PEC4: Análisis de Varianza y Repaso del Curso

Alex Stwart Porras Palma

2025-01-17

Contents

1 Preprocesado

1.1 Variable Region

```
df <- read.csv("Life-Expectancy-Data-Updated.csv", header = TRUE, sep = ',')
```

```
head(df,5)
```

```
##   Country         Region Year Infant_deaths Under_five_deaths Adult_mortality
## 1 Turkiye    Middle East 2015          11.1              13.0        105.8240
## 2   Spain European Union 2015           2.7               3.3         57.9025
## 3   India           Asia 2007          51.5              67.9        201.0765
## 4  Guyana  South America 2006          32.8              40.5        222.1965
## 5  Israel    Middle East 2012           3.4               4.3         57.9510
##   Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria Incidents_HIV
## 1                1.32          97      65 27.8    97         97          0.08
## 2               10.35          97      94 26.0    97         97          0.09
## 3                1.57          60      35 21.2    67         64          0.13
## 4                5.68          93      74 25.3    92         93          0.79
## 5                2.89          97      89 27.0    94         94          0.08
##   GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          11006          78.53                         4.9
## 2          25742          46.44                         0.6
## 3           1076        1183.21                        27.1
## 4           4146           0.75                         5.7
## 5          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
## 1                      4.8       7.8                       0
## 2                      0.5       9.7                       1
## 3                     28.0       5.0                       0
## 4                      5.5       7.9                       0
## 5                      1.1      12.8                       1
##   Economy_status_Developing Life_expectancy
## 1                         1            76.5
## 2                         0            82.8
## 3                         1            65.4
## 4                         1            67.0
## 5                         0            81.7
```

```
summary(df)
```

```
##    Country             Region               Year      Infant_deaths
##  Length:2864        Length:2864        Min.   :2000   Min.   :  1.80
##  Class :character   Class :character   1st Qu.:2004   1st Qu.:  8.10
##  Mode  :character   Mode  :character   Median :2008   Median : 19.60
##                                        Mean   :2008   Mean   : 30.36
##                                        3rd Qu.:2011   3rd Qu.: 47.35
##                                        Max.   :2015   Max.   :138.10
##  Under_five_deaths Adult_mortality  Alcohol_consumption  Hepatitis_B
##  Min.   :  2.300   Min.   : 49.38   Min.   : 0.000      Min.   :12.00
##  1st Qu.:  9.675   1st Qu.:106.91   1st Qu.: 1.200      1st Qu.:78.00
##  Median : 23.100   Median :163.84   Median : 4.020      Median :89.00
##  Mean   : 42.938   Mean   :192.25   Mean   : 4.821      Mean   :84.29
##  3rd Qu.: 66.000   3rd Qu.:246.79   3rd Qu.: 7.777      3rd Qu.:96.00
##  Max.   :224.900   Max.   :719.36   Max.   :17.870      Max.   :99.00
```

```
##     Measles           BMI             Polio        Diphtheria
##  Min.   :10.00   Min.   :19.80   Min.   : 8.0   Min.   :16.00
##  1st Qu.:64.00   1st Qu.:23.20   1st Qu.:81.0   1st Qu.:81.00
##  Median :83.00   Median :25.50   Median :93.0   Median :93.00
##  Mean   :77.34   Mean   :25.03   Mean   :86.5   Mean   :86.27
##  3rd Qu.:93.00   3rd Qu.:26.40   3rd Qu.:97.0   3rd Qu.:97.00
##  Max.   :99.00   Max.   :32.10   Max.   :99.0   Max.   :99.00
##  Incidents_HIV     GDP_per_capita   Population_mln
##  Min.   : 0.0100   Min.   :   148   Min.   :   0.080
##  1st Qu.: 0.0800   1st Qu.:  1416   1st Qu.:   2.098
##  Median : 0.1500   Median :  4217   Median :   7.850
##  Mean   : 0.8943   Mean   : 11541   Mean   :  36.676
##  3rd Qu.: 0.4600   3rd Qu.: 12557   3rd Qu.:  23.688
##  Max.   :21.6800   Max.   :112418   Max.   :1379.860
##  Thinness_ten_nineteen_years Thinness_five_nine_years   Schooling
##  Min.   : 0.100              Min.   : 0.1             Min.   : 1.100
##  1st Qu.: 1.600              1st Qu.: 1.6             1st Qu.: 5.100
##  Median : 3.300              Median : 3.4             Median : 7.800
##  Mean   : 4.866              Mean   : 4.9             Mean   : 7.632
##  3rd Qu.: 7.200              3rd Qu.: 7.3             3rd Qu.:10.300
##  Max.   :27.700              Max.   :28.6             Max.   :14.100
##  Economy_status_Developed Economy_status_Developing Life_expectancy
##  Min.   :0.0000           Min.   :0.0000            Min.   :39.40
##  1st Qu.:0.0000           1st Qu.:1.0000            1st Qu.:62.70
##  Median :0.0000           Median :1.0000            Median :71.40
##  Mean   :0.2067           Mean   :0.7933            Mean   :68.86
##  3rd Qu.:0.0000           3rd Qu.:1.0000            3rd Qu.:75.40
##  Max.   :1.0000           Max.   :1.0000            Max.   :83.80
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
df <- df %>%
  mutate(Region = case_when(
    Region == "European Union" ~ "Europe",
    Region == "Rest of Europe" ~ "Europe",
    Region %in% c("Central America and Caribbean", "South America") ~
      "South-Central America",
    TRUE ~ Region
  ))
```

```
head(df,5)
```

```
##   Country               Region Year Infant_deaths Under_five_deaths
## 1 Turkiye           Middle East 2015          11.1              13.0
```

```
## 2   Spain                Europe 2015           2.7              3.3
## 3   India                  Asia 2007          51.5             67.9
## 4  Guyana South-Central America 2006          32.8             40.5
## 5  Israel           Middle East 2012           3.4              4.3
##   Adult_mortality Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria
## 1        105.8240                1.32          97      65 27.8    97         97
## 2         57.9025               10.35          97      94 26.0    97         97
## 3        201.0765                1.57          60      35 21.2    67         64
## 4        222.1965                5.68          93      74 25.3    92         93
## 5         57.9510                2.89          97      89 27.0    94         94
##   Incidents_HIV GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          0.08          11006          78.53                         4.9
## 2          0.09          25742          46.44                         0.6
## 3          0.13           1076        1183.21                        27.1
## 4          0.79           4146           0.75                         5.7
## 5          0.08          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
## 1                      4.8       7.8                        0
## 2                      0.5       9.7                        1
## 3                     28.0       5.0                        0
## 4                      5.5       7.9                        0
## 5                      1.1      12.8                        1
##   Economy_status_Developing Life_expectancy
## 1                         1            76.5
## 2                         0            82.8
## 3                         1            65.4
## 4                         1            67.0
## 5                         0            81.7
```

1.2 Variable Life_Expectancy

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.4.2
```

```
Q1 <- quantile(df$Life_expectancy, 0.25, na.rm = TRUE)
Q3 <- quantile(df$Life_expectancy, 0.75, na.rm = TRUE)
IQR <- Q3 - Q1
lower_bound <- Q1 - 1.5 * IQR
upper_bound <- Q3 + 1.5 * IQR
df <- df%>%
  mutate(Outliers = Life_expectancy < lower_bound |
           Life_expectancy > upper_bound)
outliers <- df %>%
  filter(Outliers) %>%
  select(Country, Region, Year, Life_expectancy) %>%
  arrange(Country, Year)
```

```
head(df,5)
```

```
##   Country                Region Year Infant_deaths Under_five_deaths
## 1 Turkiye           Middle East 2015          11.1              13.0
## 2   Spain                Europe 2015           2.7               3.3
## 3   India                  Asia 2007          51.5              67.9
## 4  Guyana South-Central America 2006          32.8              40.5
```

```
## 5  Israel           Middle East 2012           3.4               4.3
##   Adult_mortality Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria
## 1        105.8240                1.32          97      65 27.8    97         97
## 2         57.9025               10.35          97      94 26.0    97         97
## 3        201.0765                1.57          60      35 21.2    67         64
## 4        222.1965                5.68          93      74 25.3    92         93
## 5         57.9510                2.89          97      89 27.0    94         94
##   Incidents_HIV GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          0.08          11006          78.53                         4.9
## 2          0.09          25742          46.44                         0.6
## 3          0.13           1076        1183.21                        27.1
## 4          0.79           4146           0.75                         5.7
## 5          0.08          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
## 1                      4.8       7.8                        0
## 2                      0.5       9.7                        1
## 3                     28.0       5.0                        0
## 4                      5.5       7.9                        0
## 5                      1.1      12.8                        1
##   Economy_status_Developing Life_expectancy Outliers
## 1                         1           76.5    FALSE
## 2                         0           82.8    FALSE
## 3                         1           65.4    FALSE
## 4                         1           67.0    FALSE
## 5                         0           81.7    FALSE
```

```
head(outliers,5)
```

```
##    Country Region Year Life_expectancy
## 1 Eswatini Africa 2003           43.4
## 2 Eswatini Africa 2004           42.7
## 3 Eswatini Africa 2005           42.5
## 4 Eswatini Africa 2006           42.7
## 5 Eswatini Africa 2007           43.3
```

1.3 Valores perdidos

```
library(VIM)
```

```
## Warning: package 'VIM' was built under R version 4.4.2
## Loading required package: colorspace
## Warning: package 'colorspace' was built under R version 4.4.2
## Loading required package: grid
## VIM is ready to use.
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
## The following object is masked from 'package:datasets':
##
##     sleep
```

```r
missing_values <- colSums(is.na(df))
print(missing_values)
```

```
##                     Country                      Region
##                           0                           0
##                        Year               Infant_deaths
##                           0                           0
##           Under_five_deaths             Adult_mortality
##                           0                           0
##         Alcohol_consumption                Hepatitis_B
##                           0                           0
##                    Measles                         BMI
##                           0                           0
##                      Polio                 Diphtheria
##                           0                           0
##               Incidents_HIV             GDP_per_capita
##                           0                           0
##             Population_mln Thinness_ten_nineteen_years
##                           0                           0
##    Thinness_five_nine_years                   Schooling
##                           0                           0
##    Economy_status_Developed   Economy_status_Developing
##                           0                           0
##             Life_expectancy                    Outliers
##                           0                           0
```

Como se puede obervar, no existen valores nulos, por lo que se continua, con el análisis. Y se filtran los valores por el año 2015.

1.4 Subconjunto de datos. Análisis del año 2015

```r
df_2015 <- df %>%
  filter(Year == 2015)
head(df_2015, 5)
```

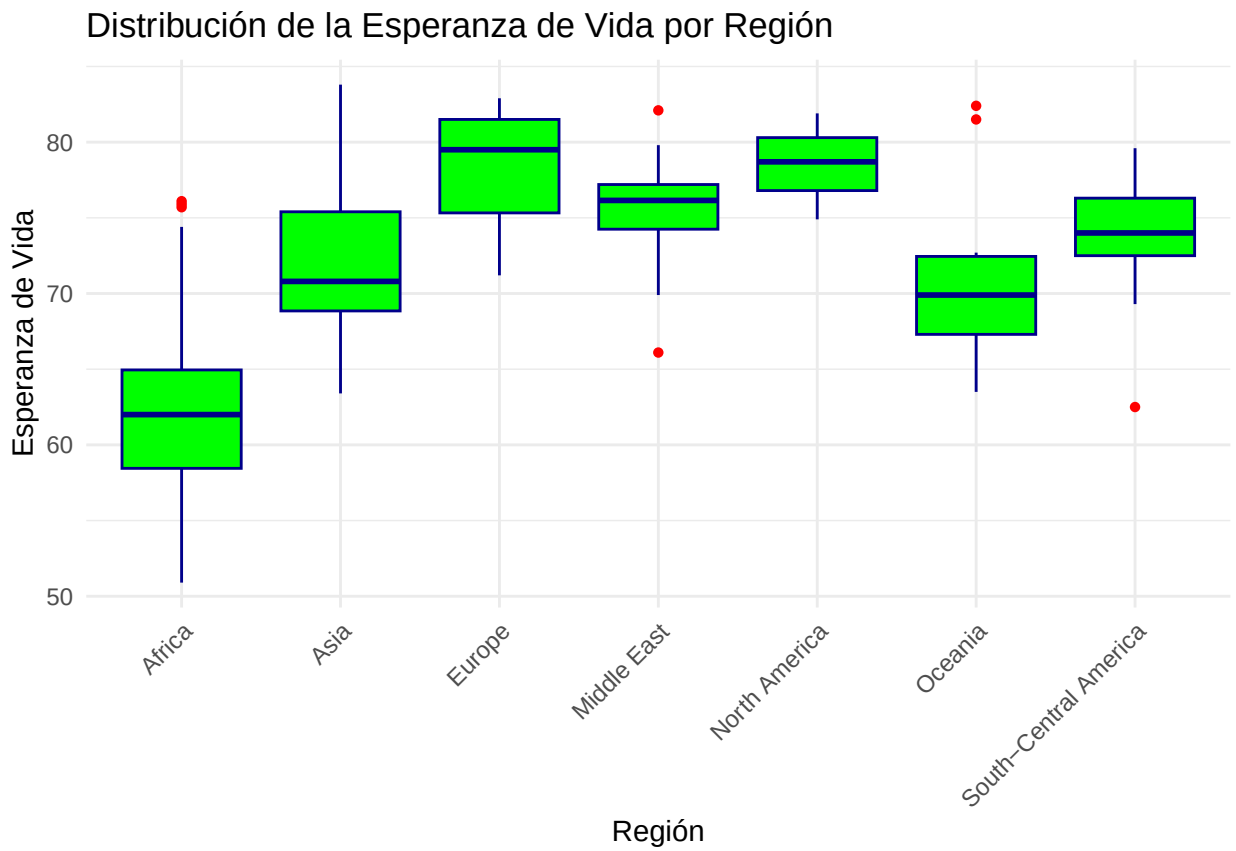
```
##              Country      Region Year Infant_deaths Under_five_deaths
## 1            Turkiye Middle East 2015          11.1              13.0
## 2              Spain      Europe 2015           2.7               3.3
## 3 Russian Federation      Europe 2015           6.6               8.2
## 4           Cameroon      Africa 2015          57.0              88.0
## 5        Gambia, The      Africa 2015          39.7              59.8
##   Adult_mortality Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria
## 1        105.8240                1.32          97      65 27.8    97         97
## 2         57.9025               10.35          97      94 26.0    97         97
## 3        223.0000                8.06          97      97 26.2    97         97
## 4        340.1265                4.55          84      64 24.3    77         84
## 5        261.7065                2.69          97      64 23.9    96         97
##   Incidents_HIV GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          0.08          11006          78.53                         4.9
## 2          0.09          25742          46.44                         0.6
## 3          0.08           9313         144.10                         2.3
## 4          1.12           1383          23.30                         5.6
## 5          0.96            661           2.09                         7.3
##   Thinness_five_nine_years Schooling Economy_status_Developed
```

```
## 1                      4.8       7.8                      0
## 2                      0.5       9.7                      1
## 3                      2.3      12.0                      0
## 4                      5.5       6.1                      0
## 5                      7.2       3.4                      0
##   Economy_status_Developing Life_expectancy Outliers
## 1                         1            76.5    FALSE
## 2                         0            82.8    FALSE
## 3                         1            71.2    FALSE
## 4                         1            57.6    FALSE
## 5                         1            60.9    FALSE
```

2 Estadística Descriptiva

2.1 Esperanza de vida por región

```
ggplot(df_2015, aes(x = Region, y = Life_expectancy)) +
  geom_boxplot(fill = "green", color = "darkblue", outlier.color = "red",
               outlier.shape = 16) +
  theme_minimal() +
  labs(
    title = "Distribución de la Esperanza de Vida por Región",
    x = "Región",
    y = "Esperanza de Vida"
  ) +
  theme(
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```



La esperanza de vida en África tiene la menor mediana de todas las regiones, valor que era esperado, además la caja es pequeña lo que indica poca variabilidad en las esperanzas de vida, la mediana mayor se encuentra en Europa, sin ningún valor atípico con su caja relativamente pequeña también lo que de nuevo indica poca variabilidad (en los cuartiles) en sus esperanzas de vida. Adicionalmente se destaca Oceania que con bigotes tan pequeños muestra que la mayoria de los datos están concentrados cerca de la mediana

2.2 Correlación

```
num_variables <- df_2015 %>%
    select(where(is.numeric))
```

```
corr_matrix <- cor(num_variables)
```

```
## Warning in cor(num_variables): the standard deviation is zero
```

```
corr_matrix <- round(corr_matrix, 3)
head(corr_matrix,5)
```

```
##                     Year Infant_deaths Under_five_deaths Adult_mortality
## Year                   1           NA                NA              NA
## Infant_deaths         NA        1.000             0.990           0.854
## Under_five_deaths     NA        0.990             1.000           0.842
## Adult_mortality       NA        0.854             0.842           1.000
## Alcohol_consumption   NA       -0.469            -0.439          -0.291
##                     Alcohol_consumption Hepatitis_B Measles    BMI  Polio
## Year                                 NA          NA      NA     NA     NA
## Infant_deaths                    -0.469      -0.506  -0.586 -0.593 -0.652
## Under_five_deaths                -0.439      -0.529  -0.585 -0.604 -0.665
```

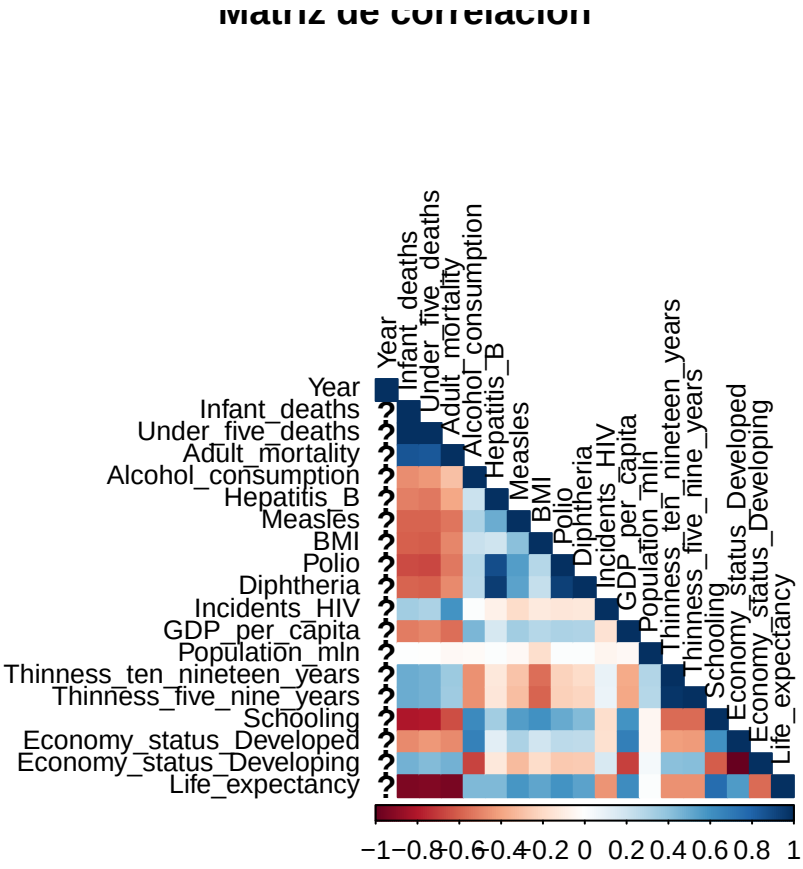
```
## Adult_mortality                 -0.291      -0.390  -0.533 -0.490 -0.525
## Alcohol_consumption              1.000       0.210   0.314  0.226  0.288
##                      Diphtheria Incidents_HIV GDP_per_capita Population_mln
## Year                         NA           NA             NA             NA
## Infant_deaths            -0.581        0.341         -0.514          0.000
## Under_five_deaths        -0.595        0.312         -0.481         -0.006
## Adult_mortality          -0.476        0.591         -0.558         -0.040
## Alcohol_consumption       0.276        0.018          0.457         -0.021
##                      Thinness_ten_nineteen_years Thinness_five_nine_years
## Year                                          NA                       NA
## Infant_deaths                              0.496                    0.495
## Under_five_deaths                          0.479                    0.472
## Adult_mortality                            0.360                    0.357
## Alcohol_consumption                       -0.456                   -0.459
##                      Schooling Economy_status_Developed
## Year                        NA                       NA
## Infant_deaths           -0.802                   -0.476
## Under_five_deaths       -0.796                   -0.435
## Adult_mortality         -0.647                   -0.477
## Alcohol_consumption      0.633                    0.679
##                      Economy_status_Developing Life_expectancy
## Year                                        NA             NA
## Infant_deaths                            0.476         -0.931
## Under_five_deaths                        0.435         -0.921
## Adult_mortality                          0.477         -0.949
## Alcohol_consumption                     -0.679          0.449
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.4.2
```

```
## corrplot 0.95 loaded
```

```
corrplot(corr_matrix, method = 'color', type = 'lower',
         tl.col = 'black', tl.cex = 0.8, title = "Matriz de correlación")
```



Se elmimina la columan Year ya que al tener valores constantes nos da el error de que la desviación standar es 0

```
df_2015_filtered <- sapply(df_2015, function(x) is.numeric(x) &&
                           sd(x, na.rm = TRUE) == 0)
```

```
df_2015 <- df[, !df_2015_filtered]
head(df_2015,5)
```

```
##   Country               Region Infant_deaths Under_five_deaths Adult_mortality
## 1 Turkiye           Middle East          11.1              13.0        105.8240
## 2   Spain                Europe           2.7               3.3         57.9025
## 3   India                  Asia          51.5              67.9        201.0765
## 4  Guyana South-Central America          32.8              40.5        222.1965
## 5  Israel           Middle East           3.4               4.3         57.9510
##   Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria Incidents_HIV
## 1                1.32          97      65 27.8    97         97          0.08
## 2               10.35          97      94 26.0    97         97          0.09
## 3                1.57          60      35 21.2    67         64          0.13
## 4                5.68          93      74 25.3    92         93          0.79
## 5                2.89          97      89 27.0    94         94          0.08
##   GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          11006          78.53                         4.9
## 2          25742          46.44                         0.6
## 3           1076        1183.21                        27.1
## 4           4146           0.75                         5.7
## 5          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
```

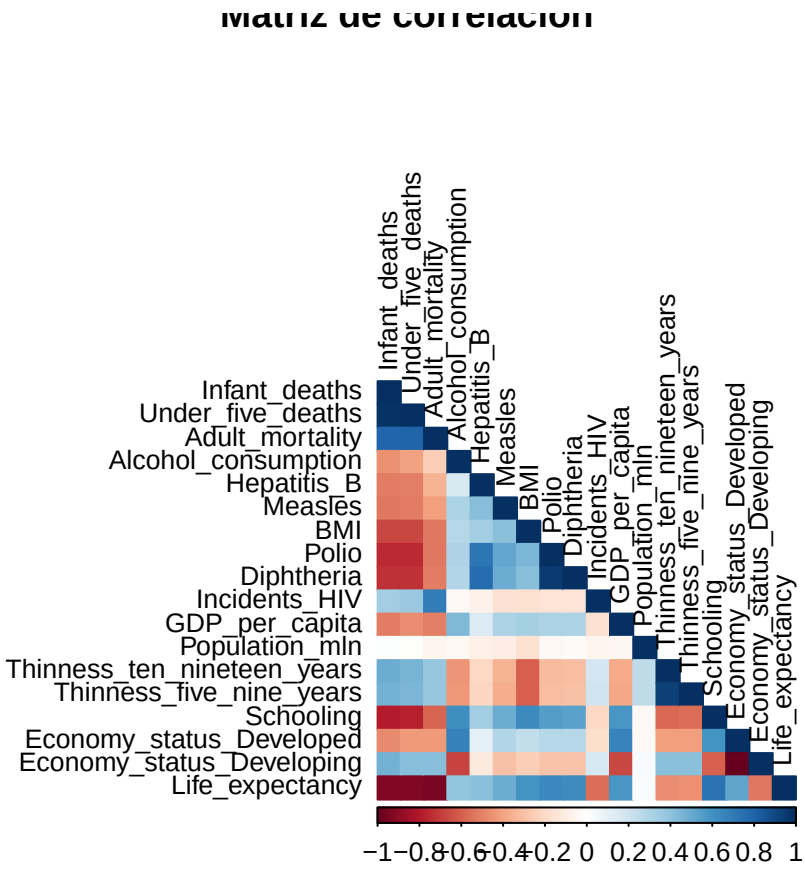
```
## 1                     4.8       7.8                     0
## 2                     0.5       9.7                     1
## 3                    28.0       5.0                     0
## 4                     5.5       7.9                     0
## 5                     1.1      12.8                     1
##   Economy_status_Developing Life_expectancy Outliers
## 1                         1            76.5    FALSE
## 2                         0            82.8    FALSE
## 3                         1            65.4    FALSE
## 4                         1            67.0    FALSE
## 5                         0            81.7    FALSE
```

```
num_variables <- df_2015 %>%
    select(where(is.numeric))
corr_matrix <- cor(num_variables)
corr_matrix <- round(corr_matrix, 3)
```

```
head(corr_matrix,5)
```

```
##                     Infant_deaths Under_five_deaths Adult_mortality
## Infant_deaths               1.000             0.986           0.795
## Under_five_deaths           0.986             1.000           0.802
## Adult_mortality             0.795             0.802           1.000
## Alcohol_consumption        -0.455            -0.409          -0.245
## Hepatitis_B                -0.513            -0.507          -0.345
##                     Alcohol_consumption Hepatitis_B Measles    BMI  Polio
## Infant_deaths                    -0.455      -0.513  -0.526 -0.662 -0.741
## Under_five_deaths                -0.409      -0.507  -0.513 -0.665 -0.743
## Adult_mortality                  -0.245      -0.345  -0.416 -0.523 -0.524
## Alcohol_consumption               1.000       0.168   0.319  0.284  0.302
## Hepatitis_B                       0.168       1.000   0.429  0.345  0.724
##                     Diphtheria Incidents_HIV GDP_per_capita Population_mln
## Infant_deaths           -0.722         0.349         -0.512          0.008
## Under_five_deaths       -0.725         0.370         -0.470         -0.005
## Adult_mortality         -0.514         0.699         -0.510         -0.054
## Alcohol_consumption      0.299        -0.034          0.444         -0.039
## Hepatitis_B              0.762        -0.076          0.159         -0.082
##                     Thinness_ten_nineteen_years Thinness_five_nine_years
## Infant_deaths                             0.491                    0.478
## Under_five_deaths                         0.467                    0.451
## Adult_mortality                           0.382                    0.380
## Alcohol_consumption                      -0.446                   -0.433
## Hepatitis_B                              -0.208                   -0.214
##                     Schooling Economy_status_Developed
## Infant_deaths          -0.789                  -0.476
## Under_five_deaths      -0.773                  -0.427
## Adult_mortality        -0.581                  -0.429
## Alcohol_consumption     0.616                   0.670
## Hepatitis_B             0.348                   0.114
##                     Economy_status_Developing Life_expectancy
## Infant_deaths                           0.476          -0.920
## Under_five_deaths                       0.427          -0.920
## Adult_mortality                         0.429          -0.945
## Alcohol_consumption                    -0.670           0.399
## Hepatitis_B                            -0.114           0.418
```

```
corrplot(corr_matrix, method = 'color', type = 'lower',
         tl.col = 'black', tl.cex = 0.8, title = "Matriz de correlación")
```



2.3 Coordenadas paralelas

```
mean_values_per_region <- df_2015 %>%
  group_by(Region) %>%
  summarise(
    Thinnes_five_nine_years = mean(Thinness_five_nine_years, na.rm = TRUE),
    Schooling = mean(Schooling, na.rm = TRUE),
    Alcohol_consumption = mean(Alcohol_consumption, na.rm = TRUE),
    Life_expectancy = mean(Life_expectancy, na.rm = TRUE)
  )
```

```
head(mean_values_per_region,5)
```

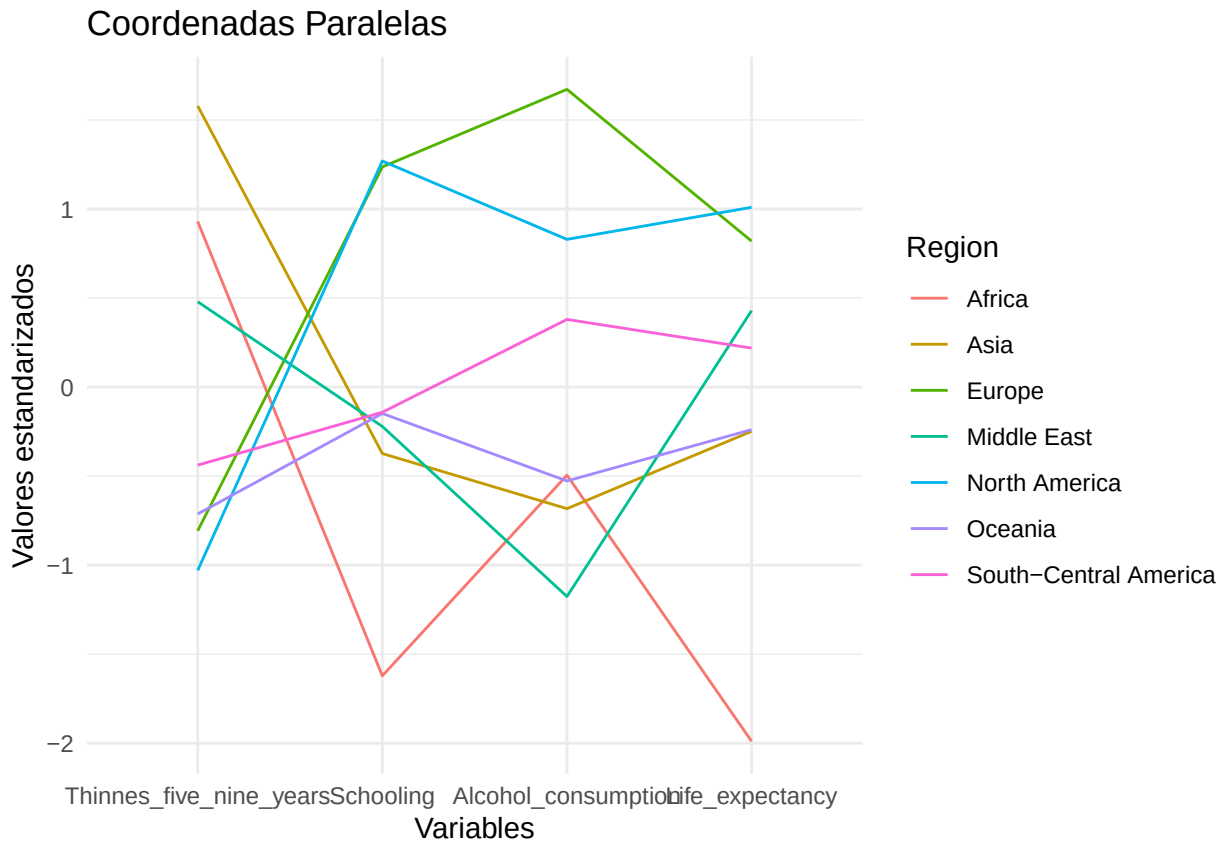
```
## # A tibble: 5 x 5
##   Region    Thinnes_five_nine_ye~1 Schooling Alcohol_consumption Life_expectancy
##   <chr>                      <dbl>     <dbl>               <dbl>           <dbl>
## 1 Africa                      7.20      4.59               2.98             57.8
## 2 Asia                        9.28      7.38               2.40             69.5
## 3 Europe                      1.62     11.0                9.64             76.6
## 4 Middle E~                   5.75      7.71               0.884            74.0
## 5 North Am~                   0.904    11.0                7.05             77.8
## # i abbreviated name: 1: Thinnes_five_nine_years
```

```
library(GGally)
```

```
## Warning: package 'GGally' was built under R version 4.4.2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
ggparcoord(
  data = mean_values_per_region,
  columns = 2:5,
  groupColumn = 1,
  scale = "std",
  title = "Coordenadas Paralelas",
) +
  theme_minimal() +
  labs(
    x = "Variables",
    y = "Valores estandarizados",
    color = "Region"
  )
```



3 Estadística Inferencial

¿Podemos aceptar que la esperanza de vida media de los países en África es inferior a la del resto de países? Considerad un nivel de confianza del 99%

3.1 Hipótesis del contraste

3.1.1 Hipótesis Nula (H_0)

La esperanza de vida media de los países en África es igual a la del resto de países

$$H_0 : \mu_{\text{Á}frica} = \mu_{Resto\_del\_mundo}$$

3.1.2 Hipótesis Alternativa (H_1)

La esperanza de vida media de los países Africanos no es igual a la del resto de países.

$$H_1 : \mu_{\text{Á}frica} \neq \mu_{Resto\_del\_mundo}$$

3.2 Test

Para saber con seguridad que test usar primero se realizará un test de Levene. Se hará uso de la librería de R car. Como se mencionó en la entrega 2, este test se usa para saber la igualdad entre 2 o más grupos. Se usará para saber si se puede usar un t-student, ya que asume varianzas poblacionales de las muestras iguales. El test de Levene se define según:

$$L = \frac{N-k}{k-1} \frac{\sum_{i=1}^{k} N_i (Z_i - Z..)^2}{\sum_{i=1}^{k} \sum_{j=1}^{N_i} (Z_{ij} - Zi.)^2}$$

donde:

- L es el resultado de la prueba
- k es el número de diferentes grupos a los que pertencen los valores
- N es el número total de casos en todos los grupos
- N_i es el número de casos en el grupo i
- Y_{ij} es el valor de la variable medida para el jésimo caso del iésimo grupo
- Y_i^- es la media o la mediana del iésimo grupo
- Z.. $= \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{N_i} Z_{ij}$ es la medida de Z_{ij}
- $Z_i. = \frac{1}{N_i} \sum_{j=1}^{N_i} Z_{ij}$ es la media de Z_{ij} para el grupo i.

```
library(car)
```

```
## Warning: package 'car' was built under R version 4.4.2
```

```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.4.2
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
levene_test <- leveneTest(Life_expectancy ~ Region, data = df_2015)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.
```

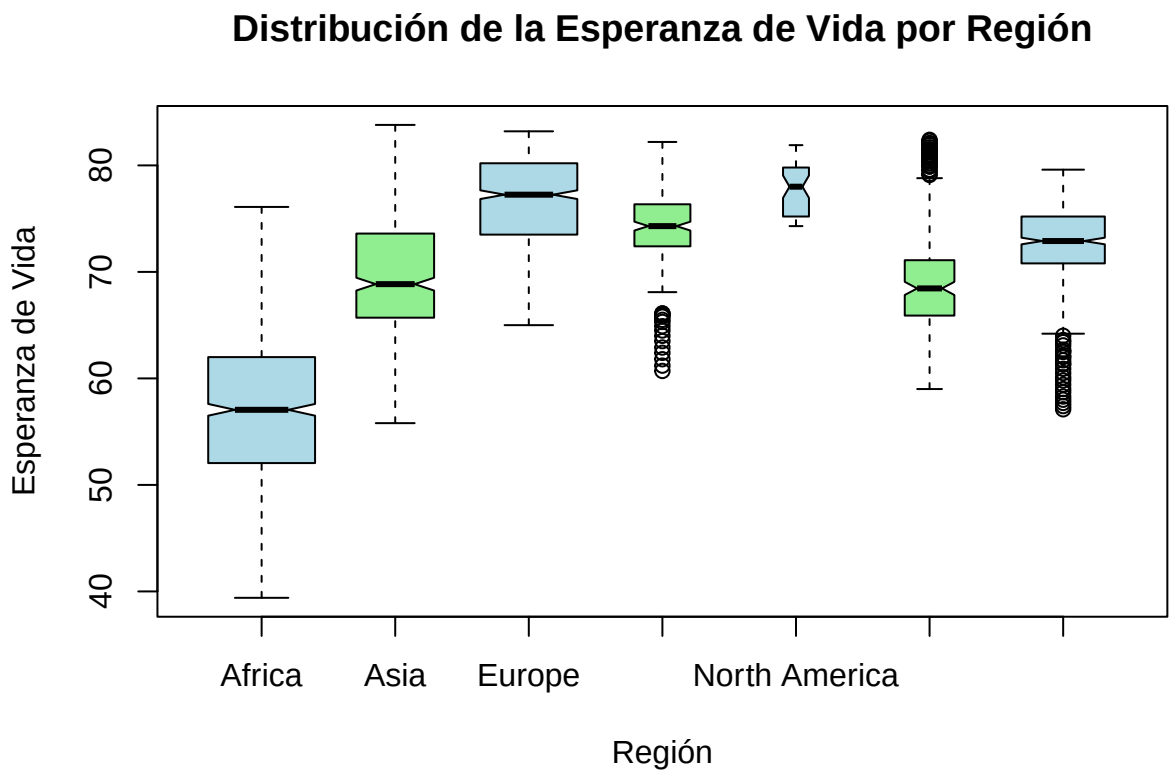
```
print(levene_test)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    6  68.318 < 2.2e-16 ***
##       2857
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

El valor de p es de 2.2e-16, que es mucho menor que $\alpha = 0.01$, ya que el intervalo de confianza es del 99%, por ello se rechaza la hipótesis nula a favor de la alternativa. Las varianzas no son iguales. Entonces se va a seguir con el test t-student ajustado para varinzas desiguales. De todas formas se visualiza.

```
boxplot(Life_expectancy ~ Region, data = df_2015,
        main = "Distribución de la Esperanza de Vida por Región",
        xlab = "Región", ylab = "Esperanza de Vida",
        col = c("lightblue", "lightgreen"),
        border = "black", notch = TRUE,
        varwidth = TRUE)
```



Gracias al diagraba de cajas podemos observar que efectivamente la varianza, difiere entre África y el resto del mundo, ya que tanto las cajas como los "bigotes" son de diferente tamaño. Esto indica que la dispersión es difernte por lo que las varianzas también. Continuamos con el t-student.

```
Africa <- df_2015 %>% filter(Region == "Africa") %>% pull(Life_expectancy)
non_Africa <- df_2015 %>% filter(Region != "Africa") %>% pull(Life_expectancy)
```

```
t_student <- t.test(Africa, non_Africa, alternative = "two.sided",
                    var.equal = FALSE, conf.level = 0.99)
print(t_student)
```

```
##
##  Welch Two Sample t-test
##
## data:  Africa and non_Africa
```

```
## t = -49.765, df = 1119.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 99 percent confidence interval:
##  -16.19329 -14.59686
## sample estimates:
## mean of x mean of y
##  57.84730  73.24238
```

p = 2.2e-16 < 0.01, por lo tanto se rechaza la hipótesis nula, a favor de la hipótesis alternativa, la espereanza de vida media de los países Africanos no es igual a la del resto de países. Algo que antes de realizar el test ya se sabía, ya que hay que tener en cuenta la calidad de vida en estos países con respecto al resto del mundo. Un planteamiento más directo hubiera sido plantear la hipótesis alternativa como

$$H_1 : \mu_{\acute{A}frica} < \mu_{Resto\_del\_mundo}$$

Lo que igualmente, acabaría fallando a favor de la hipótesis nula.

4 Modelo de regresión Lineal

4.1 Modelo de regresión lineal múltiple

Estimad un modelo de regresión lineal múltiple considerando todas las variables (excepto Country). Interpretad el modelo lineal ajustado y valorad la calidad del ajuste.

```
model_rl <- lm(Life_expectancy ~ . - Country, data = df_2015)
summary(model_rl)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ . - Country, data = df_2015)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3169 -0.8434  0.0131  0.7651  6.2682
##
## Coefficients: (1 not defined because of singularities)
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  8.373e+01  5.816e-01 143.974  < 2e-16 ***
## RegionAsia                   2.161e-01  1.018e-01   2.123  0.03382 *
## RegionEurope                -8.563e-02  1.222e-01  -0.701  0.48352
## RegionMiddle East            7.623e-02  1.243e-01   0.613  0.53963
## RegionNorth America          8.104e-01  2.135e-01   3.796  0.00015 ***
## RegionOceania               -8.544e-01  1.305e-01  -6.546 6.97e-11 ***
## RegionSouth-Central America  1.728e+00  1.003e-01  17.221  < 2e-16 ***
## Infant_deaths               -5.285e-02  5.729e-03  -9.225  < 2e-16 ***
## Under_five_deaths           -5.138e-02  3.611e-03 -14.228  < 2e-16 ***
## Adult_mortality             -4.689e-02  5.622e-04 -83.404  < 2e-16 ***
## Alcohol_consumption         -1.544e-02  1.043e-02  -1.481  0.13876
## Hepatitis_B                 -6.578e-03  2.341e-03  -2.810  0.00499 **
## Measles                      2.323e-03  1.558e-03   1.490  0.13622
## BMI                         -1.417e-01  2.011e-02  -7.049 2.26e-12 ***
## Polio                        8.091e-03  5.246e-03   1.542  0.12311
## Diphtheria                  -8.566e-03  5.313e-03  -1.612  0.10706
## Incidents_HIV                8.584e-02  1.692e-02   5.072 4.19e-07 ***
## GDP_per_capita               2.354e-05  2.106e-06  11.180  < 2e-16 ***
## Population_mln              -1.841e-04  1.829e-04  -1.007  0.31418
```

```
## Thinness_ten_nineteen_years -3.523e-02  1.533e-02  -2.298  0.02163 *
## Thinness_five_nine_years     2.608e-02  1.521e-02   1.715  0.08643 .
## Schooling                    1.354e-01  1.636e-02   8.276  < 2e-16 ***
## Economy_status_Developed     1.814e+00  1.130e-01  16.059  < 2e-16 ***
## Economy_status_Developing           NA         NA      NA       NA
## OutliersTRUE                 2.619e-01  3.073e-01   0.852  0.39403
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.214 on 2840 degrees of freedom
## Multiple R-squared:  0.9835, Adjusted R-squared:  0.9833
## F-statistic:  7345 on 23 and 2840 DF,  p-value: < 2.2e-16
```

El valor de R2 = 0.9835 nos dice, que las variables del modelo pueden explicar el 98.35% de la variabilidad de Life Expectancy. Las columnas Infant_deaths, Under_five_deaths, Adult_mortality, BMI, etc.. muestran un valores muy bajos, por debajo de 0.01, que es nuestro alpha, lo que indican con un 99% de confianza que estas variables tienen un efecto importante sobre la esperanza de vida. Incidents HIV, es la más importante en este caso. También es importante fijasrse en el valor residual 1.214, ya que al ser este bajo indicaría que las predicciones serían precisas.

4.2 Análisis de multicolinealidad y elección del modelo final

Analizad posibles problemas de multicolinealidad (alta correlación entre variables explicativas) mediante la interpretación del factor de inflación de la varianza (vif).

Para poder realizar este apartado es necesario, quitar las columnas, que esten altamente relacionadas entre ellas, debido a que si no, nos saldrá un error de coeficientes 'aliased' que singifica que hay coeficientes que son linealmente dependientes. Se va a repetir el análsis varias veces hasta terminar con columnas con valores de GVIF menores de 10, que indican una fuerte colinealidad.

```r
highly_correlated <- which(abs(corr_matrix) > 0.8, arr.ind = TRUE)
highly_correlated <- highly_correlated[highly_correlated[,1]
                                       != highly_correlated[,2], ]
highly_correlated_vars <- apply(highly_correlated, 1, function(x) {
  rowname <- rownames(corr_matrix)[x[1]]
  colname <- colnames(corr_matrix)[x[2]]
  c(rowname, colname)
})

print(highly_correlated_vars)
```

```
##      Under_five_deaths   Life_expectancy   Infant_deaths
## [1,] "Under_five_deaths" "Life_expectancy" "Infant_deaths"
## [2,] "Infant_deaths"     "Infant_deaths"   "Under_five_deaths"
##      Adult_mortality     Life_expectancy     Under_five_deaths
## [1,] "Adult_mortality"   "Life_expectancy"   "Under_five_deaths"
## [2,] "Under_five_deaths" "Under_five_deaths" "Adult_mortality"
##      Life_expectancy   Diphtheria   Polio        Thinness_five_nine_years
## [1,] "Life_expectancy" "Diphtheria" "Polio"      "Thinness_five_nine_years"
## [2,] "Adult_mortality" "Polio"      "Diphtheria" "Thinness_ten_nineteen_years"
##      Thinness_ten_nineteen_years   Economy_status_Developing
## [1,] "Thinness_ten_nineteen_years" "Economy_status_Developing"
## [2,] "Thinness_five_nine_years"    "Economy_status_Developed"
##      Economy_status_Developed    Infant_deaths     Under_five_deaths
## [1,] "Economy_status_Developed"  "Infant_deaths"   "Under_five_deaths"
## [2,] "Economy_status_Developing" "Life_expectancy" "Life_expectancy"
```

```
##      Adult_mortality
## [1,] "Adult_mortality"
## [2,] "Life_expectancy"
```

```
df_reduced <- df_2015[, !names(df_2015) %in% c("Under_five_deaths",
                                               "Infant_deaths",
                                      "Adult_mortality",
                                      "Thinness_ten_nineteen_years",
                                      "Economy_status_Developing",
                                      "Outliers", "Hepatitis_B", "Country")]
```

```
model_updated <- lm(Life_expectancy ~ ., data = df_reduced)
```

```
summary(model_updated)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ ., data = df_reduced)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.4838  -2.4732   0.0868   2.4060   9.8182
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  3.001e+01  1.249e+00  24.025  < 2e-16 ***
## RegionAsia                   3.743e+00  2.747e-01  13.628  < 2e-16 ***
## RegionEurope                 4.029e+00  3.354e-01  12.014  < 2e-16 ***
## RegionMiddle East            4.274e+00  3.456e-01  12.368  < 2e-16 ***
## RegionNorth America          3.176e+00  5.988e-01   5.303 1.23e-07 ***
## RegionOceania                2.431e+00  3.615e-01   6.726 2.10e-11 ***
## RegionSouth-Central America  6.213e+00  2.687e-01  23.126  < 2e-16 ***
## Alcohol_consumption         -2.144e-01  2.948e-02  -7.273 4.52e-13 ***
## Measles                      1.648e-02  4.428e-03   3.721 0.000202 ***
## BMI                          6.479e-01  5.541e-02  11.692  < 2e-16 ***
## Polio                        1.144e-01  1.488e-02   7.691 2.00e-14 ***
## Diphtheria                   4.550e-02  1.408e-02   3.232 0.001245 **
## Incidents_HIV               -1.195e+00  3.283e-02 -36.415  < 2e-16 ***
## GDP_per_capita               9.701e-05  5.742e-06  16.894  < 2e-16 ***
## Population_mln               2.921e-03  5.231e-04   5.584 2.57e-08 ***
## Thinness_five_nine_years     8.629e-02  2.202e-02   3.919 9.12e-05 ***
## Schooling                    5.343e-01  4.386e-02  12.182  < 2e-16 ***
## Economy_status_Developed     3.795e+00  3.195e-01  11.879  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.513 on 2846 degrees of freedom
## Multiple R-squared:  0.8613, Adjusted R-squared:  0.8605
## F-statistic:  1040 on 17 and 2846 DF,  p-value: < 2.2e-16
```

```
vif_values <- vif(model_updated)
print(vif_values)
```

```
##                              GVIF Df GVIF^(1/(2*Df))
## Region                  17.467373  6        1.269168
## Alcohol_consumption      3.196904  1        1.787989
```

```
## Measles                   1.583357  1        1.258315
## BMI                       3.428294  1        1.851565
## Polio                    11.674373  1        3.416778
## Diphtheria               11.095152  1        3.330939
## Incidents_HIV             1.417619  1        1.190638
## GDP_per_capita            2.193571  1        1.481071
## Population_mln            1.182604  1        1.087476
## Thinness_five_nine_years  2.303333  1        1.517674
## Schooling                 4.488609  1        2.118634
## Economy_status_Developed  3.884802  1        1.970990
```

```
df_reduced <- df_reduced[, !colnames(df_reduced) %in% c('Region', 'Polio',
                                                        'Diphtheria')]
```

```
model_updated <- lm(Life_expectancy ~ ., data = df_reduced)
```

```
summary(model_updated)
```

```
##
## Call:
## lm(formula = Life_expectancy ~ ., data = df_reduced)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.5133  -3.0308  -0.0248   3.5556  13.7501
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               2.894e+01  1.355e+00  21.360  < 2e-16 ***
## Alcohol_consumption      -1.018e-01  3.105e-02  -3.279  0.00105 **
## Measles                   5.900e-02  5.183e-03  11.384  < 2e-16 ***
## BMI                       1.075e+00  5.626e-02  19.106  < 2e-16 ***
## Incidents_HIV            -1.544e+00  3.640e-02 -42.427  < 2e-16 ***
## GDP_per_capita            1.029e-04  6.929e-06  14.848  < 2e-16 ***
## Population_mln            4.001e-03  6.373e-04   6.278 3.94e-10 ***
## Thinness_five_nine_years  1.563e-01  2.547e-02   6.135 9.72e-10 ***
## Schooling                 1.019e+00  4.725e-02  21.561  < 2e-16 ***
## Economy_status_Developed  2.130e+00  3.411e-01   6.244 4.89e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.406 on 2854 degrees of freedom
## Multiple R-squared:  0.7812, Adjusted R-squared:  0.7805
## F-statistic:  1132 on 9 and 2854 DF,  p-value: < 2.2e-16
```

```
vif_values <- vif(model_updated)
print(vif_values)
```

```
##      Alcohol_consumption                 Measles                     BMI
##                 2.254277                1.379139                2.246503
##            Incidents_HIV          GDP_per_capita           Population_mln
##                 1.107918                2.029952                1.115465
## Thinness_five_nine_years            Schooling Economy_status_Developed
##                 1.958910                3.311273                2.813343
```

Terminado así el análisis, las variables seleccionadas son: Alcohol_consumption Measles, BMI, Indicents_HIV,

GDP_per_capita, Population_mln, Thinnes_five_nine_years, Schooling, Economy_status_Developed.

5 Regresión Logística

5.1 Modelo Predictivo

Para continuar con los ejercicios se vuelve al data frame original con los valores filtrados para 2015.

```
head(df_2015,5)
```

```
##   Country               Region Infant_deaths Under_five_deaths Adult_mortality
## 1 Turkiye           Middle East          11.1              13.0        105.8240
## 2   Spain                Europe           2.7               3.3         57.9025
## 3   India                  Asia          51.5              67.9        201.0765
## 4  Guyana South-Central America          32.8              40.5        222.1965
## 5  Israel           Middle East           3.4               4.3         57.9510
##   Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria Incidents_HIV
## 1                1.32          97      65 27.8    97         97          0.08
## 2               10.35          97      94 26.0    97         97          0.09
## 3                1.57          60      35 21.2    67         64          0.13
## 4                5.68          93      74 25.3    92         93          0.79
## 5                2.89          97      89 27.0    94         94          0.08
##   GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          11006          78.53                         4.9
## 2          25742          46.44                         0.6
## 3           1076        1183.21                        27.1
## 4           4146           0.75                         5.7
## 5          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
## 1                      4.8       7.8                       0
## 2                      0.5       9.7                       1
## 3                     28.0       5.0                       0
## 4                      5.5       7.9                       0
## 5                      1.1      12.8                       1
##   Economy_status_Developing Life_expectancy Outliers
## 1                         1           76.5    FALSE
## 2                         0           82.8    FALSE
## 3                         1           65.4    FALSE
## 4                         1           67.0    FALSE
## 5                         0           81.7    FALSE
```

```
median_adult_mortaility <- median(df_2015$Adult_mortality)

df_2015$High_adult_mortality <- ifelse(df_2015$Adult_mortality >
                                       median_adult_mortaility, 1,0)
head(df_2015,5)
```

```
##   Country               Region Infant_deaths Under_five_deaths Adult_mortality
## 1 Turkiye           Middle East          11.1              13.0        105.8240
## 2   Spain                Europe           2.7               3.3         57.9025
## 3   India                  Asia          51.5              67.9        201.0765
## 4  Guyana South-Central America          32.8              40.5        222.1965
## 5  Israel           Middle East           3.4               4.3         57.9510
##   Alcohol_consumption Hepatitis_B Measles  BMI Polio Diphtheria Incidents_HIV
## 1                1.32          97      65 27.8    97         97          0.08
```

```
## 2               10.35          97      94 26.0    97         97          0.09
## 3                1.57          60      35 21.2    67         64          0.13
## 4                5.68          93      74 25.3    92         93          0.79
## 5                2.89          97      89 27.0    94         94          0.08
##   GDP_per_capita Population_mln Thinness_ten_nineteen_years
## 1          11006          78.53                         4.9
## 2          25742          46.44                         0.6
## 3           1076        1183.21                        27.1
## 4           4146           0.75                         5.7
## 5          33995           7.91                         1.2
##   Thinness_five_nine_years Schooling Economy_status_Developed
## 1                      4.8       7.8                        0
## 2                      0.5       9.7                        1
## 3                     28.0       5.0                        0
## 4                      5.5       7.9                        0
## 5                      1.1      12.8                        1
##   Economy_status_Developing Life_expectancy Outliers High_adult_mortality
## 1                         1           76.5    FALSE                    0
## 2                         0           82.8    FALSE                    0
## 3                         1           65.4    FALSE                    1
## 4                         1           67.0    FALSE                    1
## 5                         0           81.7    FALSE                    0
```

```r
model <- glm(High_adult_mortality ~ Hepatitis_B + Measles + BMI + Polio +
               Diphtheria, data = df_2015, family = binomial)
summary(model)
```

```
##
## Call:
## glm(formula = High_adult_mortality ~ Hepatitis_B + Measles +
##     BMI + Polio + Diphtheria, family = binomial, data = df_2015)
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) 17.858771   0.807535  22.115  < 2e-16 ***
## Hepatitis_B  0.035714   0.005804   6.153 7.59e-10 ***
## Measles     -0.026907   0.003310  -8.128 4.36e-16 ***
## BMI         -0.375390   0.026514 -14.158  < 2e-16 ***
## Polio       -0.075718   0.014546  -5.206 1.93e-07 ***
## Diphtheria  -0.029505   0.014364  -2.054     0.04 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3970.3  on 2863  degrees of freedom
## Residual deviance: 2712.0  on 2858  degrees of freedom
## AIC: 2724
##
## Number of Fisher Scoring iterations: 5
```

Los coeficientes indican que:

Hepatitis_B es un factor de riesgo, ya que un aumento en su valor incrementa la probabilidad de alta mortalidad adulta. Measles, BMI, Polio y Diphtheria actúan como factores inversamente proporcionales, ya que sus aumentos se asocian con una disminución en la probabilidad de alta mortalidad. Los p-valores son

muy bajos para todas las variables, lo que indica una alta significancia estadística, y el modelo tiene un buen ajuste con una desviación residual de 2712.0

5.2 Matriz de confusión

Analizad la precisión del modelo, comparando la predicción del modelo sobre los mismos datos del conjunto de datos. Asumiremos que la predicción del modelo es 1 si la probabilidad del modelo de regresión logística es superior o igual a 0.5 y 0 en caso contrario. Calculad la matriz de confusión. Indicad los valores de sensibilidad y especificidad e interpretadlos.

```
predictions <- predict(model, type = 'response')
classify_predictions <- ifelse(predictions >= 0.5,1,0)
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.4.2
```

```
## Loading required package: lattice
```

```
conf_matrix <- confusionMatrix(as.factor(classify_predictions),
                               as.factor(df_2015$High_adult_mortality))

print(conf_matrix)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1245  447
##          1  187  985
##
##                Accuracy : 0.7786
##                  95% CI : (0.763, 0.7937)
##     No Information Rate : 0.5
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5573
##
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.8694
##             Specificity : 0.6878
##          Pos Pred Value : 0.7358
##          Neg Pred Value : 0.8404
##              Prevalence : 0.5000
##          Detection Rate : 0.4347
##    Detection Prevalence : 0.5908
##       Balanced Accuracy : 0.7786
##
##        'Positive' Class : 0
##
```

El modelo tiene una precisión del 77.86%, aunque su sensibilidad es del 86.94% lo que indica que el modelo tiene una alta capacidad para identificar países con alta mortalidad adulta. Por otro lado tiene un Specificity, del 68.78%, lo que indica que el modelo tiene una capacidad limitada para predecir países con baja mortalidad. El modelo se considera como válido y robusto, aunque tiene margen de mejora.

5.3 Predicción

Aplicad el modelo de regresión logística para predecir la probabilidad que un país tenga una mortalidad de adultos superior a la mediana considerando las siguientes características: el 87.1% de los niños de 1 año están vacunados contra la Hepatitis B, el 80.2% de los niños de 1 año están vacunados con la primera dosis contra el Sarampión, el índice de masa corporal medio en adultos es de 25.6, el 88.3% de los niños de 1 año están vacunados con la primera dosis contra la Polio, y el 87.9% de los niños de 1 año vacunados con toxoide diftérico, tétánico y tos ferina. Haced los cálculos sin usar la función predict.

```
model
```

```
##
## Call:  glm(formula = High_adult_mortality ~ Hepatitis_B + Measles +
##     BMI + Polio + Diphtheria, family = binomial, data = df_2015)
##
## Coefficients:
## (Intercept)  Hepatitis_B      Measles          BMI        Polio   Diphtheria
##    17.85877      0.03571     -0.02691     -0.37539     -0.07572     -0.02950
##
## Degrees of Freedom: 2863 Total (i.e. Null);  2858 Residual
## Null Deviance:       3970
## Residual Deviance: 2712  AIC: 2724
```

Regresión Logística

$$logit(p) = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 ... \beta_i \cdot X_i$$

donde p es la probabilidad, y los valores de beta los coeficientes del modelo, Xi, son los valores de las características del enunciado y p se calcula según:

$$p = \frac{1}{1 + exp(-logit(p))}$$

```
intercept <- 17.85877
coef_hepatitis_B <- 0.03571
coef_measles <- -0.02691
coef_BMI <- -0.37539
coef_polio <- -0.07572
coef_diphtheria <- -0.02950
```

```
hepatitis_B <- 0.871
measles <- 0.802
BMI <- 25.6
polio <- 0.883
diphtheria <- 0.879
```

```
logit <- intercept + coef_hepatitis_B * hepatitis_B +
        coef_measles * measles +
        coef_BMI * BMI +
        coef_polio * polio +
        coef_diphtheria * diphtheria
```

```
probabilidad <- 1 / (1 + exp(-logit))
print(probabilidad)
```

```
## [1] 0.9997158
```

```r
datos <- data.frame(
  Hepatitis_B = 0.871,
  Measles = 0.802,
  BMI = 25.6,
  Polio = 0.883,
  Diphtheria = 0.879
)
```

```r
probabilidad <- predict(model, datos, type = "response")
print(probabilidad)
```

```
##         1
## 0.9997158
```

Se ha podido confirmar el resultado tanto con el calculo sin usar la función predict, como con ella, aproximadamente 1.

6 ANOVA unifactorial

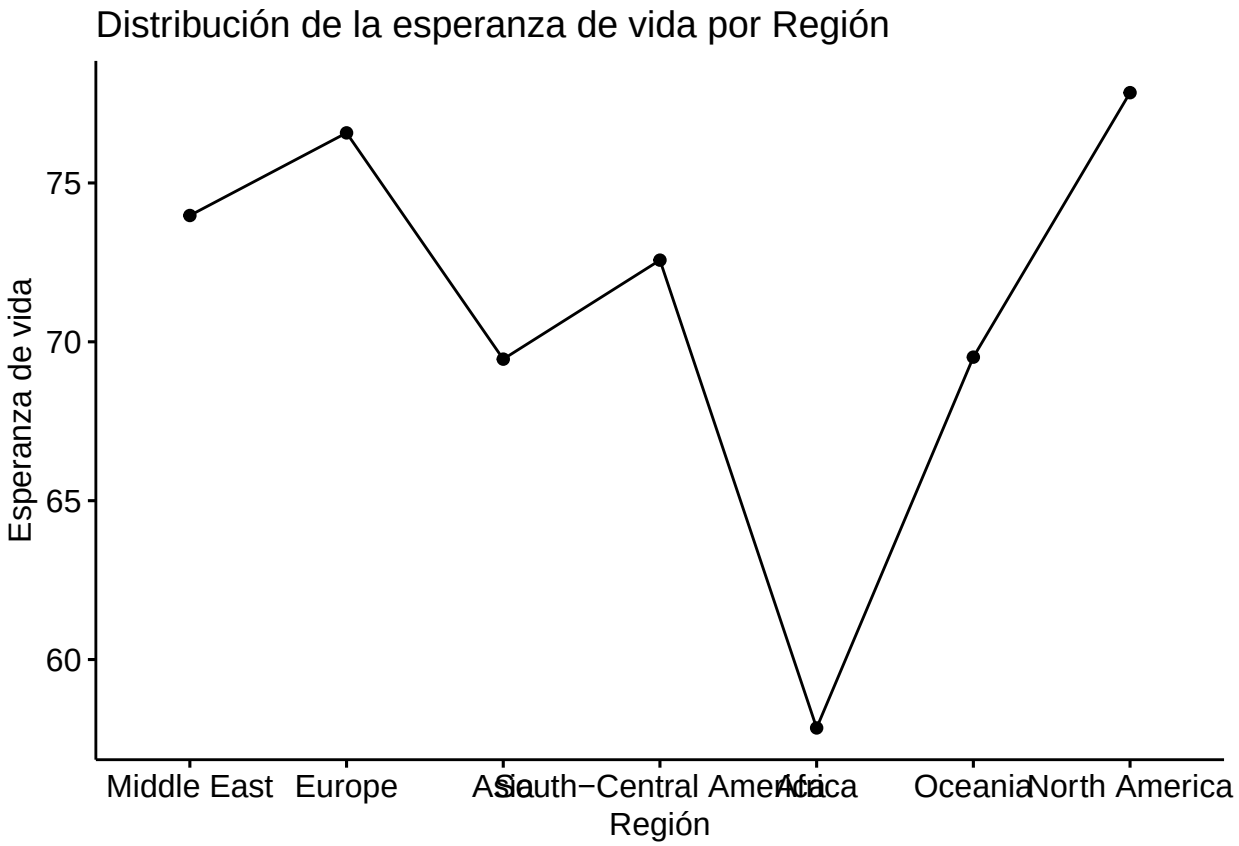
A continuación se realizará un análisis de varianza, donde se desea comparar la esperanza de vida para las distintas regiones. El análisis de varianza consiste en evaluar si la variabilidad de una variable dependiente puede explicarse a partir de una o varias variables independientes, denominadas factores. Nos interesa evaluar si la variabilidad de la variable Life_expectancy puede explicarse por la región. 1. ¿Existen diferencias en la esperanza de vida entre las diferentes regiones? 2. Si existen diferencias, ¿entre qué regiones se dan estas diferencias?

6.1 Visualización gráfica

```r
library(ggpubr)
```

```
## Warning: package 'ggpubr' was built under R version 4.4.2
```

```r
ggline(df_2015, x="Region", y = "Life_expectancy",
       add = "mean",
       xlab = "Región",
       ylab = "Esperanza de vida",
       title = "Distribución de la esperanza de vida por Región")
```



Se puede apreciar una clara variación de la esperanza de vida al nacer en las diferentes regiones del mundo, Norte América y Europa, como en el análsiis anterior contienen los valores más altos. Como también África, presenta la esperanza de vida más baja junto América del sur, mostrándo de nuevo una brecha entre los países desarrollados y en desarrollo.

6.2 Hipótesis Nula y Alternativa

Hipótesis Nula

No existen diferencias en la esperanza de vida entre los diferentes países

$$H_0 : \mu_1 = \mu_2 = \mu_i$$

donde μ_i son las medias de las esperanzas de vida por región.

Hipótesis Alternativa

Existen diferencias significativas en la esperanza de vida entre al menos dos regiones

$$H_1 : \mu_i \neq \mu_j$$

Donde μ_i y μ_j son las medias de las esperanzas de vida para al menos dos regiones.

6.3 Modelo

Calculad el análisis de varianza, usando la función aov o lm. Interpretad el resultado del análisis.

```r
anova_model <- aov(Life_expectancy ~ Region, data = df_2015)
summary(anova_model)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## Region         6 155763   25960   760.6 <2e-16 ***
## Residuals   2857  97514      34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Como era de esperar el valor de P es muy pequeño 2e-16, por lo que se falla a favor de la hipótesis alternativa, ya que el valor es menor que el intervalo de confianza del 95% y del 99%. También se tiene un valor de F alto lo que hace referencia a una alta variabilidad entre las medidas de las regiones con respecto a la variabilidad de las regiones en sí, es decir, dentro de ellas.

6.4 Efectos de los niveles del factor y fuerza de relación

Proporcionad la estimación del efecto de los niveles del factor Life_expectancy. Interpretad los resultados. Calculad la parte de la variabilidad de la esperanza de vida explicada por el efecto de los niveles (fuerza de relación), es decir, calculad $\eta^2 = SSB/SST$ del modelo. Interpretad los resultados.

```r
anova_results <- summary(anova_model)[[1]]
SSB <- anova_results[1, "Sum Sq"]
SSW <- anova_results[2, "Sum Sq"]

SST <- SSB + SSW

eta_sq <- SSB / SST

print(eta_sq)
```

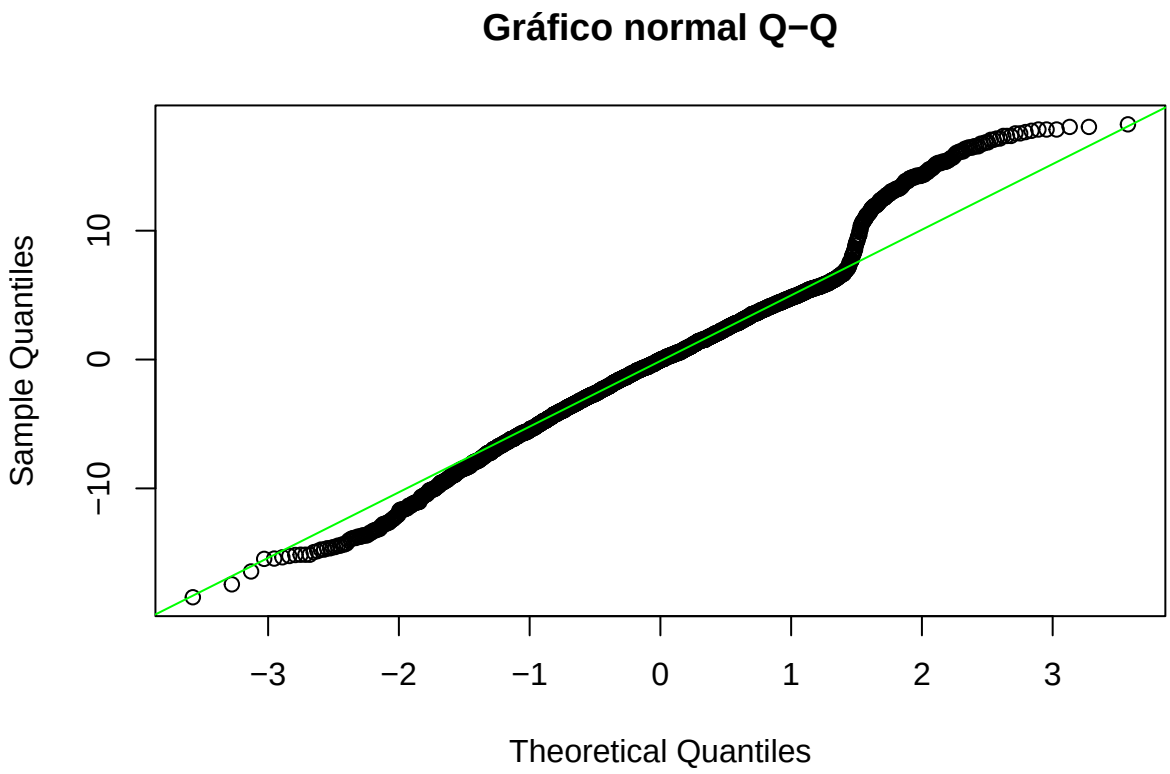
```
## [1] 0.6149901
```

El valor de eta cuadrado es aproximadamente el 61.5% lo que quiere decir que explica o que puede explicar el 61.5% de la varibilidad de la esperanza de vida por las diferencias entre las regiones. Se podría afirmar también que que el hecho de pertenecer a una región u otra explica +50% de la varibilidad observada con respecto a la esperanza de vida.

6.5 Diagnosis del modelo

Usad el gráfico Normal Q-Q y el test Shapiro-Wilk para evaluar la normalidad de los residuos. Podéis usar las funciones de R correspondientes para hacer el gráfico y el test. El gráfico "Residuals vs Fitted" proporciona información sobre la homocedasticidad de los residuos. Mostrad e interpretad este gráfico.

```r
residuals_model <- residuals(anova_model)
qqnorm(residuals_model, main = "Gráfico normal Q-Q")
qqline(residuals_model, col = "green")
```

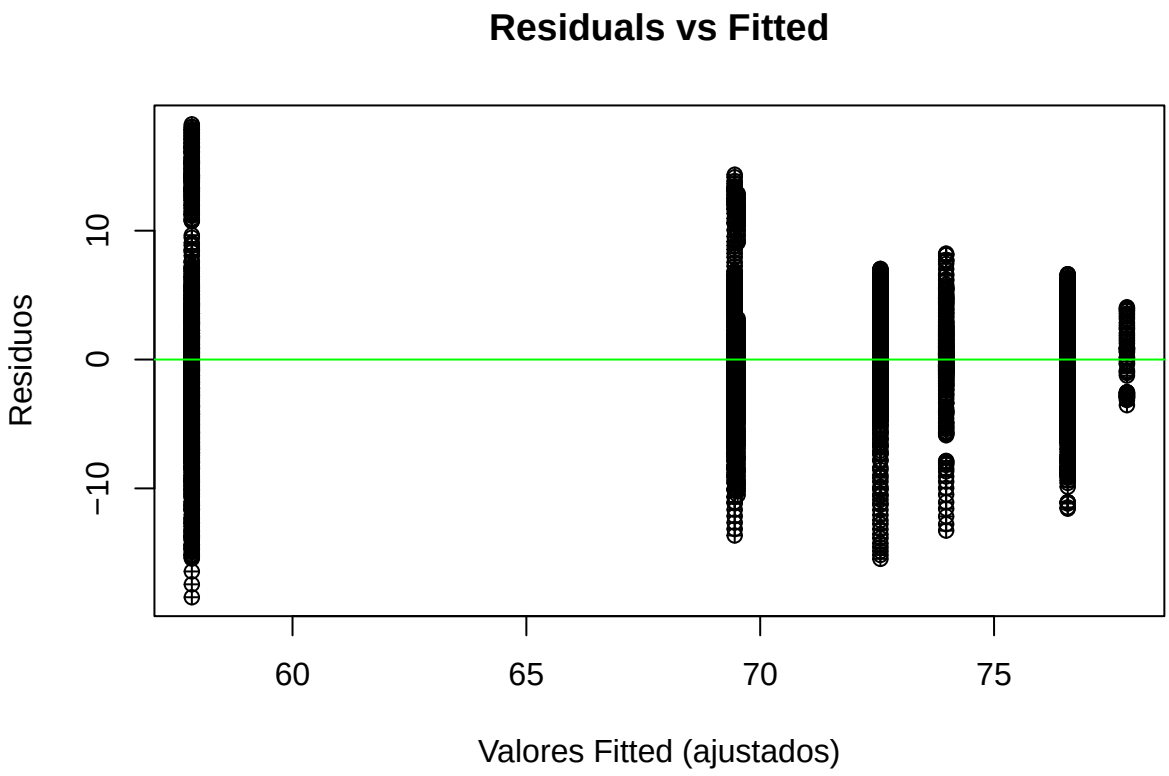


El gráfico cuantil cuantil se usa para saber si un conjunto de datos se ajusta a una dist normal. La linea diagonal (verde) representa los valores esperados si los datos estuviesen perfectamente distribuidos y los puntos son la representación de los datos orgiginales. La mayoría de los puntos se encuentran en la diagonal lo que como se ha explicado antes indica que estos puntos siguen una dis normal con algunas colas.

```
shapiro <- shapiro.test(residuals_model)
print(shapiro)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals_model
## W = 0.97947, p-value < 2.2e-16
```

```
plot(fitted(anova_model), residuals_model,
     main = "Residuals vs Fitted",
     xlab = "Valores Fitted (ajustados)",
     ylab = "Residuos",
     pch = 10)
abline(h = 0, col = "green")
```



La homocedasticidad, implica que la varianza de los errores (residuales) es constante a lo largo de todos los valores predichos del modelo. Lo que se hubiese repesentado mediante una nube de puntos aleatoriamente situados, dado que ese no es el caso y hay patrones claros, estamos ante un caso de heterocedasticidad, y la varianza de los errores residuales aumenta a medida que aumentan los valores ajustados

7 Comparaciones múltiples

Independientemente del resultado obtenido en el apartado anterior, realizad un test de comparación múltiple entre los grupos con corrección de Bonferroni. Este test se aplica cuando el test ANOVA devuelve rechazar la hipótesis nula de igualdad de medias. Por lo tanto, procederemos como si el test ANOVA hubiera dado como resultado el rechazo de la hipótesis nula. Calculad las comparaciones entre grupos con la corrección Bonferroni. Interpretad los resultados.

```
multiple_comparison <- pairwise.t.test(df_2015$Life_expectancy,
                                       df_2015$Region,
                                       p.adjust.method = "bonferroni")
print(multiple_comparison)
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  df_2015$Life_expectancy and df_2015$Region
##
##                      Africa  Asia    Europe  Middle East North America Oceania
## Asia                 < 2e-16 -       -       -           -             -
## Europe               < 2e-16 < 2e-16 -       -           -             -
```

```
## Middle East           < 2e-16 < 2e-16 1.9e-07 -           -             -
## North America         < 2e-16 < 2e-16 1.00000 0.00069     -             -
## Oceania               < 2e-16 1.00000 < 2e-16 1.0e-12     < 2e-16       -
## South-Central America < 2e-16 1.6e-14 < 2e-16 0.05955     5.6e-08       6.1e-08
##
## P value adjustment method: bonferroni
```

Se muestran las diferencias de las medias de la esperanza de vida por regiones, Los valores tan bajos, de nuevo 2e-16 muestran que las comparaciones son significativas 2e-16 < 0.05 & 2e-16 <0.01, de nuevo África es la más notoria, con respecto a las demás regiones. En conclusión las diferencias son relevantes para entender que los factores económicos y sanitarios influyen en la salud de las diferentes regiones del mundo. Europa, América del Norte, y Oceanía presentan esperanzas de vida más altas.

8 ANOVA multifactorial

A continuación, se desea evaluar el efecto sobre la esperanza de vida de la región combinado con si es un país desarrollado o no (Economy_status_Developed).

8.1 Análisis visual de los efectos principales y posibles interacciones

```
datos_agrupados <- df_2015 %>%
  group_by(Region, Economy_status_Developed) %>%
  summarise(Mean_Life_Expectancy = mean(Life_expectancy, na.rm = TRUE)) %>%
  arrange(Region, Economy_status_Developed)
```

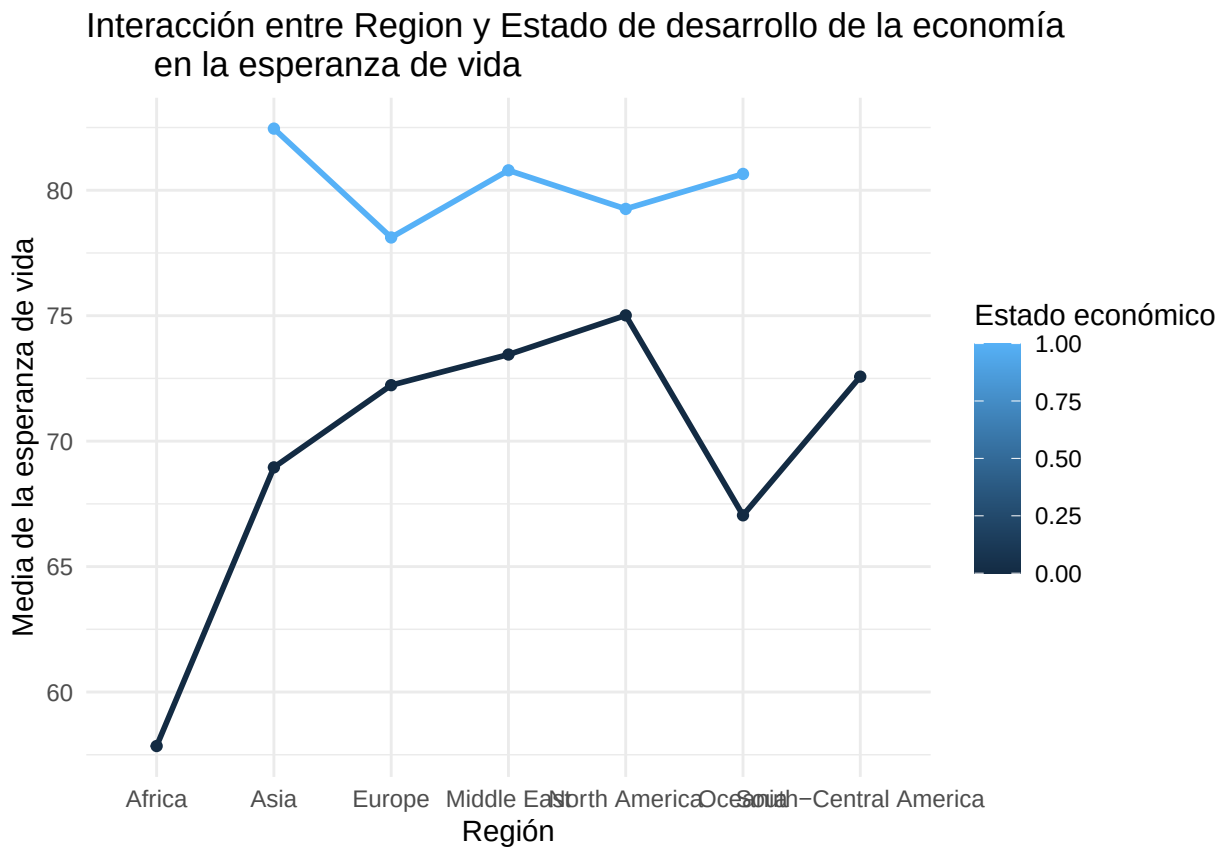
```
## `summarise()` has grouped output by 'Region'. You can override using the
## `.groups` argument.
```

```
head(datos_agrupados,5)
```

```
## # A tibble: 5 x 3
## # Groups:   Region [3]
##   Region Economy_status_Developed Mean_Life_Expectancy
##   <chr>                    <int>                <dbl>
## 1 Africa                       0                 57.8
## 2 Asia                         0                 69.0
## 3 Asia                         1                 82.5
## 4 Europe                       0                 72.2
## 5 Europe                       1                 78.1
```

```
ggplot(datos_agrupados, aes(x = Region, y = Mean_Life_Expectancy,
                            color = Economy_status_Developed,
                            group = Economy_status_Developed)) +
  geom_line(linewidth = 1) +
  geom_point(linewidth = 3) +
  labs(title = "Interacción entre Region y Estado de desarrollo de la economía
       en la esperanza de vida",
       x = "Región",
       y = "Media de la esperanza de vida",
       color = "Estado económico") +
  theme_minimal()
```

```
## Warning in geom_point(linewidth = 3): Ignoring unknown parameters: `linewidth`
```



Gracias al resultado se confirma de nuevo que el desarrollo económico afecta a la esperanza de vida. La linea azul representa la esperanza de vida y la línea negra representa el estado de desarrollo econímico, lo que además nos indica que el desarollo económico no es determinista, ya que el valor más alto de desarollo lo tiene Norte América, y sin embargo la longevidad más alta se la lleva Asia, es decir hay factores externos que influyen de la esperanza de vida.

8.2 Cálculo del modelo

```
modelo_anova <- aov(Life_expectancy ~ Region * Economy_status_Developed,
                    data = df_2015)
summary(modelo_anova)
```

```
##                                  Df Sum Sq Mean Sq F value   Pr(>F)
## Region                            6 155763   25960  877.62  < 2e-16 ***
## Economy_status_Developed          1  11169   11169  377.57  < 2e-16 ***
## Region:Economy_status_Developed   4   1982     496   16.75 1.39e-13 ***
## Residuals                      2852  84363      30
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

La región tiene un valor de p de 2e-16 y un valor de F de 877 esto junto, con los valores de los estados de desarrollo económico muestran que los países desarrollados tienden a tener mayor esperanza de vida. La interacción entre las regiones y estado económico es significativa lo que implica que el impacto económico no es uniforme para todas las regiones, el SS residuals 84363 indica que que los factores y su interacción explican la varibilidad de la esperanza de vida o al menos una gran parte. Con este análisis se destaca la necesidad de salud en países en desarollo así como la creación de políticas que faciliten el acceso a la sanidad, priorizando deciones económicas estratégicas que aumenten la economía regional y maximizar su efectivdad.