Modelo Lineal Múlitple

Trabajo 1: Modelos Lineales

Bladimir Morales Torrez

Enero 2022

Contents

1 Datos a ser estudiados

En el presente trabajo se realizará un modelo de regresión lineal múltiple que explique la variable de respuesta Esperanza de vida de los diferentes países del mundo y que estan registrados en el Banco Mundial, cabe mencionar que los datos fueron procesados antes de su utilización tomando en cuenta que hacen referencia al año 2018 y solo se tomo en cuenta los países que cuentan con la información tanto la variable de respuesta como las explicativas teniendo en total 120 países a nivel mundial que representarán la muestra para el desarrollo del modelo.

La variable de respuesta es:

Y =Esperanza de vida al nacer, total (años).

Las variables explicativas X_i son:

- Tasa de mortalidad, adultos (por cada 1.000 adultos)
- Tasa de fertilidad, total (nacimientos por cada mujer)
- Superficie (kilómetros cuadrados)
- Crecimiento del PIB (% anual)
- Acceso a la electricidad (% de población)
- Desempleo, total (% de la población activa total) (estimación modelado OIT)
- Emisiones de CO2 (kt)
- Emisiones de metano (kt de equivalente de CO2)
- Emisiones de óxido nitroso (miles de toneladas métricas de equivalente de CO2)
- Gasto público en educación, total (% del PIB)

La información fue obtenida del banco de datos que tiene el Banco Mundial y fue descargada del siguiente enlace (https://databank.bancomundial.org/source/world-development-indicators).

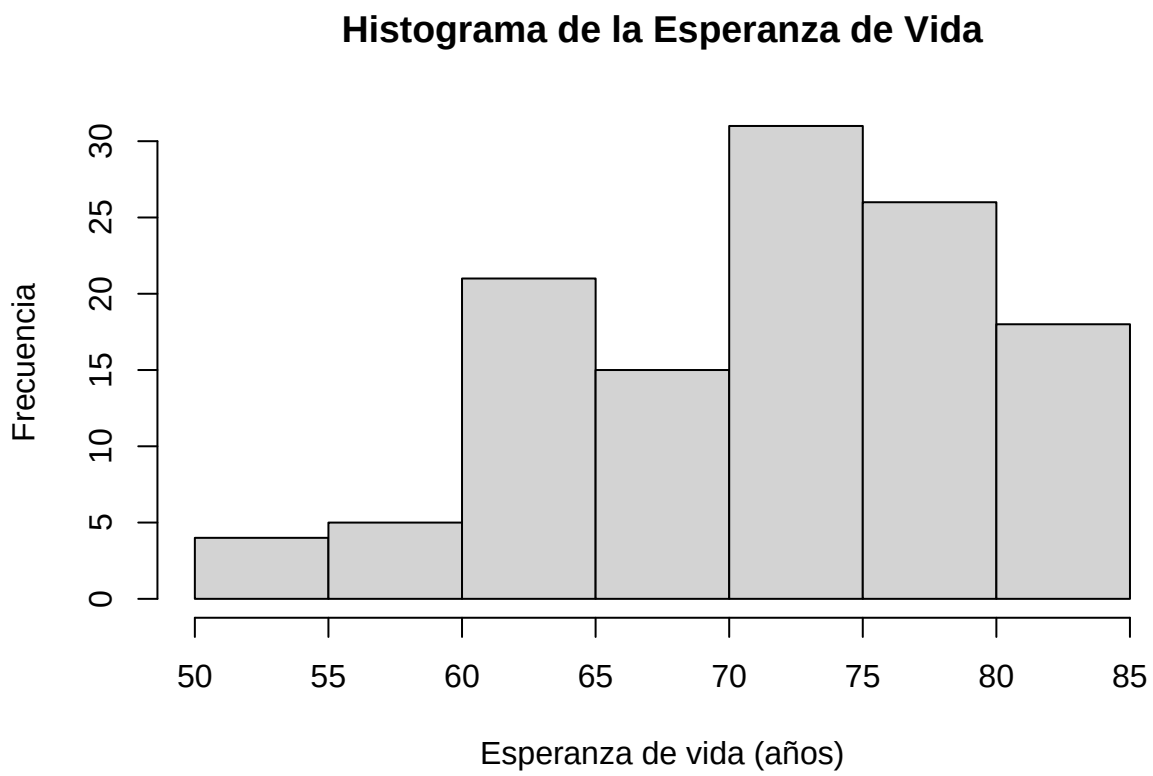
Primero se realizará un análisis descriptivo de las variables en estudio, luego se estimará un modelo de regresión lineal múltiple encontrando el que mejor se ajuste discriminando covariables, para luego realizar el diagnóstico del modelo validando todos los supuestos del modelo, se realizará posteriormente la predicción de datos para que finalmente detallar las conclusiones.

2 Análisis descriptivo

```
bd_pob1<-read_xlsx("./data/bd_poblacion.xlsx")
bd_esp_vida<-bd_pob1[,3:ncol(bd_pob1)]
```

En primer lugar se visualizará el comportamiento de la variable respuesta Esperanza de vida al nacer mediante el histograma.

```
hist(bd_esp_vida$esp_vida,main = "Histograma de la Esperanza de Vida",
     xlab = "Esperanza de vida (años)",
     ylab = "Frecuencia")
```



La esperanza de vida tiene los siguiente estadísticos:

```
summary(bd_esp_vida$esp_vida)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   52.80   64.88   73.12   71.45   76.80   83.75
```

La media de la esperanza de vida de todos los países en estudio es de 71 años aproximadamente, con un máximo de 83 y mínimo de 53 aproximadamente, lo cual se ve reflejado en el histograma de frecuencias donde la mayoría de los países tiene este indicador entre los 70 a 75 años de edad.

En el siguiente cuadro se se observa las correlaciones que tienen las variables en estudio.

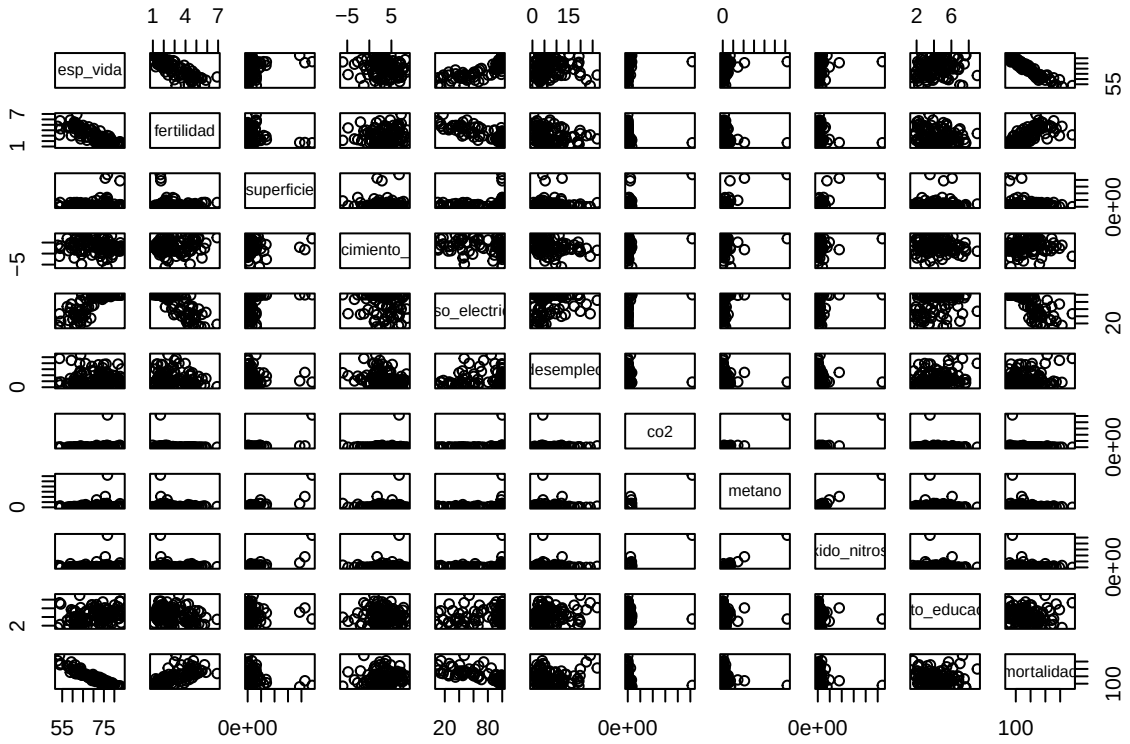
```
cor(bd_esp_vida)
```

```
##                       esp_vida fertilidad   superficie crecimiento_pib
## esp_vida            1.00000000 -0.8337722  0.097108433    -0.159705902
## fertilidad         -0.83377219  1.0000000 -0.058692402     0.130598023
## superficie          0.09710843 -0.0586924  1.000000000    -0.007519633
## crecimiento_pib    -0.15970590  0.1305980 -0.007519633     1.000000000
## acceso_electricidad  0.81630543 -0.8276405  0.076346040    -0.139233006
```

```
## desempleo           -0.01355599 -0.1568483  0.004480272    -0.296474595
## co2                  0.11255860 -0.1294939  0.637828070     0.094004929
## metano               0.09758181 -0.1152920  0.801056355     0.086497462
## oxido_nitroso        0.07486037 -0.0878121  0.808261171     0.099278939
## gasto_educacion      0.22300587 -0.2173722  0.039475663    -0.121371920
## mortalidad          -0.95411977  0.7147177 -0.105701294     0.102297152
##                     acceso_electricidad    desempleo         co2      metano
## esp_vida                      0.81630543 -0.013555988  0.11255860  0.09758181
## fertilidad                   -0.82764053 -0.156848316 -0.12949392 -0.11529199
## superficie                    0.07634604  0.004480272  0.63782807  0.80105635
## crecimiento_pib              -0.13923301 -0.296474595  0.09400493  0.08649746
## acceso_electricidad           1.00000000  0.148661600  0.10758814  0.10942308
## desempleo                     0.14866160  1.000000000 -0.03838354 -0.05818316
## co2                           0.10758814 -0.038383545  1.00000000  0.93281498
## metano                        0.10942308 -0.058183162  0.93281498  1.00000000
## oxido_nitroso                 0.06833360 -0.058317428  0.93701023  0.98564851
## gasto_educacion               0.22673477  0.219033651 -0.04034134 -0.04990278
## mortalidad                   -0.76893603  0.106615112 -0.14004220 -0.13180976
##                     oxido_nitroso gasto_educacion mortalidad
## esp_vida               0.07486037      0.22300587 -0.9541198
## fertilidad            -0.08781210     -0.21737220  0.7147177
## superficie             0.80826117      0.03947566 -0.1057013
## crecimiento_pib        0.09927894     -0.12137192  0.1022972
## acceso_electricidad    0.06833360      0.22673477 -0.7689360
## desempleo             -0.05831743      0.21903365  0.1066151
## co2                    0.93701023     -0.04034134 -0.1400422
## metano                 0.98564851     -0.04990278 -0.1318098
## oxido_nitroso          1.00000000     -0.04343349 -0.1020348
## gasto_educacion       -0.04343349      1.00000000 -0.1227976
## mortalidad            -0.10203477     -0.12279756  1.0000000
```

Se puede observar que la variable de respuesta, cuenta con un alto grado de correlación con las variables de fertilidad, acceso_electricidad y mortalidad, lo cual se puede visualizar en el siguiente gráfico que explica la relación lineal que pueden tener las variables en estudio.

```
plot(bd_esp_vida)
```



3 Estimación

En esta sección se estimará un modelo de regresión lineal clásico con las 10 covariables propuestas, donde la variable respuesta es la esperanza de vida, el fin será obtener un modelo óptimo el cual explique de mejor manera la variable Y y así también que sea parsimonioso. Para esto se tomará los siguientes criterios:

- Pruebas de hipótesis para los coeficientes
- Menor Criterio de Información de Akaike AIC
- Estimación de modelo restringido
- Análisis de varianza ANOVA

3.1 Pruebas de hipótesis para los coeficientes

Primero se estima el modelo con toda la información.

```
mod<-lm(esp_vida~.,data=bd_esp_vida)
summary(mod)
```

```
##
## Call:
## lm(formula = esp_vida ~ ., data = bd_esp_vida)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.3836 -0.8878  0.2011  0.9053  3.0928
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          8.718e+01  1.403e+00  62.141  < 2e-16 ***
## fertilidad          -1.859e+00  1.923e-01  -9.664  2.5e-16 ***
## superficie           3.923e-07  2.094e-07   1.873  0.06373 .
## crecimiento_pib     -1.055e-01  5.331e-02  -1.978  0.05041 .
## acceso_electricidad -5.145e-03  1.012e-02  -0.508  0.61236
## desempleo           -1.532e-02  2.872e-02  -0.533  0.59492
## co2                  4.723e-07  5.244e-07   0.901  0.36979
## metano              -1.852e-05  6.881e-06  -2.692  0.00823 **
## oxido_nitroso        2.291e-05  1.795e-05   1.277  0.20442
## gasto_educacion      2.881e-01  9.217e-02   3.126  0.00227 **
## mortalidad          -6.381e-02  2.634e-03 -24.229  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.46 on 109 degrees of freedom
## Multiple R-squared:  0.9671, Adjusted R-squared:  0.9641
## F-statistic: 320.8 on 10 and 109 DF,  p-value: < 2.2e-16
```

En el modelo lineal se tiene un coeficiente de determinación ajustado de $R_a^2 = 0.9641$, concluyendo que el modelo tiene una alta calidad. También se puede observar que se tiene un estadístico F grande el cual rechaza la hipóstesis nula, teniendo así evidencia estadística para rechazar H_0.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_p = 0$$
$$H_1 : \beta_j \neq 0 \ \text{; para algún } j = 1, ..., p$$

Ahora para cada uno de los coeficientes β la prueba de hipótesis es la siguiente:

$$H_0 : \beta_j = 0$$
$$H_0 : \beta_j \neq 0$$

Observando los p-valores existen covariables que no rechazan la hipótesis nula teniendo así evidencia estadística para decir que no son significativos, las cuales son: superficie, crecimiento del PIB, acceso a la electricidad, emisiones de CO^2 y emisiones de oxido nitroso.

3.2 Menor Criterio de Información de Akaike AIC

Por otro lado el software R con la función `stepAIC()` realiza iteraciones combinando todos los modelos lineales posibles, tomando en cuenta el menor valor del Criterio de Información de Akaike (AIC).

```
stepAIC(mod,trace = F)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + superficie + crecimiento_pib +
##     metano + oxido_nitroso + gasto_educacion + mortalidad, data = bd_esp_vida)
##
```

```
## Coefficients:
##     (Intercept)       fertilidad       superficie  crecimiento_pib
##       8.641e+01       -1.772e+00        2.688e-07       -1.001e-01
##          metano    oxido_nitroso  gasto_educacion       mortalidad
##      -1.790e-05        3.212e-05        2.811e-01       -6.386e-02
```

Posterior de la búsqueda del menor AIC, se proporciona las siguientes variables explicativas: tasa de fertilidad, superficie, crecimiento del PIB, emisiones de metano, emisiones de oxido nitroso, gasto en educación y mortalidad.

Tomando en cuenta las mismas se ajusta nuevamente el modelo.

```
mod1<-lm(esp_vida ~ fertilidad + superficie + crecimiento_pib + metano + oxido_nitroso +
    gasto_educacion+mortalidad, data = bd_esp_vida)
summary(mod1)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + superficie + crecimiento_pib +
##     metano + oxido_nitroso + gasto_educacion + mortalidad, data = bd_esp_vida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.3025 -0.9141  0.1452  0.9828  3.2672
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.641e+01  5.926e-01 145.804  < 2e-16 ***
## fertilidad      -1.772e+00  1.505e-01 -11.775  < 2e-16 ***
## superficie       2.688e-07  1.639e-07   1.640  0.10381
## crecimiento_pib -1.001e-01  5.081e-02  -1.969  0.05139 .
## metano          -1.790e-05  6.574e-06  -2.723  0.00751 **
## oxido_nitroso    3.212e-05  1.547e-05   2.076  0.04018 *
## gasto_educacion  2.811e-01  8.979e-02   3.131  0.00222 **
## mortalidad      -6.386e-02  2.148e-03 -29.737  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.449 on 112 degrees of freedom
## Multiple R-squared:  0.9667, Adjusted R-squared:  0.9646
## F-statistic: 464.8 on 7 and 112 DF,  p-value: < 2.2e-16
```

Aún persiste la no significancia del 5% de las variables superficie, crecimiento del PIB y oxido nitroso, eliminando las mismas se ajusta nuevamente el modelo.

```
mod2<-lm(esp_vida ~ fertilidad + metano + gasto_educacion+mortalidad, data = bd_esp_vida)
summary(mod2)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + metano + gasto_educacion +
##     mortalidad, data = bd_esp_vida)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9495 -0.9213  0.2880  1.0082  3.1891
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.583e+01  5.860e-01 146.477  < 2e-16 ***
## fertilidad      -1.744e+00  1.552e-01 -11.240  < 2e-16 ***
## metano          -1.916e-06  1.142e-06  -1.678 0.096010 .
## gasto_educacion  3.328e-01  9.212e-02   3.613 0.000451 ***
## mortalidad      -6.384e-02  2.223e-03 -28.715  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.51 on 115 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9616
## F-statistic: 746.5 on 4 and 115 DF,  p-value: < 2.2e-16
```

3.3 Estimación del modelo restringido

Resultando ahora no significativo a la covariable emisiones de metano. Antes de eliminar esta covariable se realizara una prueba de significancia para conjuntos de parámtros, en el primer modelo se tendra todas covariables del anterior modelo y se estimará otro modelo sin la variable de emisiones de metano.

Las hipótesis de interés serán:

$$H_0 : \beta_{p-q+1} = \beta_{p-q+2} = ... = \beta_p = 0$$

$$H_1 : H_0 \text{ es falso}$$

Entonces el estadístico de prueba será:

$$F_c = \frac{(SCE_R - SCE)/q}{SCE/(n-p-1)} \sim F_{(q;n-p-1)}$$

```
# Modelo Sin restringir
mod2_1<-mod2
summary(mod2_1)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + metano + gasto_educacion +
##     mortalidad, data = bd_esp_vida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9495 -0.9213  0.2880  1.0082  3.1891
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.583e+01  5.860e-01 146.477  < 2e-16 ***
## fertilidad      -1.744e+00  1.552e-01 -11.240  < 2e-16 ***
## metano          -1.916e-06  1.142e-06  -1.678 0.096010 .
```

```
## gasto_educacion  3.328e-01  9.212e-02   3.613 0.000451 ***
## mortalidad      -6.384e-02  2.223e-03 -28.715  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.51 on 115 degrees of freedom
## Multiple R-squared:  0.9629, Adjusted R-squared:  0.9616
## F-statistic: 746.5 on 4 and 115 DF,  p-value: < 2.2e-16
```

```
# Modelo restringido
mod2_2<-lm(esp_vida ~ fertilidad + gasto_educacion + mortalidad, data = bd_esp_vida)
summary(mod2_2)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + gasto_educacion + mortalidad,
##     data = bd_esp_vida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9198 -0.8747  0.2262  1.0437  3.1932
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     85.633528   0.578341 148.068  < 2e-16 ***
## fertilidad      -1.732991   0.156251 -11.091  < 2e-16 ***
## gasto_educacion  0.344270   0.092584   3.718  0.00031 ***
## mortalidad      -0.063586   0.002235 -28.445  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.522 on 116 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.961
## F-statistic: 979.1 on 3 and 116 DF,  p-value: < 2.2e-16
```

Entonces:

$$SCE = S^2 * (n - p - 1) = (1.51)^2 * 115 = 262.2115$$

$$SCE_R = S^2 * (n - p - 1) = (1.522)^2 * 116 = 268.7121$$

Así

$$F_c = \frac{(268.7121 - 262.2115)/1}{262.2115/115} = 2.851015$$

```
qf(0.95,1,115)
```

```
## [1] 3.923599
```

Así se tiene

$$F_c = 2.851015 < 3.923599 = F_{(1-\alpha;1;115)}$$

Así no se rechaza la hipótesis nula donde el $\beta_q = 0$ que corresponde a la emisión del metano.

```r
pf(2.851015,1,115)
```

```
## [1] 0.9059738
```

Para el valor-p es 0.9059738> 0.05, así no se rechaza la hipótesis nula.

Por lo tanto se tiene evidencia estadística para decir que el $\beta_q = 0$ correspondiente a la emisión del metano, no teniendo así significancia en el modelo y así se decide eliminanr el mismo para ajustar el nuevo modelo.

```r
mod2<-lm(esp_vida ~ fertilidad + gasto_educacion + mortalidad, data = bd_esp_vida)
summary(mod2)
```

```
##
## Call:
## lm(formula = esp_vida ~ fertilidad + gasto_educacion + mortalidad,
##     data = bd_esp_vida)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9198 -0.8747  0.2262  1.0437  3.1932
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     85.633528   0.578341 148.068  < 2e-16 ***
## fertilidad      -1.732991   0.156251 -11.091  < 2e-16 ***
## gasto_educacion  0.344270   0.092584   3.718  0.00031 ***
## mortalidad      -0.063586   0.002235 -28.445  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.522 on 116 degrees of freedom
## Multiple R-squared:  0.962,  Adjusted R-squared:  0.961
## F-statistic: 979.1 on 3 and 116 DF,  p-value: < 2.2e-16
```

Finalmente se encuentra un modelo con covariables significativas que explican de manera óptima el modelo ya que se tiene un $R_a^2 = 0.961$. Por tanto el modelo será:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X i_3 + \varepsilon_i$$

Donde:

Y_i : Esperanza de vida al nacer.
X_{i1} : Tasa de fertilidad, total (nacimientos por cada mujer).
X_{i2} : Gasto público en educación, total (% del PIB).
X_{i3} : Tasa de mortalidad, adultos (por cada 1.000 adultos).

El modelo estimado será:

$$\text{Esp. vida}_i = 85.63 - 1.73 \text{ fertilidad}_i + 0.34 \text{ gasto educación}_i - 0.06 \text{mortalidad}_i$$

En la interpretación de los coeficientes se tiene:

- Si la tasa de fertilidad aumenta en un punto se espera que la esperanza de vida al nacer disminuya en 1.73 años aproximadamente, cuando las otras covariables son 0.

- Si el gasto público en educación aumenta en 1% se espera que la esperanza de vida aumente en 0.34 años aproximadamente, cuando las otras covariables son 0.
- Si la tasa de mortalidad aumenta en un punto se espera que la esperanza de vida al nacer disminuya en 0.06 años aproximadamente, cuando las otras covariables son 0.

3.4 Análisis de varianza ANOVA

Para este modelo que posiblemente sea óptimo y parsimonioso, se realizará un análisis de varianza ANOVA.

```
anova(mod2)
```

```
## Analysis of Variance Table
##
## Response: esp_vida
##                  Df Sum Sq Mean Sq   F value  Pr(>F)
## fertilidad        1 4915.5  4915.5 2122.5423 < 2e-16 ***
## gasto_educacion   1   12.9    12.9    5.5905 0.01972 *
## mortalidad        1 1873.8  1873.8  809.1114 < 2e-16 ***
## Residuals       116  268.6     2.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede observar que para las covariables seleccionadas almenos una si explica de mejor manera a la esperanza de vida.

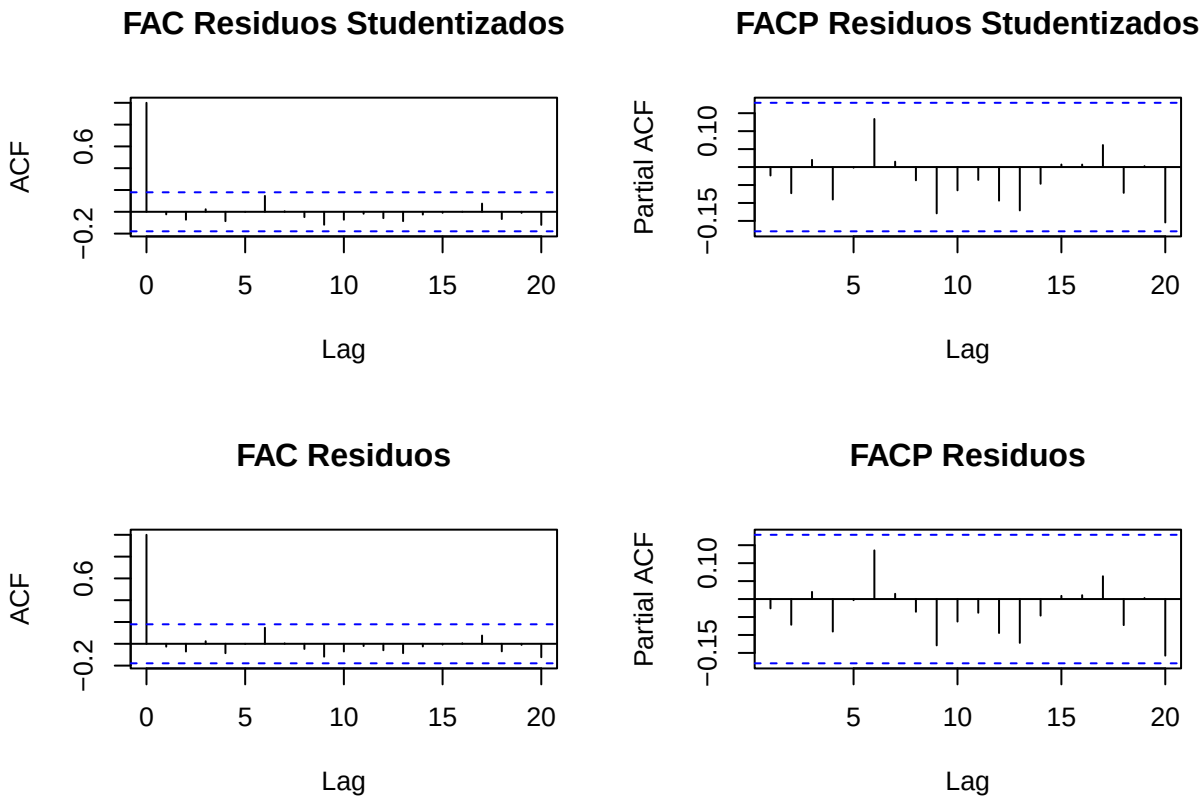
4 Análisis de residuos

```
bd_mod<-bd_pob1 %>%
  dplyr::select(pais,codigo,esp_vida,fertilidad,gasto_educacion,mortalidad) %>%
  mutate(residuos=residuals(mod2),
         residuos_student=rstudent(mod2),
         ajustados=fitted(mod2))
```

Se realizará el análisis de residuos para verificar los supuestos de normalidad, para este cometido se utilizarán los residuos obtenidos por la regresión y así también los Studentizados.

Primero se grafica la función de autocorrelación, el cual muestra que se asemeja al comportamiento de ruido blando vale decir media 0 y varianza constante tanto en los residuos de la regresión como los studentizados.

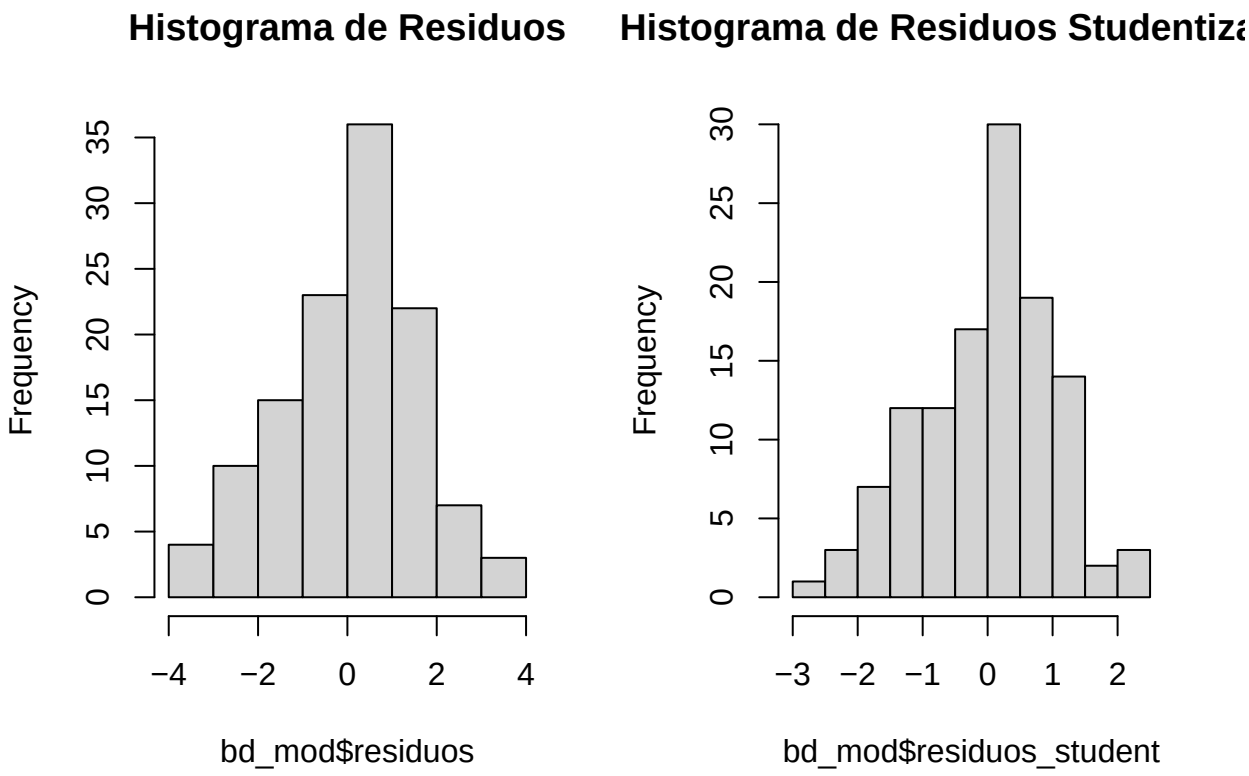
```
#Análisis de residuos
par(mfrow=c(2,2))
acf(bd_mod$residuos_student,main="FAC Residuos Studentizados")
pacf(bd_mod$residuos_student,main="FACP Residuos Studentizados")
acf(bd_mod$residuos,main="FAC Residuos")
pacf(bd_mod$residuos,main="FACP Residuos")
```



4.1 Supuesto de normalidad

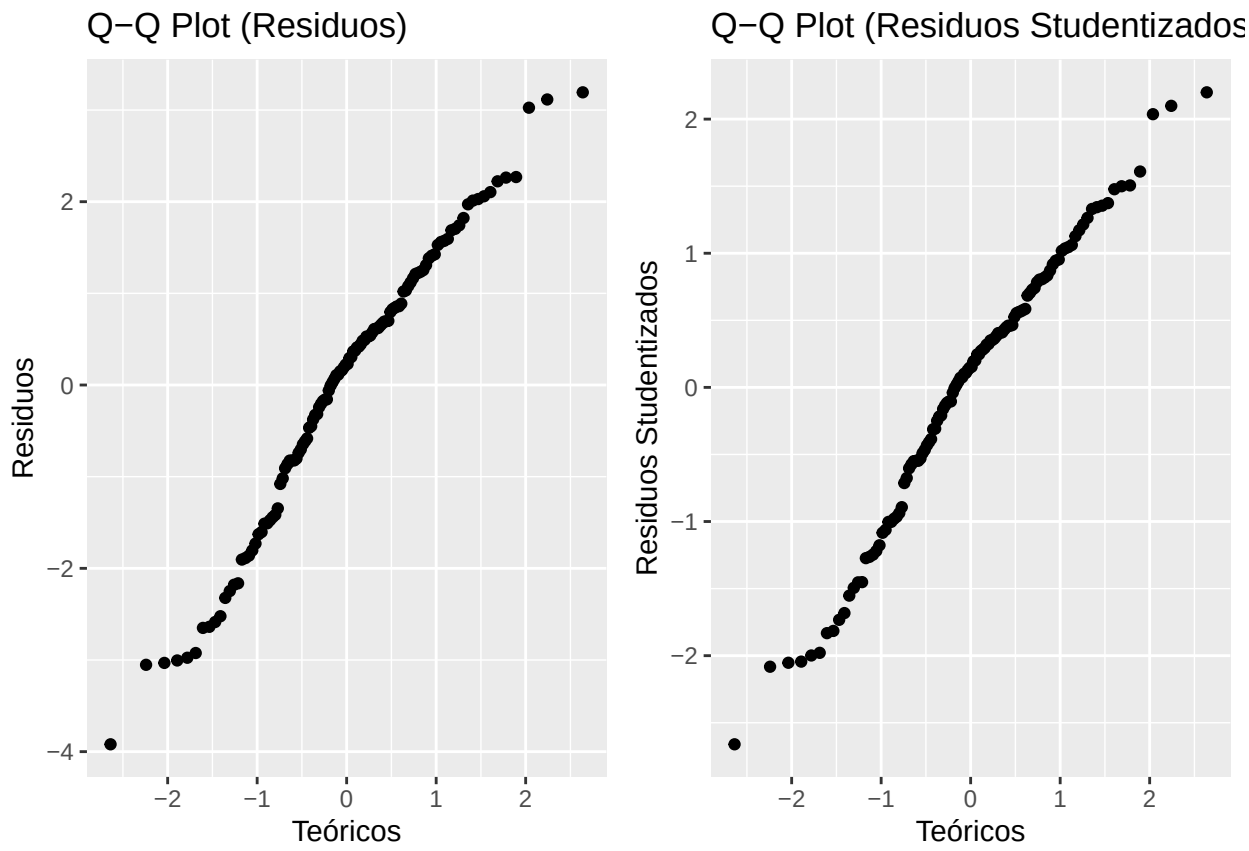
Se grafica el histograma de los residuos de la regresión y los studentizados, se observa que posiblemente tengan un comportamiento normal ya que los coeficientes estimados son significativamente iguales a 0.

```
#Normalidad
par(mfrow=c(1,2))
hist(bd_mod$residuos,main="Histograma de Residuos")
hist(bd_mod$residuos_student,main="Histograma de Residuos Studentizados")
```



Se grafica el qq-plot, donde se puede observar que el diagrama es casi lineal lo cual muestra un posible comportamiento normal

```
grid.arrange(
  ggplot(bd_mod)+
    stat_qq(aes(sample=residuos))+
    labs(title="Q-Q Plot (Residuos)",y="Residuos",x="Teóricos")
  ,
  ggplot(bd_mod)+
    stat_qq(aes(sample=residuos_student))+
    #geom_abline(color="blue")+
    labs(title="Q-Q Plot (Residuos Studentizados)",y="Residuos Studentizados",x="Teóricos")
  ,ncol=2
)
```



Para tener evidencia estadística se realizará la dócima de Jarque-Bera para modelos de regresión.

H_0 : residuos son normales

```
jb.norm.test(bd_mod$residuos)
```

```
##
##  Jarque-Bera test for normality
##
## data:  bd_mod$residuos
## JB = 3.2034, p-value = 0.127
```

```
jb.norm.test(bd_mod$residuos_student)
```

```
##
##  Jarque-Bera test for normality
##
## data:  bd_mod$residuos_student
## JB = 3.0108, p-value = 0.1615
```

Tanto para los residuos de la regresión como los studentizados el p-valor es mayor a 0.05 teniendo evidencia estadística para no rechazar H_0, así los residuos son normales.

4.2 Supuestos de multicolinealidad

Se muestra las correlaciones de las covariables, donde las variables más correlacionadas estan entre la tasa de fertilidad y la tasa de mortalidad.

```
cor(bd_mod[,4:6])
```

```
##                 fertilidad gasto_educacion mortalidad
## fertilidad       1.0000000      -0.2173722  0.7147177
## gasto_educacion -0.2173722       1.0000000 -0.1227976
## mortalidad       0.7147177      -0.1227976  1.0000000
```

Se construirá regresiones auxiliares, para determinar la significancia entre ellas.

Si tenemos la tasa de fertilidad como variable respuesta, se tiene que la tasa de mortalidad es significativa para el modelo, pero su $R_a^2 = 0.5198$ lo cual indica que tiene un ajuste relativamente bajo.

```
mod_aux1<-lm(fertilidad ~ gasto_educacion+mortalidad,bd_mod)
summary(mod_aux1)
```

```
##
## Call:
## lm(formula = fertilidad ~ gasto_educacion + mortalidad, data = bd_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5410 -0.5338 -0.0859  0.5145  3.2987
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.5850417  0.3092271   5.126 1.18e-06 ***
## gasto_educacion -0.1106384  0.0538164  -2.056    0.042 *
## mortalidad       0.0101613  0.0009311  10.914  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9004 on 117 degrees of freedom
## Multiple R-squared:  0.5279, Adjusted R-squared:  0.5198
## F-statistic: 65.41 on 2 and 117 DF,  p-value: < 2.2e-16
```

Si tenemos el gasto público en educación (% PIB) como variable respuesta, se tiene que la tasa de fertilidad es significativa para el modelo, pero su $R_a^2 = 0.0.03317$ lo cual indica que tiene un ajuste bajo.

```
mod_aux2<-lm(gasto_educacion~ fertilidad +mortalidad ,bd_mod)
summary(mod_aux2)
```

```
##
## Call:
## lm(formula = gasto_educacion ~ fertilidad + mortalidad, data = bd_mod)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -2.7023 -1.1218 -0.1299  1.0691  4.6664
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.047363   0.340248  14.834   <2e-16 ***
## fertilidad  -0.315123   0.153281  -2.056    0.042 *
## mortalidad   0.001152   0.002230   0.517    0.606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.52 on 117 degrees of freedom
## Multiple R-squared:  0.04942,    Adjusted R-squared:  0.03317
## F-statistic: 3.041 on 2 and 117 DF,  p-value: 0.05157
```

Si tenemos la tasa de mortalidad como variable respuesta, se tiene que la tasa de fertilidad es significativa para el modelo, pero su $R_a^2 = 0.5036$ lo cual indica que tiene un ajuste relativamente bajo.

```
mod_aux3<-lm(mortalidad~ fertilidad +gasto_educacion,bd_mod)
summary(mod_aux3)
```

```
##
## Call:
## lm(formula = mortalidad ~ fertilidad + gasto_educacion, data = bd_mod)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -133.428  -36.319   -7.849   29.092  291.536
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)       20.655     23.842   0.866    0.388
## fertilidad        49.645      4.549  10.914   <2e-16 ***
## gasto_educacion    1.975      3.825   0.517    0.606
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 62.94 on 117 degrees of freedom
## Multiple R-squared:  0.5119, Adjusted R-squared:  0.5036
## F-statistic: 61.36 on 2 and 117 DF,  p-value: < 2.2e-16
```

Para determinar también la multicolinealidad se utilizará el factor de inflación de la varianza.

$$VIF = \frac{1}{1 - R_j^2} \qquad ; j = 1, 2, 3$$

donde el R_j^2 es el coeficiente de determinación de las regresiones auxiliares.

```
r1<-summary(mod_aux1);1/(1-r1$r.squared)
```

```
## [1] 2.11809
```

```r
r2<-summary(mod_aux2);1/(1-r2$r.squared)
```

```
## [1] 1.051987
```

```r
r3<-summary(mod_aux3);1/(1-r3$r.squared)
```

```
## [1] 2.048905
```

El software R también tiene la función `vif()` que calcula el VIF

```r
car::vif(mod2)
```

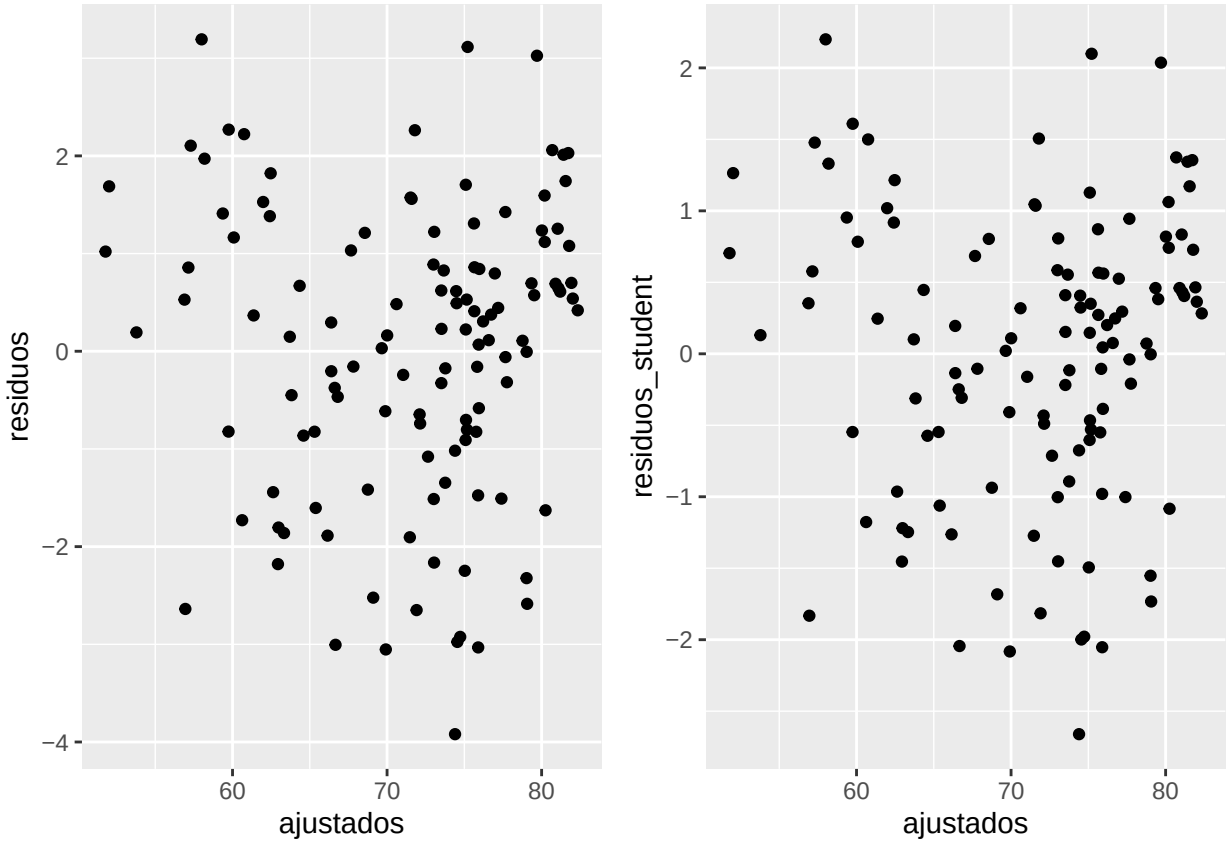
```
##      fertilidad gasto_educacion      mortalidad
##        2.118090        1.051987        2.048905
```

En todos los casos el $VIF < 10$ lo que indica que no existe multicolinealidad entre las covariables.

4.3 Supuesto de homocedasticidad

Primero se hará un análisis gráfico:

```r
#Homocedasticidad
grid.arrange(
  ggplot(bd_mod,aes(ajustados,residuos))+
    geom_point()
  ,
    ggplot(bd_mod,aes(ajustados,residuos_student))+
    geom_point(),
ncol=2)
```



Ahora se utilizara el contraste de Breuch-Pagan, donde la hipótesis son:

$$H_0 : \text{Errores son homocedasticos}$$
$$H_0 : \text{Errores son heterocedasticos}$$

```
mod_res<-lm(residuos^2~ fertilidad + gasto_educacion + mortalidad, data = bd_mod)

estadistico<-nrow(bd_mod)*summary(mod_res)$r.squared
valorp<-pchisq(estadistico,df=3,lower.tail = F)
cbind(estadistico,valorp)
```

```
##      estadistico    valorp
## [1,]    3.498525 0.3209535
```

```
bptest(mod2)
```

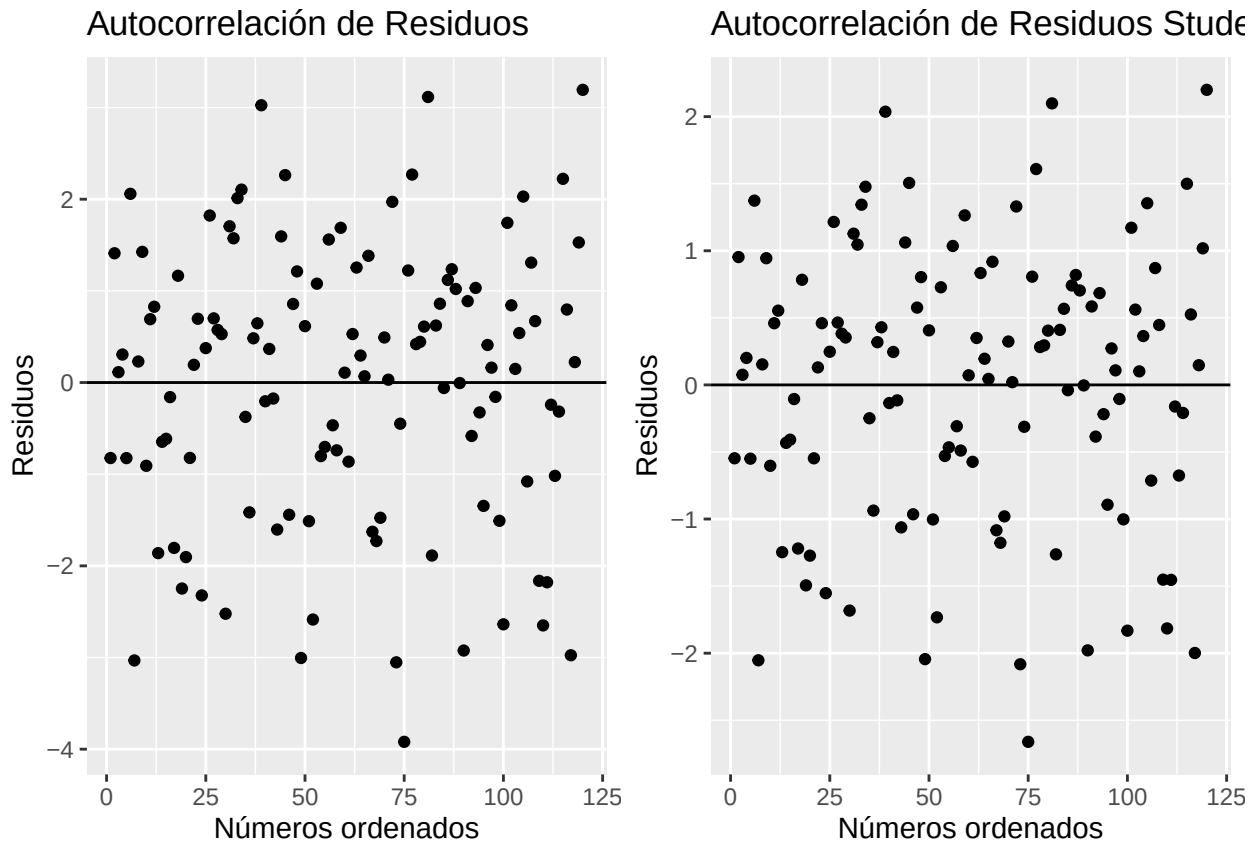
```
##
##  studentized Breusch-Pagan test
##
## data:  mod2
## BP = 3.4985, df = 3, p-value = 0.321
```

El p-valor es mayor que 0.05 con lo cual existe evidencia estadística para no rechazar la hipótesis nula, vale decir que los errores son homocedásticos.

4.4 Supuesto de autocorrelación

Se gráfica los residuos de la regresión y los residuos studentizados, los cuales no siguen un patron sistemático lo cual mostraria que no son correlacionados.

```
grid.arrange(
  bd_mod %>%
    ggplot(aes(x=c(1:nrow(.)),y=residuos))+
      geom_point()+
      #geom_line(color="blue")+
      geom_hline(yintercept = 0)+
      labs(title="Autocorrelación de Residuos",
           y="Residuos",
           x="Números ordenados"),
  bd_mod %>%
    ggplot(aes(x=c(1:nrow(.)),y=residuos_student))+
      geom_point()+
      #geom_line(color="blue")+
      geom_hline(yintercept = 0)+
      labs(title="Autocorrelación de Residuos Studentizados",
           y="Residuos",
           x="Números ordenados")
  ,ncol=2)
```



Se utilizará ahora el contraste de Durbin Watson, que tiene las siguientes hipótesis:

$$H_0 : \phi = 0 \text{ (no existe autocorrelación)}$$

$$H_1 : \phi \neq 0 \text{ (existe autocorrelación)}$$

```
dwtest(mod2)
```

```
##
##  Durbin-Watson test
##
## data:  mod2
## DW = 2.0115, p-value = 0.5243
## alternative hypothesis: true autocorrelation is greater than 0
```

El $DW = 2.0115$ y el p-valor es mayor a 0.05, con lo cual no rechazamos la hipótesis nula así tenemos evidencia estadística para decir que no existe autocorrelación de grado 1 en los residuos.

Ahora se utilizará ahora el contraste de Breuch Godfrey, que tiene las siguientes hipótesis:

$$H_0 : \varepsilon_i \sim N(0, \sigma^2) \text{ son ruido blanco}$$
$$H_1 : \varepsilon_i \text{ es un proceso } AR(p)$$

```
bgtest(mod2)
```

```
##
##  Breusch-Godfrey test for serial correlation of order up to 1
##
## data:  mod2
## LM test = 0.085577, df = 1, p-value = 0.7699
```

El p-valor es mayor a 0.05, con lo cual no rechazamos la hipótesis nula así tenemos evidencia estadística para decir que no existe autocorrelación de grado p en los residuos.

5 Predicción

Como el modelo

$$\text{Esp. vida}_i = 85.63 - 1.73 \text{ fertilidad}_i + 0.34 \text{ gasto educación}_i - 0.06\text{mortalidad}_i$$

cumplió todos los supuestos podemos pasar a realizar las predicciones, como los respectivos intervalos de confianza, para los siguientes datos.

$$\text{Esp. vida}_i = 85.63 - 1.73(3)_i + 0.34(10) - 0.06(100)$$

```
x0<-data.frame(fertilidad=3,gasto_educacion=10,mortalidad=100)
predict(mod2,x0,interval = "confidence")
```

```
##        fit      lwr      upr
## 1 77.51866 76.36937 78.66795
```

Se espera que cuando tenemos una tasa de fertilidad de 3, un gasto en educación en porcentajes del PIB del 10% y una tasa de mortalidad de 100 cada 1000 adultos, la esperanza de vida de un país será de 77.52 años de vida, que puede variar entre 76.37 a 78.67 años según el valor medio de las predicciones.

```
predict(mod2,x0,interval = "prediction")
```

```
##        fit      lwr      upr
## 1 77.51866 74.29288 80.74444
```

Se espera que cuando tenemos una tasa de fertilidad de 3, un gasto en educación en porcentajes del PIB del 10% y una tasa de mortalidad de 100 cada 1000 adultos, la esperanza de vida de un país será de 77.52 años de vida, que puede variar entre 74.29 a 80.74 años para el valor de la predicción.

Ahora se genera diferentes valores para ver el comportamiento de las predicciones.

```
x<-data.frame(fertilidad=c(1,5),gasto_educacion=c(10,2),mortalidad=c(20,50))
predict(mod2,x,interval="confidence")
```

```
##        fit      lwr      upr
## 1 86.07152 84.96627 87.17677
## 2 74.47781 73.30729 75.64834
```

```
predict(mod2,x,interval="prediction")
```

```
##        fit      lwr      upr
## 1 86.07152 82.86116 89.28187
## 2 74.47781 71.24440 77.71122
```

La esperanza de vida al nacer sea mayor si tiene una tasa de fertilidad y mortalidad baja, y un gasto en educación alto. Mientras que la esperanza de vida sera menor si pasa lo contrario a lo mencionado anteriormente.

6 Conclusiones

En este trabajo se estudio el comportamiento de la variable de respuesta esperanza de vida al nacer teniendo en un principio diez covariables obtenidas de diferentes paises del mundo con datos oficiales del Banco Mundial, de las cuales en la búsqueda del modelo lineal múltiple óptimo se determino que existen tres variables significativas para la variable de respuesta los cuales son:

- Tasa de mortalidad, adultos (por cada 1.000 adultos)
- Tasa de fertilidad, total (nacimientos por cada mujer)
- Gasto público en educación, total (% del PIB)

El modelo ajustado con estas tres covariables tiene un coeficiente de determinación ajustado de 0.961, lo cual indica la calidad del modelo.

Posteriormente se hizo la validación del modelo realizando pruebas visuales y estadísticas para verificar los supuestos de normalidad, multicolinealidad, homocedasticidad y autocorrelación de los residuos estimados por la regresión y los residuos studentizados, donde se cumplieron todos los supuestos.

Finalmente se realizaron predicciones bajo el modelo de regresión, donde cualquier país puede trabajar en aumentar o reducir estas covariables para tener una esperanza de vida al nacer alta o baja, mostrando así que si se quiere aumentar el indicador se deberia reducir lo más posible la tasa de fertilidad y mortalidad y aumentar el porcentaje de gasto público en educación con respecto al porcentaje del PIB.

Es necesario mencionar que la variable de respuesta tiene soporte positivo y en la gráfica del histograma tenia una cierta asimetría sesgada a la derecha, donde también existen datos atípicos para tal efecto se recomendaría indagar en modelos asimétricos, para verificar si se tiene un mejor ajuste y así una más confiable predicción.