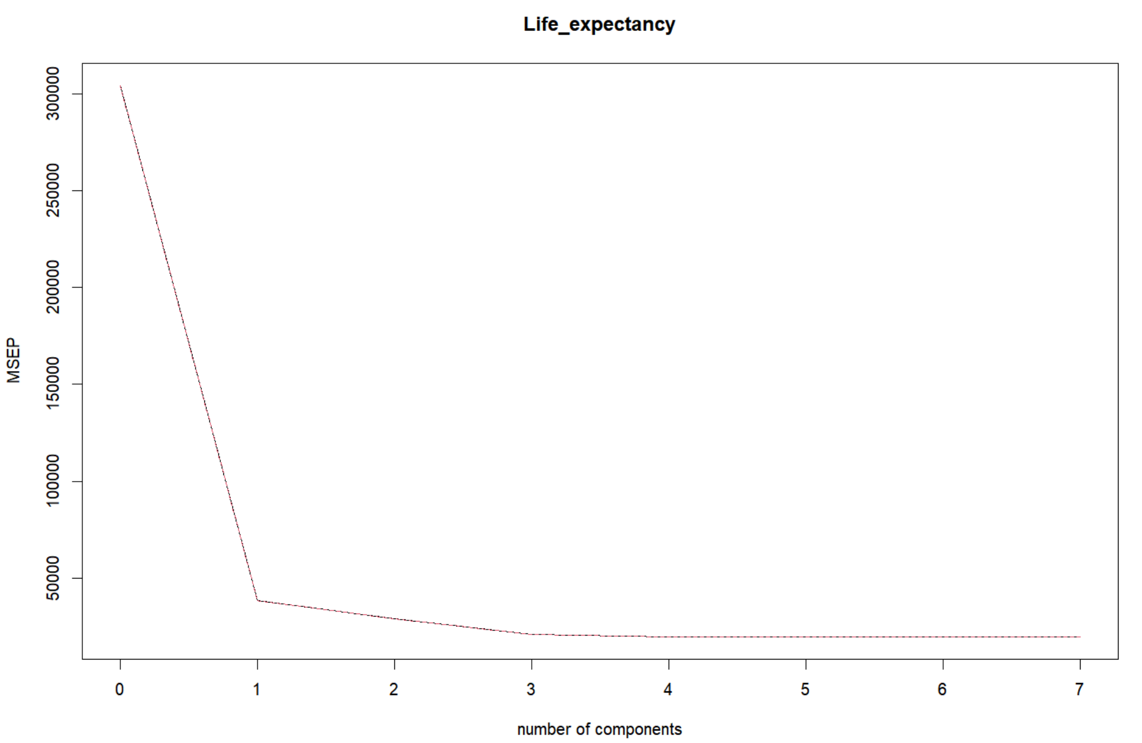
2 COMPS: 1.126047

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
X	40.51	62.50	76.01	85.87	92.58	97.33	100.00
Life_expectancy	82.03	87.41	87.54	87.69	87.91	90.51	93.74

MÉTODOS DE REDUCCION DE DIMENSIONALIDADES

PLS



1 COMP: 1.922561

3 COMPS: 0.6482499

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps
X	40.13	60.16	67.20	76.85	84.25	92.94	100.00
Life_expectancy	87.52	90.60	93.32	93.63	93.73	93.74	93.74

CONCLUSIONES

MODELO GANADOR

PLS CON 3 COMPS

```
Data:    X dimension: 1545 7
         Y dimension: 1545 1
Fit method: kernelpls
Number of components considered: 3

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps
CV             556    197.9    168.7    139.1
adjCV          556    197.9    168.7    139.0

TRAINING: % variance explained
                 1 comps  2 comps  3 comps
X                  39.40    59.25    66.10
Life_expectancy    87.37    90.88    93.83
```

Thinness_five_nine_years/Thinness_ten_nineteen_years

Under_five_deaths/Infant_deaths

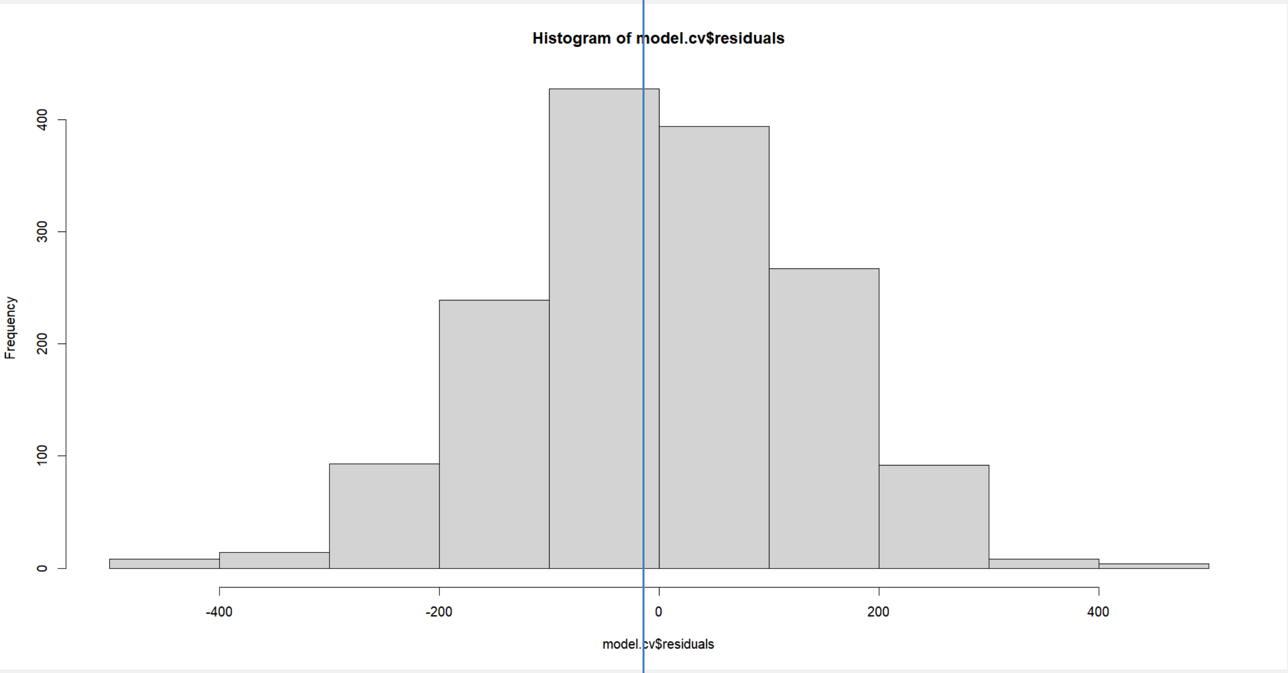
REGRESIÓN LOGÍSTICA

DIVISIÓN

```
> mediana_vida
[1] 69.1
```

PAÍSES MAS SUBDESARROLLADOS (0)

PAÍSES MENOS SUBDESARROLLADOS (1)



REGRESIÓN LOGÍSTICA

CONCLUSIONES

Accuracy: 0.9489

Sensibilidad (verdaderos positivos): 0.9433
Especificidad (verdaderos negativos): 0.9466

MATRIZ DE CONFUSION

	0	1
0	316	18
1	19	319

Accuracy: 0.945

BIBLIOGRAFÍA

Prácticas y documentos Campus Virtual

CONJUNTO DATOS: https://www.kaggle.com/datasets/lashagoch/life-expectancy-who-updated