[R] Liner-Regression

Priyank Thakkar

25/10/2021

Libraries

```
library(tidyverse)

## -- Attaching packages --------------------------------------- tidyverse 1.
3.1 --

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflict
s() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()

library(corrplot)

## corrplot 0.90 loaded

library(MASS)

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select

library(car)

## Loading required package: carData

##
## Attaching package: 'car'

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

## The following object is masked from 'package:purrr':
##
##     some
```

Dataset & Format: SWISS

Here, from previous project **"[R]-PCA"**, we have selected Swiss data set as our dataset to perform Linear Regression. This data set is about Standardized fertility measure and socio-economic indicators for each of **47 French-speaking** provinces of Switzerland at about **1888.** Each of which is in percent, i.e., in [0, 100].

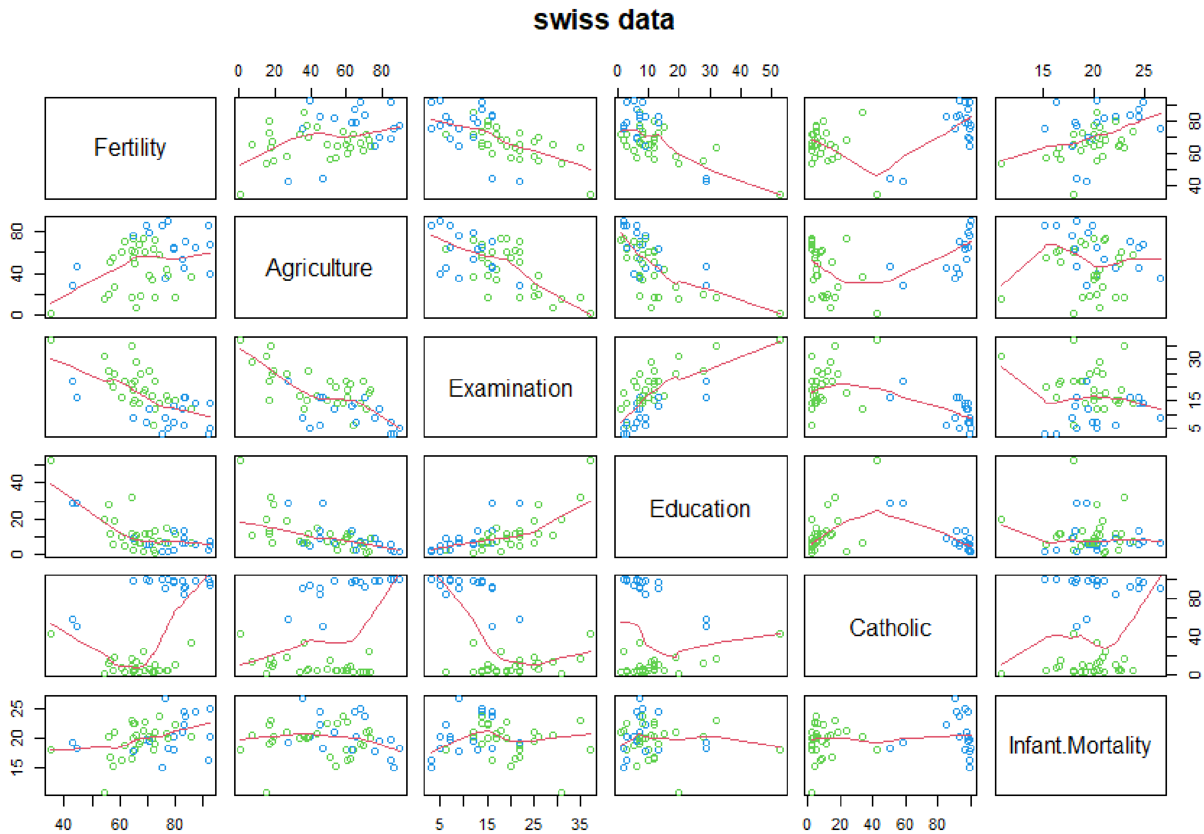
[,1] Fertility:	Ig, 'common standardized fertility measure'
[,2] Agriculture:	% of males involved in agriculture as occupation
[,3] Examination:	% draftees receiving highest mark on army examination
[,4] Education:	% education beyond primary school for draftees.
[,5] Catholic:	% 'catholic' (as opposed to 'protestant').
[,6] Infant.Mortality:	live births who live less than 1 year.

All variables but 'Fertility' give proportions of the population.

```
data(swiss)
summary(swiss)

##    Fertility      Agriculture     Examination      Education    
##  Min.   :35.00   Min.   : 1.20   Min.   : 3.00   Min.   : 1.00  
##  1st Qu.:64.70   1st Qu.:35.90   1st Qu.:12.00   1st Qu.: 6.00  
##  Median :70.40   Median :54.10   Median :16.00   Median : 8.00  
##  Mean   :70.14   Mean   :50.66   Mean   :16.49   Mean   :10.98  
##  3rd Qu.:78.45   3rd Qu.:67.65   3rd Qu.:22.00   3rd Qu.:12.00  
##  Max.   :92.50   Max.   :89.70   Max.   :37.00   Max.   :53.00  
##     Catholic       Infant.Mortality
##  Min.   :  2.150   Min.   :10.80   
##  1st Qu.:  5.195   1st Qu.:18.15   
##  Median : 15.140   Median :20.00   
##  Mean   : 41.144   Mean   :19.94   
##  3rd Qu.: 93.125   3rd Qu.:21.70   
```

Variable Distribution

```
pairs(swiss, panel = panel.smooth, main = "swiss data", col = 3 + (swiss$Cath
olic > 50))
```



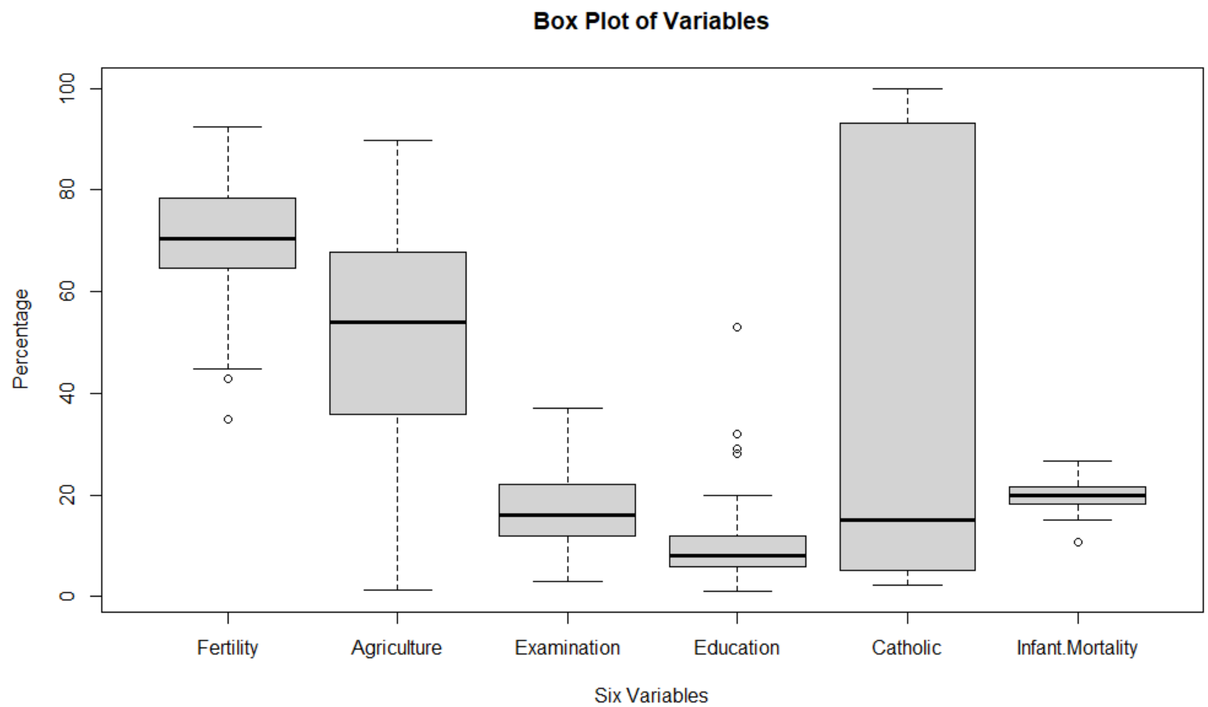
ANOVA

Analysis of Variance technique partitions the total variability in the sample data into component parts - variance explained by the regression line and the residual variance unexplained by the regression line.

ANOVA F-Test $F_0 = MS_R / MS_E$ Under null hypothesis F_0 has F-distribution with $(a-1)$ and $a(n-1)$ degrees of freedom

Lets perform ANOVA on our model from previous section

```
boxplot(swiss, xlab="Six Variables", ylab="Percentage", main="Box Plot of Var
iables")
```



Findings/Conclusion (Boxplot)

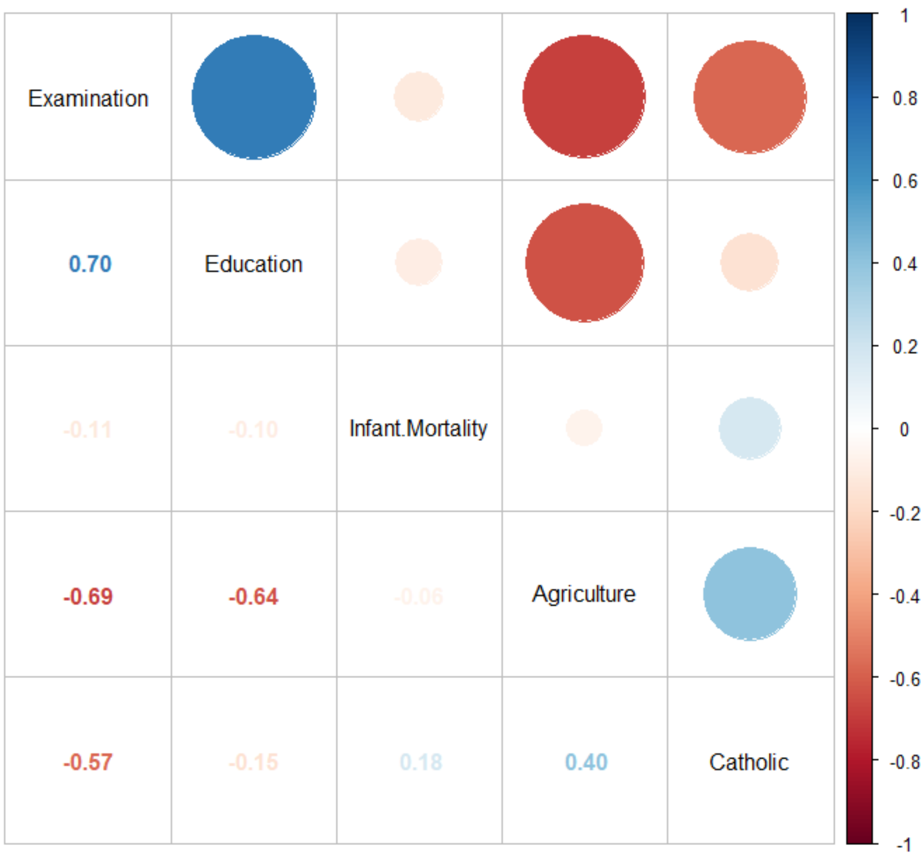
- Catholic variable covers wide range of values
- Infant.Mortality variable is very condensed
- Education and Fertility seems to have some outliers

```
fit = aov(Fertility~ ., data=swiss)
anova(fit)

## Analysis of Variance Table
##
## Response: Fertility
##                  Df  Sum Sq Mean Sq F value    Pr(>F)
## Agriculture       1  894.84  894.84 17.4288 0.0001515 ***
## Examination       1 2210.38 2210.38 43.0516 6.885e-08 ***
## Education         1  891.81  891.81 17.3699 0.0001549 ***
## Catholic          1  667.13  667.13 12.9937 0.0008387 ***
## Infant.Mortality  1  408.75  408.75  7.9612 0.0073357 **
## Residuals        41 2105.04   51.34
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

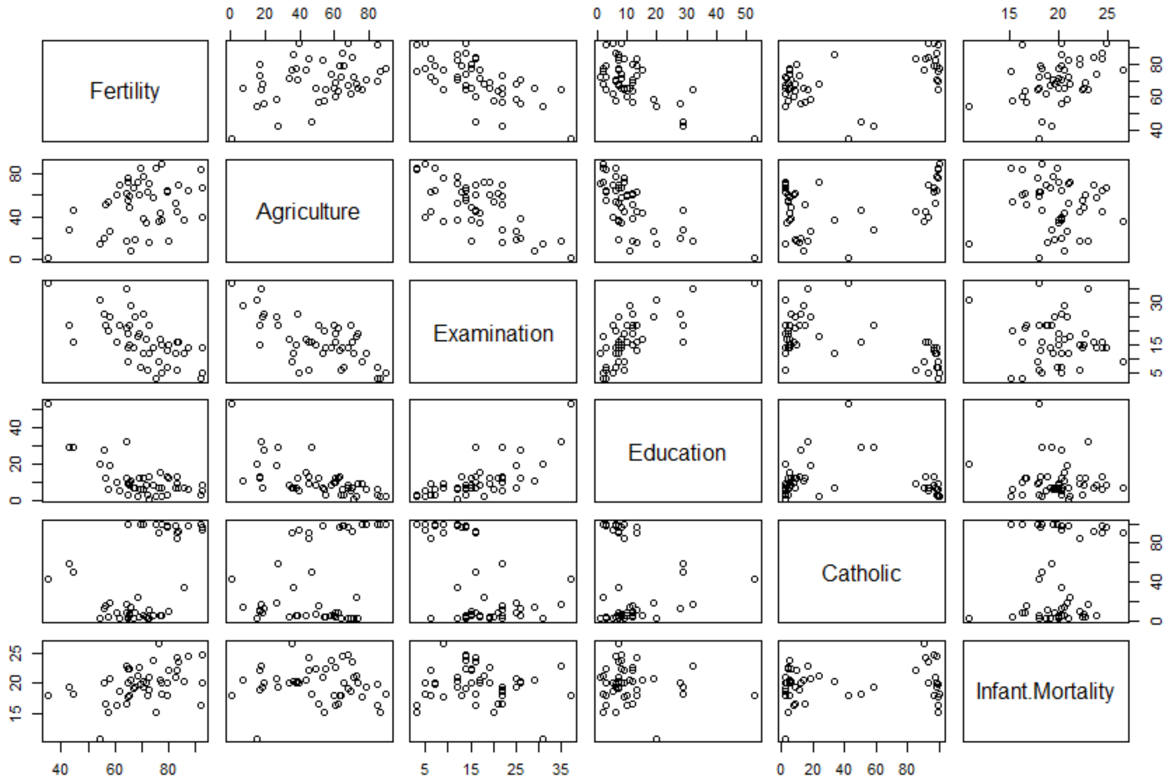
Correlation among Input Variables

Are the input variables correlated? Lets start with computing and visualizing the correlation matrix

```
cor(swiss)

##                   Fertility Agriculture Examination   Education   Catholic
## Fertility         1.0000000  0.35307918  -0.6458827 -0.66378886  0.4636847
## Agriculture       0.3530792  1.00000000  -0.6865422 -0.63952252  0.4010951
## Examination      -0.6458827 -0.68654221   1.0000000  0.69841530 -0.5727418
## Education        -0.6637889 -0.63952252   0.6984153  1.00000000 -0.1538589
## Catholic          0.4636847  0.40109505  -0.5727418 -0.15385892  1.0000000
## Infant.Mortality  0.4165560 -0.06085861  -0.1140216 -0.09932185  0.1754959
##                  Infant.Mortality
## Fertility              0.41655603
## Agriculture           -0.06085861
## Examination           -0.11402160
## Education             -0.09932185
## Catholic               0.17549591
## Infant.Mortality       1.00000000

corrplot.mixed(cor(swiss[,-1]), order="hclust", tl.col="black")
```



```
pairs(swiss)
```



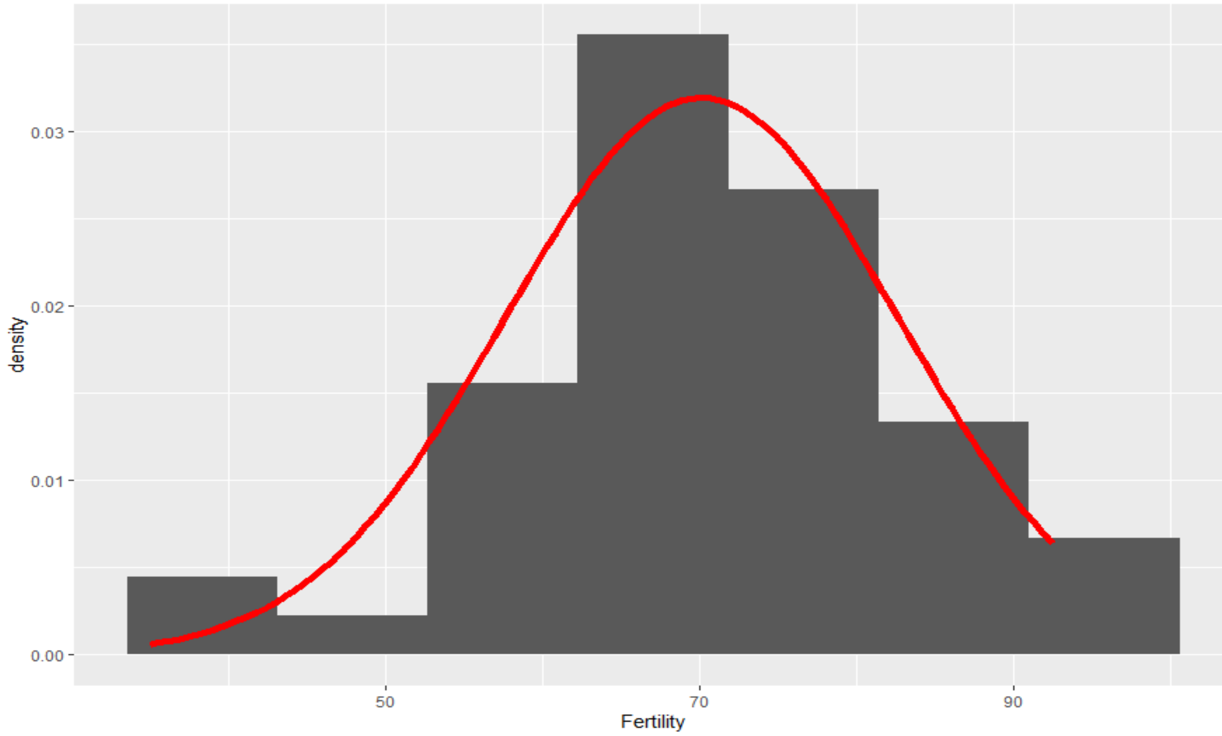
Findings/Conclusion (Correlation)

- All correlations with Fertility are less than **0.7**, indicating no signs of strong multicollinearity.
- Correlations are between **0.3-0.7**, indicating mild multicollinearity.
- Plot shows linear relationship between **Agriculture and Examination**. Moreover, also between **Examination and Education**.

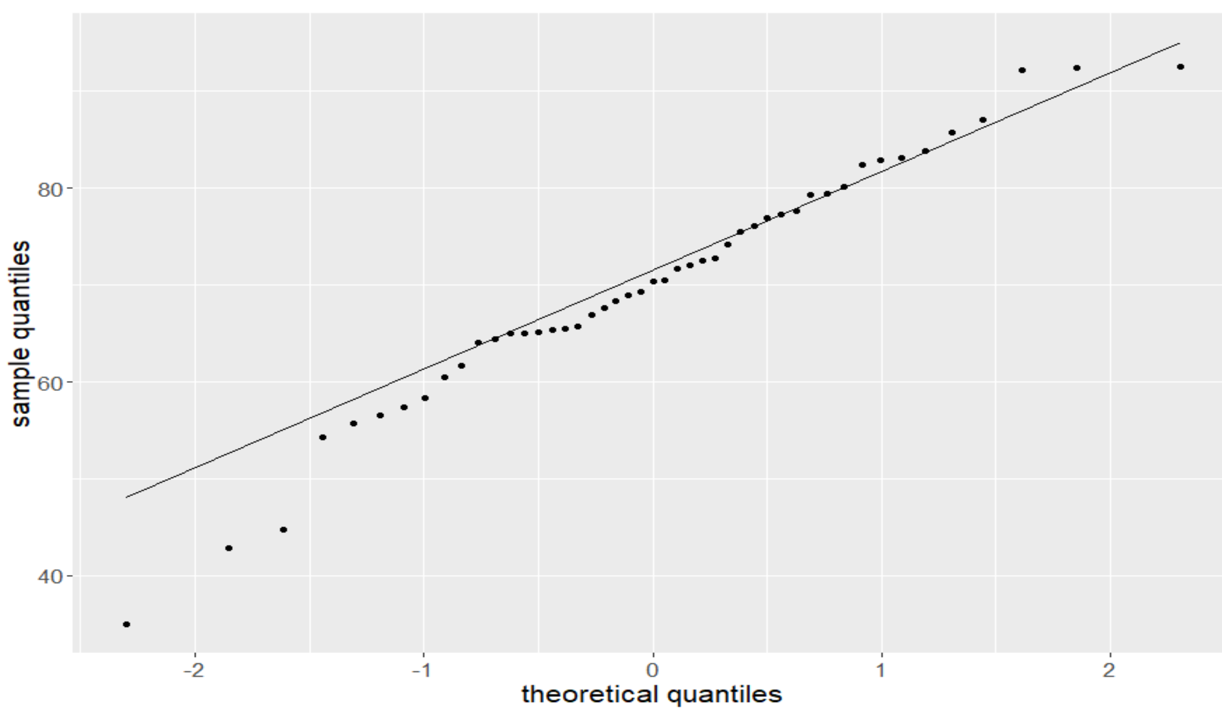
Regression

Building empirical models

```
ggplot(data = swiss, aes(Fertility)) +
  geom_histogram(mapping = aes(x = Fertility, y = stat(density)), bins = 7) +
   stat_function(
    fun = dnorm,
    args = list(mean = mean(swiss$Fertility), sd = sd(swiss$Fertility)),
    lwd = 2,
    col = 'red')
```




```r
Fertility <- ggplot(data = swiss, aes(sample = Fertility))

Fertility +
  stat_qq(distribution = stats::qnorm) + stat_qq_line() +
  labs(y = 'sample quantiles', x = 'theoretical quantiles') +
  theme(text = element_text(size = 16))
```



Findings/Conclusion (Histogram)

- In Histogram Plot is skewed or leaned towards the right side a little bit.
- Fertility rates are mostly between 60-90%

Fit a Model

```
model1 <- lm(Fertility ~ ., swiss)
summary(model1)

## 
## Call:
## lm(formula = Fertility ~ ., data = swiss)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -15.2743  -5.2617   0.5032   4.1198  15.3213 
## 
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
## Agriculture      -0.17211    0.07030  -2.448  0.01873 *  
## Examination      -0.25801    0.25388  -1.016  0.31546    
## Education        -0.87094    0.18303  -4.758 2.43e-05 ***
## Catholic          0.10412    0.03526   2.953  0.00519 ** 
## Infant.Mortality  1.07705    0.38172   2.822  0.00734 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 7.165 on 41 degrees of freedom
## Multiple R-squared:  0.7067, Adjusted R-squared:  0.671 
## F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

Residual Analysis

```
# stepAIC(model1, direction = "both")
# link of reference : https://ashutoshtr.medium.com/what-is-stepaic-in-r-a65b
71c9eeba

model2 <- step(model1)

## Start:  AIC=190.69
## Fertility ~ Agriculture + Examination + Education + Catholic + 
##     Infant.Mortality
## 
##                    Df Sum of Sq    RSS    AIC
## - Examination       1     53.03 2158.1 189.86
## <none>                          2105.0 190.69
```

```
## - Agriculture       1    307.72 2412.8 195.10
## - Infant.Mortality  1    408.75 2513.8 197.03
## - Catholic          1    447.71 2552.8 197.75
## - Education         1   1162.56 3267.6 209.36
##
## Step:  AIC=189.86
## Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##
##                    Df Sum of Sq    RSS    AIC
## <none>                          2158.1 189.86
## - Agriculture       1    264.18 2422.2 193.29
## - Infant.Mortality  1    409.81 2567.9 196.03
## - Catholic          1    956.57 3114.6 205.10
## - Education         1   2249.97 4408.0 221.43
```

Findings/Conclusion (Residual Analysis)

- The final AIC of model Achieved = 189.86
- And select the model with having the lowest AIC number, means Small lose of data while performing that model.

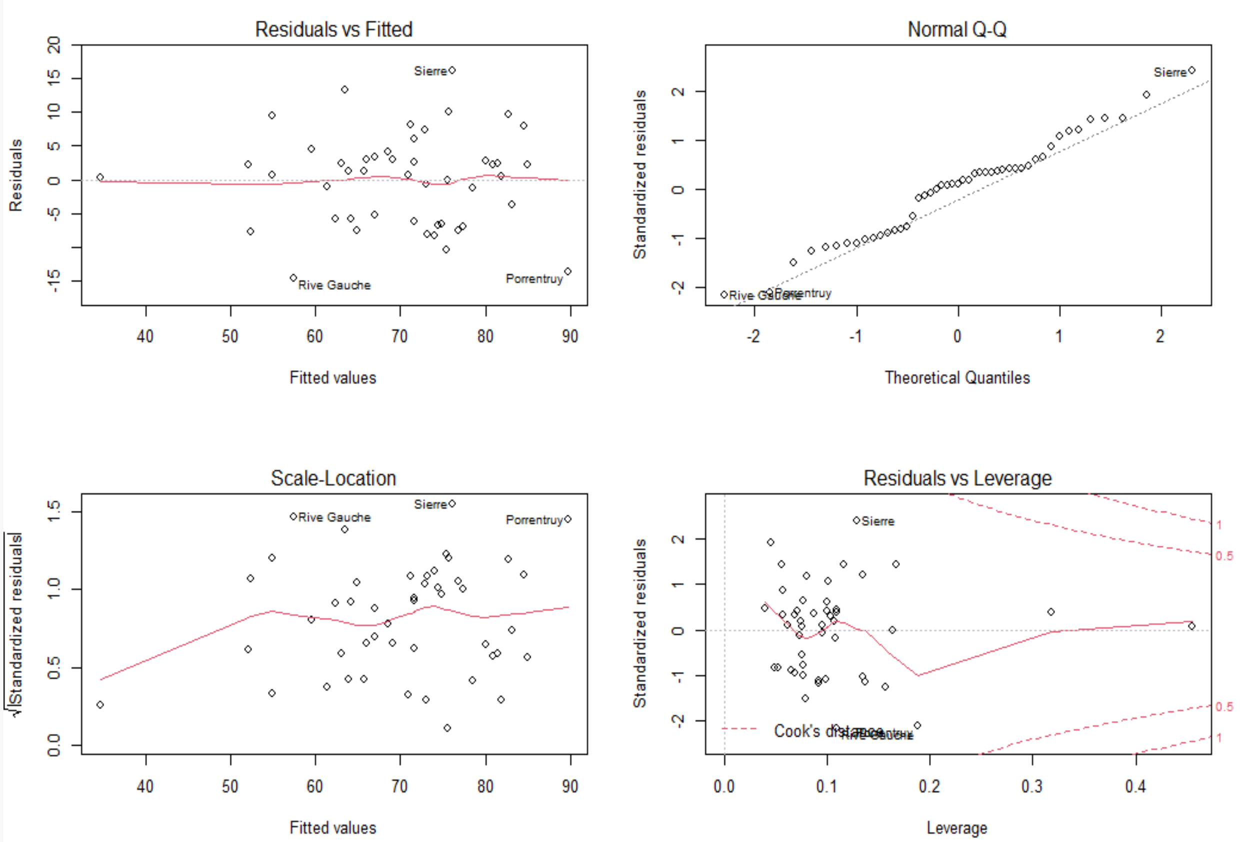
```
summary(model2)

##
## Call:
## lm(formula = Fertility ~ Agriculture + Education + Catholic +
##     Infant.Mortality, data = swiss)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.6765  -6.0522   0.7514   3.1664  16.1422
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      62.10131    9.60489   6.466 8.49e-08 ***
## Agriculture      -0.15462    0.06819  -2.267  0.02857 *
## Education        -0.98026    0.14814  -6.617 5.14e-08 ***
## Catholic          0.12467    0.02889   4.315 9.50e-05 ***
## Infant.Mortality  1.07844    0.38187   2.824  0.00722 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.168 on 42 degrees of freedom
## Multiple R-squared:  0.6993, Adjusted R-squared:  0.6707
## F-statistic: 24.42 on 4 and 42 DF,  p-value: 1.717e-10

anova(model1, model2)

## Analysis of Variance Table
##
## Model 1: Fertility ~ Agriculture + Examination + Education + Catholic +
```

```
##     Infant.Mortality
## Model 2: Fertility ~ Agriculture + Education + Catholic + Infant.Mortality
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     41 2105.0
## 2     42 2158.1 -1   -53.027 1.0328 0.3155

# drop1(model1,test='F')
```

Findings/Conclusion (Model1 and Model2)

- By both a T-test and an ANOVA F test we find **Examination** does not have significant effect on **Fertility**.
- Final Model2 ($Fertility\ (Agriculture + Education + Catholic + Infant.Mortality), data = swiss$)

```
par(mfrow=c(2,2))
plot(model2)
```



Other Diagnostic functions

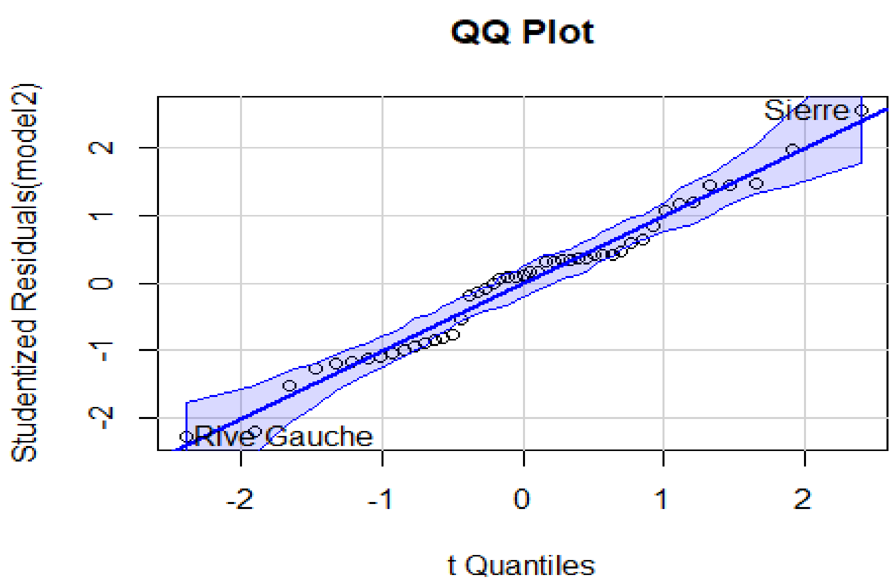
Other functions you can try:

- coefficients(model2) # model coefficients
- confint(model2, level=0.95) # CIs for model parameters
- fitted(model2) # predicted values
- residuals(model2) # residuals
- anova(model2) # anova table
- vcov(model2) # covariance matrix for model parameters
- influence(model2) # regression diagnostics

```
confint(model2, level=0.95)

##                        2.5 %      97.5 %
## (Intercept)      42.71786664 81.48475647
## Agriculture      -0.29223032 -0.01700466
## Education        -1.27921575 -0.68131191
## Catholic          0.06635694  0.18297584
## Infant.Mortality  0.30780496  1.84907938
```

Additional Regression Diagnostics

```
# Assessing Outliers
outlierTest(model2) # Bonferonni p-value for most extreme obs

## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##        rstudent unadjusted p-value Bonferroni p
## Sierre 2.570011           0.013901      0.65335

qqPlot(model2, main="QQ Plot") #qq plot for studentized resid
```



```
##       Sierre Rive Gauche
##           37          47

leveragePlots(model2) # leverage plots
```