---

title: “swiss” author: “Juan Manuel Cabrera” date: “2023-08-09” output: pdf_document: latex_engine: xelatex —

Objetivo del ejercicio

El objetivo del ejercicio es buscar un modelo que explique la variable fertiliy a partir de las otras variables.

Librerias

```
library(ggplot2)
library(MASS)
library(ppcor)
library(GGally)
```

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

```
library(relaimpo)
```

```
## Loading required package: boot

## Loading required package: survey

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##
## Attaching package: 'survival'

## The following object is masked from 'package:boot':
##
##     aml

##
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':
##
##     dotchart

## Loading required package: mitools

## This is the global version of package relaimpo.

## If you are a non-US user, a version with the interesting additional metric pmvd is available

## from Ulrike Groempings web site at prof.beuth-hochschule.de/groemping.
```

```r
library(car)
```

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

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

```
## The following object is masked from 'package:boot':
##
##     logit
```

1. Carga dataframe

```r
data <- swiss
attach(data)
data
```

```
##              Fertility Agriculture Examination Education Catholic
## Courtelary        80.2        17.0          15        12     9.96
## Delemont          83.1        45.1           6         9    84.84
## Franches-Mnt      92.5        39.7           5         5    93.40
## Moutier           85.8        36.5          12         7    33.77
## Neuveville        76.9        43.5          17        15     5.16
## Porrentruy        76.1        35.3           9         7    90.57
## Broye             83.8        70.2          16         7    92.85
## Glane             92.4        67.8          14         8    97.16
## Gruyere           82.4        53.3          12         7    97.67
## Sarine            82.9        45.2          16        13    91.38
## Veveyse           87.1        64.5          14         6    98.61
## Aigle             64.1        62.0          21        12     8.52
## Aubonne           66.9        67.5          14         7     2.27
## Avenches          68.9        60.7          19        12     4.43
## Cossonay          61.7        69.3          22         5     2.82
## Echallens         68.3        72.6          18         2    24.20
## Grandson          71.7        34.0          17         8     3.30
## Lausanne          55.7        19.4          26        28    12.11
## La Vallee         54.3        15.2          31        20     2.15
## Lavaux            65.1        73.0          19         9     2.84
## Morges            65.5        59.8          22        10     5.23
## Moudon            65.0        55.1          14         3     4.52
## Nyone             56.6        50.9          22        12    15.14
## Orbe              57.4        54.1          20         6     4.20
## Oron              72.5        71.2          12         1     2.40
## Payerne           74.2        58.1          14         8     5.23
## Paysd'enhaut      72.0        63.5           6         3     2.56
## Rolle             60.5        60.8          16        10     7.72
## Vevey             58.3        26.8          25        19    18.46
## Yverdon           65.4        49.5          15         8     6.10
## Conthey           75.5        85.9           3         2    99.71
## Entremont         69.3        84.9           7         6    99.68
## Herens            77.3        89.7           5         2   100.00
## Martigwy          70.5        78.2          12         6    98.96
```

```
## Monthey           79.4        64.9          7         3    98.22
## St Maurice        65.0        75.9          9         9    99.06
## Sierre            92.2        84.6          3         3    99.46
## Sion              79.3        63.1         13        13    96.83
## Boudry            70.4        38.4         26        12     5.62
## La Chauxdfnd      65.7         7.7         29        11    13.79
## Le Locle          72.7        16.7         22        13    11.22
## Neuchatel         64.4        17.6         35        32    16.92
## Val de Ruz        77.6        37.6         15         7     4.97
## ValdeTravers      67.6        18.7         25         7     8.65
## V. De Geneve      35.0         1.2         37        53    42.34
## Rive Droite       44.7        46.6         16        29    50.43
## Rive Gauche       42.8        27.7         22        29    58.33
##              Infant.Mortality
## Courtelary               22.2
## Delemont                 22.2
## Franches-Mnt             20.2
## Moutier                  20.3
## Neuveville               20.6
## Porrentruy               26.6
## Broye                    23.6
## Glane                    24.9
## Gruyere                  21.0
## Sarine                   24.4
## Veveyse                  24.5
## Aigle                    16.5
## Aubonne                  19.1
## Avenches                 22.7
## Cossonay                 18.7
## Echallens                21.2
## Grandson                 20.0
## Lausanne                 20.2
## La Vallee                10.8
## Lavaux                   20.0
## Morges                   18.0
## Moudon                   22.4
## Nyone                    16.7
## Orbe                     15.3
## Oron                     21.0
## Payerne                  23.8
## Paysd'enhaut             18.0
## Rolle                    16.3
## Vevey                    20.9
## Yverdon                  22.5
## Conthey                  15.1
## Entremont                19.8
## Herens                   18.3
## Martigwy                 19.4
## Monthey                  20.2
## St Maurice               17.8
## Sierre                   16.3
## Sion                     18.1
## Boudry                   20.3
## La Chauxdfnd             20.5
```

```
## Le Locle                 18.9
## Neuchatel                23.0
## Val de Ruz               20.0
## ValdeTravers             19.5
## V. De Geneve             18.0
## Rive Droite              18.2
## Rive Gauche              19.3
```

2. Análisis de los datos

```
str(data)
```

```
## 'data.frame':    47 obs. of  6 variables:
##  $ Fertility       : num  80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 ...
##  $ Agriculture     : num  17 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 ...
##  $ Examination     : int  15 6 5 12 17 9 16 14 12 16 ...
##  $ Education       : int  12 9 5 7 15 7 7 8 7 13 ...
##  $ Catholic        : num  9.96 84.84 93.4 33.77 5.16 ...
##  $ Infant.Mortality: num  22.2 22.2 20.2 20.3 20.6 26.6 23.6 24.9 21 24.4 ...
```

El dataset está formado por 47 observaciones y 6 variables.

No es necesario hacer alguna transformación en los tipos de variable.

2.1. Comprobamos si existen datos vacios

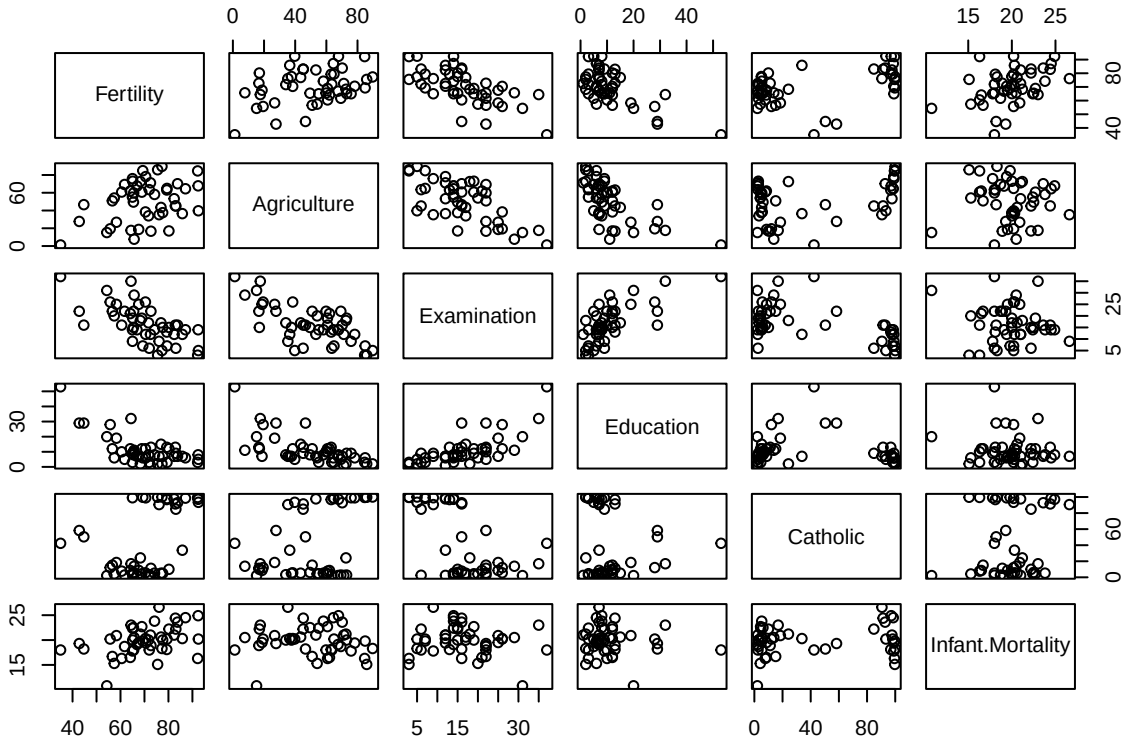
```
sum(is.na(data))
```

```
## [1] 0
```

No se observan datos vacios.

2.2. Visualizar los datos

```
plot(data)
```



1.3. Análisis Shapiro-Wilks

A continuación se comprueba si los datos están normalizados con el test Shapiro-Wilk.

```
shapiro.test(Agriculture)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Agriculture
## W = 0.96643, p-value = 0.193
```

Resultado: p-value > 0.05, rechazamos la hipótesis nula y aceptamos que existe normalidad con el 95% de confianza.

```
shapiro.test(Examination)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Examination
## W = 0.96962, p-value = 0.2563
```

Resultado: p-value > 0.05, existe normalidad.

```r
shapiro.test(Education)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Education
## W = 0.7482, p-value = 1.312e-07
```

Resultado: p-value < 0.05, aceptamos la hipótesis nula, rechazamos normalidad.

```r
shapiro.test(Catholic)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Catholic
## W = 0.7463, p-value = 1.205e-07
```

Resultado: p-value < 0.05, rechazamos normalidad.

```r
shapiro.test(Infant.Mortality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Infant.Mortality
## W = 0.97762, p-value = 0.4978
```

p-value $= 0.05$, aceptamos normalidad en los datos.

2.3. Correlación de Spearman

Analizamos la correlación entre las variables del dataframe.

```r
cor(data, method='spearman')
```

```
##                   Fertility Agriculture Examination   Education    Catholic
## Fertility         1.0000000   0.2426643 -0.66090300 -0.44325769  0.41364556
## Agriculture       0.2426643   1.0000000 -0.59885994 -0.65046381  0.28868781
## Examination      -0.6609030  -0.5988599  1.00000000  0.67460383 -0.47505753
## Education        -0.4432577  -0.6504638  0.67460383  1.00000000 -0.14441631
## Catholic          0.4136456   0.2886878 -0.47505753 -0.14441631  1.00000000
## Infant.Mortality  0.4371367  -0.1521287 -0.05915436 -0.01898137  0.06611714
##                  Infant.Mortality
## Fertility              0.43713670
## Agriculture           -0.15212866
## Examination           -0.05915436
## Education             -0.01898137
## Catholic               0.06611714
## Infant.Mortality       1.00000000
```

Se observan que existen correlaciones entre varias variables, aquellas correlaciones más significativas son las que se muestran en la siguiente tabla:

Variable 1	Variable 2	Correlación
Examination	Agriculture	-0.5989
Examination	Eduaction	0.6746
Examination	Catholic	-0.4751
Education	Agriculture	-0.6504

No se ha incluido las correlaciones con la variable Fertility ya que será nuestra variable independiente.

A continuación se realizará la prueba de hipótesis para cada correlación de la tabla anterior.

```
cor.test(Examination, Agriculture, methos='spearman')
```

```
##
##  Pearson's product-moment correlation
##
## data:  Examination and Agriculture
## t = -6.3341, df = 45, p-value = 9.952e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.8133545 -0.4974484
## sample estimates:
##        cor
## -0.6865422
```

S(45)=-6.3341, p<0.001, rs=-0.68654

Con un nivel de significación del 95% se estima que existe una correlación entre la variable Examination y Agricultura, esta correlación es negativa y fuerte.

```
cor.test(Examination, Education , methos='spearman')
```

```
##
##  Pearson's product-moment correlation
##
## data:  Examination and Education
## t = 6.5463, df = 45, p-value = 4.811e-08
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.5144218 0.8209342
## sample estimates:
##       cor
## 0.6984153
```

S(45) = 6.546, p<0.01, rs=0.6984

Existe correlación positiva fuerte entre Examination y Education.

```r
cor.test(Examination, Catholic, method='spearman')
```

```
## Warning in cor.test.default(Examination, Catholic, method = "spearman"): Cannot
## compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Examination and Catholic
## S = 25513, p-value = 0.0007403
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.4750575
```

$S(45) = 25513$, $p<0.01$, rs=-0.471

Existe correlación negativa y leve entre Examination y Catholic.

```r
cor.test(Education, Agriculture, method = 'spearman')
```

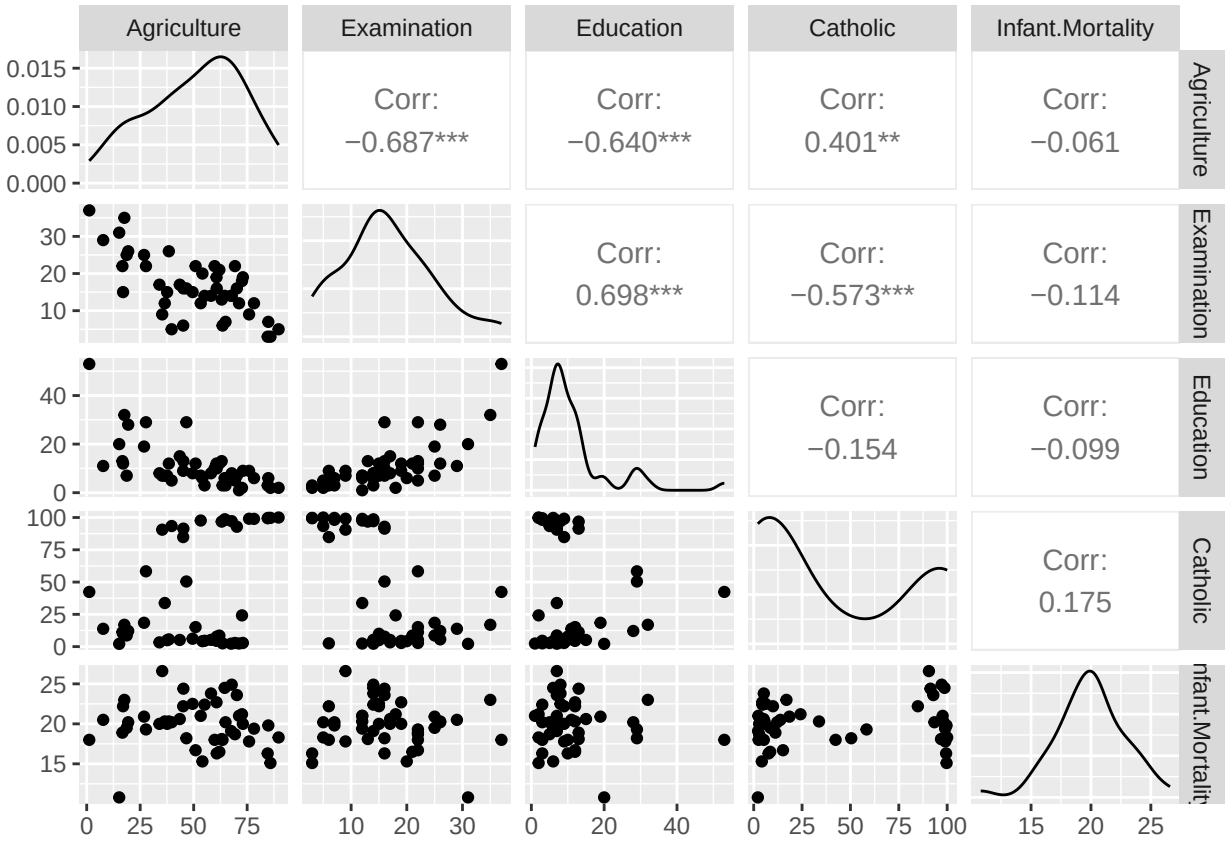
```
## Warning in cor.test.default(Education, Agriculture, method = "spearman"):
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  Education and Agriculture
## S = 28546, p-value = 7.457e-07
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##        rho
## -0.6504638
```

$S(45) = 28546$, $p<0.01$, rs=-0.6504

Existe correlación negativa fuerte entre Education y Agriculture.

```r
ggpairs(data[,-1], progress=F)
```



A continuación y a partir del gráfico anterior se analiza aquellos conjuntos de variables con una correlación superior a 0.5:

- Agriculture vs Examination: se observa una correlación negativa fuerte (corr: -0.687).
- Education vs Examination: se observa una correlación positiva fuerte (corr: 0.698).
- Education vs Catholic: se observa una correlación negativa débil (corr: -0.573).

2. Modelo lineal

Se crea un modelo sin iteraccion.

```
model <- lm(Fertility ~ Agriculture + Examination + Education + Catholic)
summary(model)
```

```
##
## Call:
## lm(formula = Fertility ~ Agriculture + Examination + Education +
##     Catholic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7813  -6.3308   0.8113   5.7205  15.5569
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 91.05542    6.94881  13.104  < 2e-16 ***
```

```
## Agriculture -0.22065    0.07360  -2.998  0.00455 **
## Examination -0.26058    0.27411  -0.951  0.34722
## Education   -0.96161    0.19455  -4.943 1.28e-05 ***
## Catholic     0.12442    0.03727   3.339  0.00177 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.736 on 42 degrees of freedom
## Multiple R-squared:  0.6498, Adjusted R-squared:  0.6164
## F-statistic: 19.48 on 4 and 42 DF,  p-value: 3.95e-09
```

2.1. Función matemática

La función matemática que define el modelo es:

$Fertility = \beta_0 + \beta_1 \cdot Agriculture + \beta_2 \cdot Examination + \beta_3 \cdot Education + \beta_4 \cdot Catholic + \beta_5 \cdot Infant.Mortality$

Y sustituyendo los predictores tenemos:

$Fertility = 66.915 - 0.172 \cdot Agriculture - 0.258 \cdot Examination - 0.871 \cdot Education + 0.104 \cdot Catholic + 1.077 \cdot Infant.Mortality$

2.2. Análisis Bondad de Ajuste

Prueba F global:

- F(5,41) = 19.76, p<0.001

Como p<0.05 se rechaza la hipótesis nula, por lo que al menos uno de los predictores está relacionado con la respuesta.

Error estándar residual:

RSE = 7.165, existe un error de 7.165 en la media estandarizada de fertilidad.

```
sigma(model)/mean(Fertility)*100
```

```
## [1] 11.02957
```

La tasa de error es del 10.22%.

Coeficiente de determinación:

R2 ajustado= 67.1%

El 67.1% de los datos pueden ser explicados por el modelo.

2.3. Coeficientes

Las variables que contribuyen al modelo son aquellos donde se rechaza la hipótesis nula (p<0.05):

- Intercepto (p<0.001)
- Education (p<0.001)
- Agriculture (p<0.01)
- Catholic (p<0.01)

La variable Examination (p>0.05) por lo que se acepta la hipótesis nula, es decir, esta variable no contribuye de manera significativa al modelo.

2.4. Generalización del modelo

Multicolinealidad

```
vif(model)
```

```
## Agriculture Examination   Education    Catholic
##    2.147410    3.675372    2.689400    1.856475
```

Ninguno de los 3 valores es mayor que 5, por lo que la multicolinealidad no es un problema.

Importancia de los predictores

```
crlm <- calc.relimp(model,
                    type=c("lmg"),
                    rela=T)
crlm
```

```
## Response variable: Fertility
## Total response variance: 156.0425
## Analysis based on 47 observations
##
## 4 Regressors:
## Agriculture Examination Education Catholic
## Proportion of variance explained by model: 64.98%
## Metrics are normalized to sum to 100% (rela=TRUE).
##
## Relative importance metrics:
##
##                    lmg
## Agriculture 0.1014550
## Examination 0.2780238
## Education   0.4319655
## Catholic    0.1885556
##
## Average coefficients for different model sizes:
##
##                      1X         2Xs         3Xs        4Xs
## Agriculture  0.1942017 -0.01698147 -0.15945704 -0.2206455
## Examination -1.0113173 -0.87947769 -0.64052222 -0.2605824
## Education   -0.8623503 -0.76351628 -0.83499925 -0.9616124
## Catholic     0.1388857  0.08922549  0.09787351  0.1244184
```

A continuación se muestra la importancia de cada variable:

Variable	Importancia(%)
Agriculture	10,14
Examintaion	27.8
Education	43.19
Catholic	18.85

Se observa que la variable más influyente es la educación, y la menos influyente agricultura.

Intervalos de confianza

```
confint(model)
```

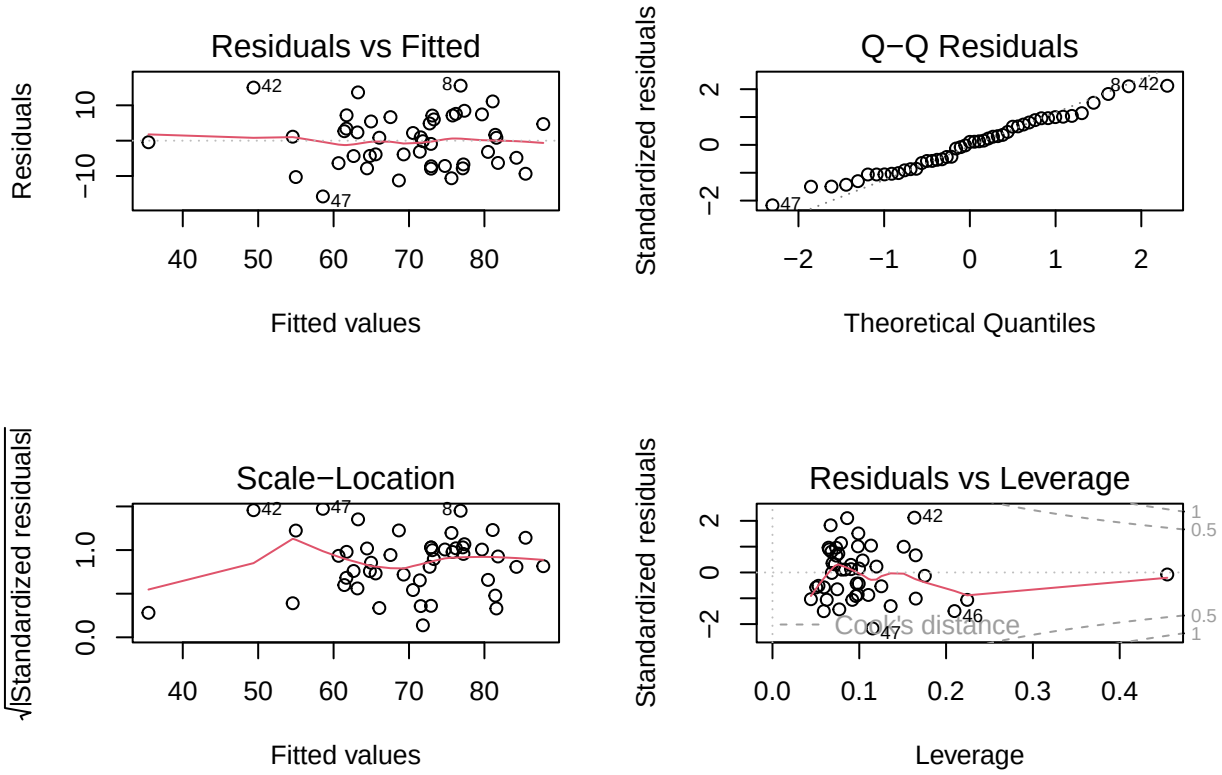
```
##                   2.5 %       97.5 %
## (Intercept) 77.03215094 105.07869687
## Agriculture -0.36917650  -0.07211452
## Examination -0.81375792   0.29259312
## Education   -1.35422113  -0.56900364
## Catholic     0.04921149   0.19962537
```

Valores de confianza:

Predictores	Tramo
Intercepto	[77.032 : 105.079]
Agriculture	[-0.369 : -0.0721]
Examination	[-0.814 : 0.293]
Education	[-1.354 : -0.569]
Catholic	[0.049 : 0.2]

Supuestos del modelo

```
par(mfrow = c(2,2))
plot(model)
```



Residuals vs fitted: se observa que los residuos no presentan tendencia, por lo que podríamos decir que **existe linealidad**.

Q-Q Residuals: las observaciones se encuentran a lo largo de la línea diagonal, por lo que podemos asumir que **existe el supuesto de normalidad**.

Scale-Location: se observa que se **cumple el supuesto de homocedasticidad**.

Residuals vs Leverage: no hay valores influyentes.

3. Predicciones

3.1. Predicciones con valores existentes

Seleccionamos de forma aleatoria valores existente en el dataframe

```
new_data <- data[c(3,8,21,26,42),]
```

Realizamos predicciones

```
predict_data <- predict(model, new_data[,-1])
predict_data
```

```
## Franches-Mnt        Glane       Morges      Payerne    Neuchatel
##     87.80550     76.84310     63.16259     67.54558     49.38524
```

Ahora vamos a ver los residuos entre los valores reales y predichos

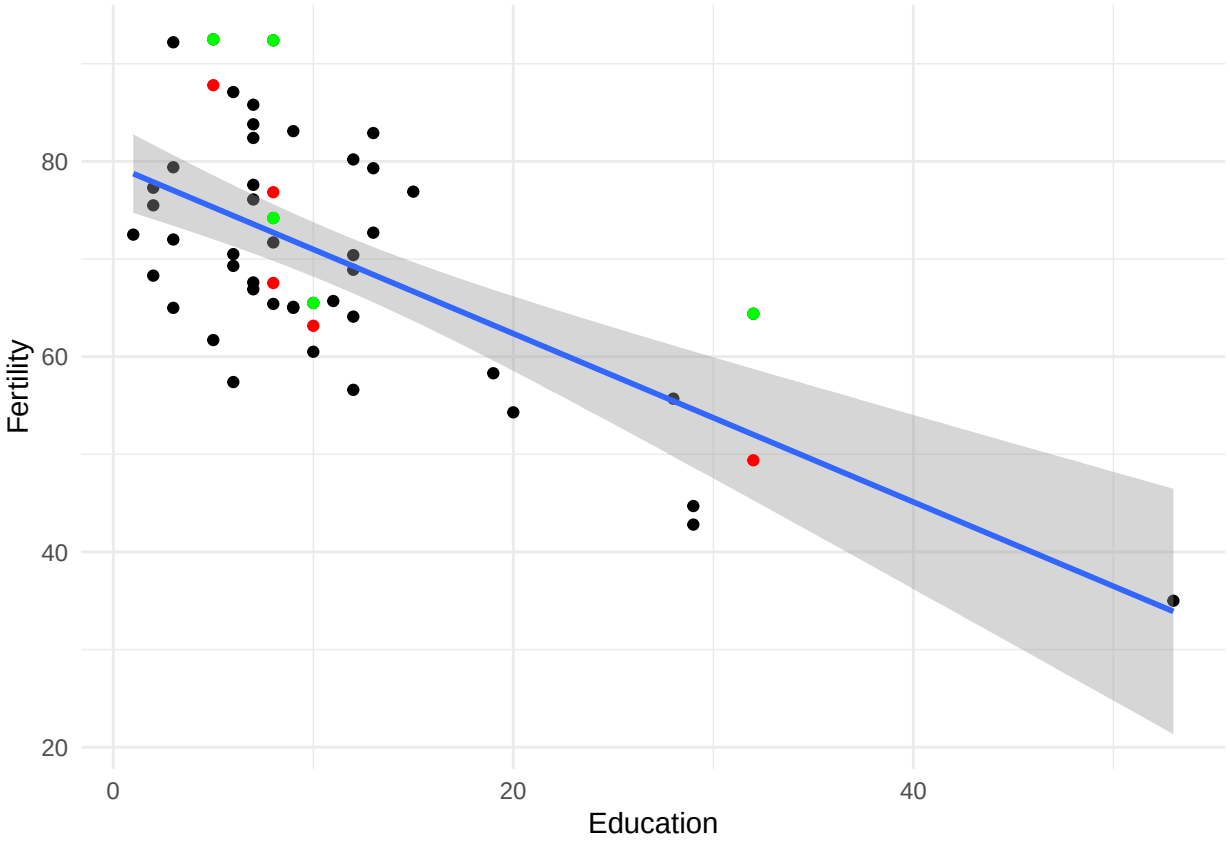
```
data[c(3,8,21,26,42), 1] - predict_data
```

```
## Franches-Mnt        Glane       Morges      Payerne    Neuchatel
##     4.694495    15.556900     2.337406     6.654425    15.014757
```

Se observa que los valores que más se alejan son Glane (15.56) y Neuchatel(15.01).

```
ggplot(data,aes(x=Education,y=Fertility))+
  geom_point()+
  geom_smooth(method="lm")+
  theme_minimal()+
  geom_point(data = new_data, mapping = aes(x=new_data$Education, y=new_data$Fertility), color='green')+
  geom_point(data = new_data, mapping = aes(x=new_data$Education, y=predict_data), color='red')
```

```
## `geom_smooth()` using formula = 'y ~ x'
```


Fertility
80
60
40
20
0
20
40
Education