Exploratory data analysis of Swiss Data

Aritra Biswas

6 April 2016

Data Set:

Swiss Fertility and Socioeconomic Indicators (1888) Data

Description:

Standardized fertility measure and socio-economic indicators for each of 47 French-speaking provinces of Switzerland at about 1888.

Format:

A data frame with 47 observations on 6 variables, each of which is in percent, i.e., in [0, 100].

Fertility : Ig, 'common standardized fertility measure'

Agriculture : % of males involved in agriculture as occupation

Examination : % draftees receiving highest mark on army examination

Education : % education beyond primary school for draftees.

Catholic : % 'catholic' (as opposed to 'protestant').

Infant.Mortality : live births who live less than 1 year.

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

Source:

Project "16P5"", pages 549-551 in

Mosteller, F. and Tukey, J. W. (1977) Data Analysis and Regression: A Second Course in Statistics. Addison-Wesley, Reading Mass.

indicating their source as "Data used by permission of Franice van de Walle. Office of Population Research, Princeton University, 1976. Unpublished data assembled under NICHD contract number No 1-HD-O-2077.""

Here we show the data type of each variable by using **str()** function from base package in R. The variable with int data type are of the format of discrete and others are continuous.

```
## '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 ...
```

A shortnote on Boxplot:

The box plot presents five sample statistics:

1) the minimum,
2) the lower quartile,
3) the median,
4) the upper quartile
5) the maximum

The length of the box is thus the interquartile range of the sample. A line is drawn across the box at the sample median. Whiskers sprout from the two ends of the box until they reach the sample maximum and minimum. The box length gives an indication of the sample variability and the line across the box shows where the sample is centered. The position of the box in its whiskers and the position of the line in the box also tells us whether the sample is symmetric or skewed, either to the right or left. For a symmetric distribution, long whiskers, relative to the box length, can betray a heavy tailed population and short whiskers, a short tailed population. So, provided the number of points in the sample is not too small, the box plot also gives us some idea of the "shape" of the sample, and by implication, the shape of the population from which it was drawn.

Application of boxplot:

a) Indicator of Centrality
b) Indicator of Spread
c) Indicator of Symmetry
d) Indicator of Tail Length
e) Outlier

One definition of outlier is any data point more than 1.5 interquartile ranges (IQRs) below the first quartile or above the third quartile.

Residual: The difference between the predicted value (based on the regression equation) and the actual, observed value.

Outlier: In linear regression, an outlier is an observation with large residual. In other words, it is an observation whose dependent-variable value is unusual given its value on the predictor variables. An outlier may indicate a sample peculiarity or may indicate a data entry error or other problem.

Leverage: An observation with an extreme value on a predictor variable is a point with high leverage. Leverage is a measure of how far an independent variable deviates from its mean. High leverage points can have a great amount of effect on the estimate of regression coefficients.

Influence: An observation is said to be influential if removing the observation substantially changes the estimate of the regression coefficients. Influence can be thought of as the product of leverage and outlierness.

Cook's distance (or Cook's D): A measure that combines the information of leverage and residual of the observation.

Histogram with a Normal Distribution Fit:

A histogram dissects the range of the variable into equal-width class intervals called bins and then plots the number of observations falling in each bin as a bar chart (i.e. the height of the bar represents the number, proportion or percentage of observations in that class).

The normal distribution is a commonly used distribution for continuous variables with many convenient properties, so let's try to fit the normal distribution to this data and examine if it is consistent with the histogram.

Stem and leaf plot: A Stem and Leaf Plot is a special table where each data value is split into a "stem" (the first digit or digits) and a "leaf" (usually the last digit). A stem and leaf plot generally helps in studying that where the majority of the values of the variable lie , like around what range of values in the x-axis. They are also useful for highlighting outliers and finding the mode. However, stem-and-leaf displays are only useful for moderately sized data sets (around 15-150 data points).

QQ plot: Normal q-q plot gives the idea of the relationship between the theoretical quantiles of the data and sample quantiles of the data.

Test for population correlation coefficients:

In cases such as these, we answer our research question concerning the existence of a linear relationship by using the t-test for testing the population correlation coefficient $H_0 : \rho = 0$.

Null hypothesis: $H_0 : \rho = 0$ **Alternative hypothesis:** $H_A : \rho \neq 0$ or $H_A : \rho < 0$ or $H_A : \rho > 0$

Test Statistic:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

As always, the P-value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if the null hypothesis were true?" The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.

Finally, we make a decision:

If the P-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor and response."

If the P-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor and response."

Shapiro-Wilk Test: (Parametric Test)

H_0 : The samples come from a parent population with Normal distribution.

H_1 : The samples do not come from a parent population with Normal distribution.

The Shapiro-Wilk test is a test of normality. Null hypothesis checks whether the sample came from a normally distributed population. The test statistic is:

$$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

where $x_{(i)}$ (with parentheses enclosing the subscript index i) is the ith order statistic, i.e., the ith-smallest number in the sample;

$\bar{x} = (x_1 + \cdots + x_n)/n$ is the sample mean;

the constants a_i are given by $(a_1, \ldots, a_n) = \frac{m^{\mathsf{T}} V^{-1}}{(m^{\mathsf{T}} V^{-1} V^{-1} m)^{1/2}}$ where $m = (m_1, \ldots, m_n)^{\mathsf{T}}$ and $m_1, \ldots, m_n$ are the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution, and V is the co variance matrix of those order statistics. The user may reject the null hypothesis if W is below a predetermined threshold.

For the Shapiro-Wilk statistic:

- If p is more than .05 $\{\alpha\}$, we can be 95% $\{(1-\alpha)100\%\}$ certain that the data are normally distributed. (In other words, we fail to reject the null hypothesis.)
- If p is less than .05 $\{\alpha\}$, we can be 95% $\{(1-\alpha)100\%\}$ certain that the data are not normally distributed. (In other words, we reject the null hypothesis.)

Kolmogorov-Smirnov test (nonparametric test):

The Kolmogorov-Smirnov test is defined by:

H_0**:** The data follow a specified distribution. (Normal in this case)

H_1 : The data do not follow the specified distribution. (Normal in this case)

Test Statistic: The Kolmogorov-Smirnov test statistic is defined as

$$D_n = \sup_x |F_n(x) - F(x)|)$$

where F is the theoretical cumulative distribution of the distribution being tested which must be a continuous distribution (i.e., no discrete distributions such as the binomial or Poisson), and it must be fully specified (i.e., the location, scale, and shape parameters cannot be estimated from the data).

Significance Level: α

Critical Values: The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value.

Test for population correlation coefficients:

In cases such as these, we answer our research question concerning the existence of a linear relationship by using the t-test for testing the population correlation coefficient $H_0 : \rho = 0$.

Null hypothesis: $H_0 : \rho = 0$

Alternative hypothesis: $H_A : \rho \neq 0$ or $H_A : \rho < 0$ or $H_A : \rho > 0$

Test Statistic:

$$t^* = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

As always, the P-value is the answer to the question "how likely is it that we'd get a test statistic t^* as extreme as we did if the null hypothesis were true?" The P-value is determined by referring to a t-distribution with n-2 degrees of freedom.

Finally, we make a decision:

If the P-value is smaller than the significance level α, we reject the null hypothesis in favor of the alternative. We conclude "there is sufficient evidence at the α level to conclude that there is a linear relationship in the population between the predictor and response."

If the P-value is larger than the significance level α, we fail to reject the null hypothesis. We conclude "there is not enough evidence at the α level to conclude that there is a linear relationship in the population between the predictor and response."

Scatter Diagrams and Regression Lines

Scatter Diagrams

If data is given in pairs then the scatter diagram of the data is just the points plotted on the xy-plane. The scatter plot is used to visually identify relationships between the first and the second entries of paired data.

If the points follow a linear pattern , then we say that there is a high linear correlation, while if the points do not follow a linear pattern, it is said to be there is no linear correlation. If the data somewhat follow a linear path, then we say that there is a moderate linear correlation.

Things which can be explored from scatter plot are:

a) Direction

b) Form

c) Strength

d) Outliers and Influntial Points

For variable Fertility:

```r
attach(swiss)
print(paste("Mean of Fertility:",paste(mean(Fertility), collapse=" ")))
print(paste("SD of Fertility:",paste(sd(Fertility), collapse=" ")))
print(paste("Median of Fertility:",paste(median(Fertility), collapse=" ")))
print(paste("IQR of Fertility:",paste(IQR(Fertility), collapse=" ")))
print(paste("Maximum of Fertility:",paste(max(Fertility), collapse=" ")))
print(paste("Minimum of Fertility:",paste(min(Fertility), collapse=" ")))
print(paste("Outliers in Fertility:",paste(boxplot.stats(Fertility)$out, collapse=" ")))
```

Descriptive Statistics

- Mean of Fertility: 70.1
- SD of Fertility: 12.49
- Median of Fertility: 70.4
- IQR of Fertility: 13.7
- Maximum of Fertility: 92.5
- Minimum of Fertility: 35
- Outliers in Fertility: 35, 42.8

Stem-and-Leaf plot of Fertility

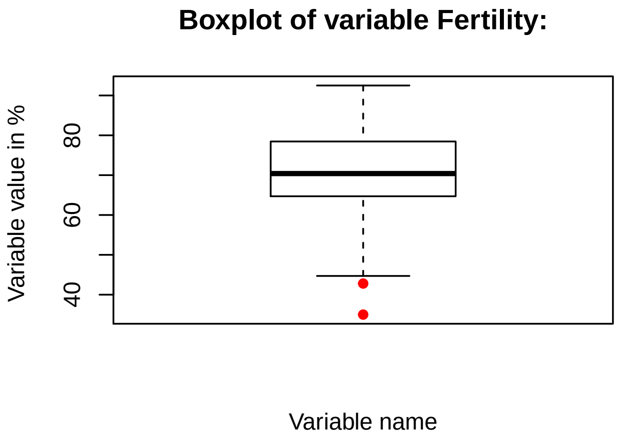
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   3 | 5
##   4 | 35
##   5 | 46778
##   6 | 124455556678899
##   7 | 01223346677899
##   8 | 0233467
##   9 | 223
```

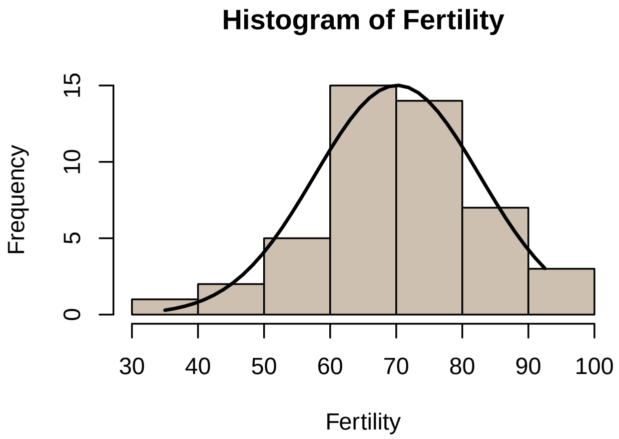
A stem and leaf plot generally helps in studying that where the majority of the values of the variable lie , like around what range of values in the x-axis.

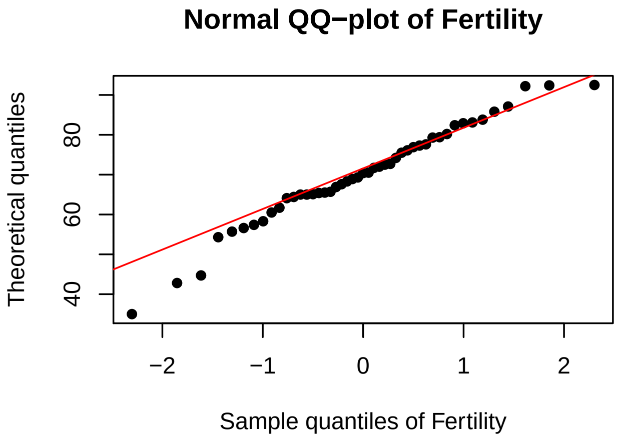
Thus we can observe that large amount of the values lies in the 50-60 range . Also we can then infer that within the range 30-100 the maximum values lies almost in the middle part and hence the data comes from the normal population.

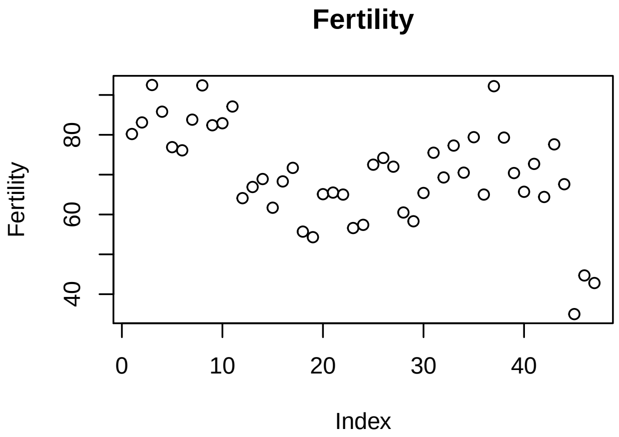
Visualization:

```r
par(mfrow=c(2,2))
#Locading dataset in R
attach(swiss)

#Plotting variable Fertility in Boxplot
boxplot(Fertility,
        col = "white",
        main="Boxplot of variable Fertility:",
        ylab ="Variable value in % ",
        xlab="Variable name",
        outcol="red",
        outpch=19)

#Histogram of Fertility variable from swiss dataset with fitted normal density curve

h1<-hist(swiss$Fertility,col="antiquewhite3",main="Histogram of Fertility",xlab="Fertility");
xfit<-seq(min(swiss$Fertility),max(swiss$Fertility),length=40)
yfit<-dnorm(xfit,mean=mean(swiss$Fertility),sd=sd(swiss$Fertility))
yfit <- yfit*diff(h1$mids[1:2])*length(swiss$Fertility)
lines(xfit, yfit, col="black", lwd=2)


#QQ plot of the variable Fertility
qqnorm(swiss$Fertility,
       pch=16,
       main="Normal QQ-plot of Fertility",
       xlab="Sample quantiles of Fertility",
       ylab="Theoretical quantiles");
qqline(swiss$Fertility, col = 2);


#Plotting of the variable Fertility with respect to its index.
plot(Fertility, main="Fertility")
```









Histogram: The histogram with the density curve of Fertility clearly shows that maximum frequency of the values lie slightly towards right and thus the variable is nearly skewed and so the data is from normal population. Also the bars are not much outside the density curve.

Boxplot: From this boxplot we can interpret that the variable Fertility is slightly left skewed which we can comment as nearly symmetric, the values of variables are not that dispersed because the length of the two ends of the whiskers is almost equal and the median also lie almost in the centre . And also two of the values are extremely small and thus lie outside of the plot which are denoted by the red dots in the plot.

QQ Plot: The above QQ plot clearly shows that most of the values lies above the normal line but more or less close to it. So we can interpret that the data is surely from a normal distribution.

Outliers: From the plot of values with index, this can be shown that there are some values away from the data cloud. The boxplot confirms that the values are outlier.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

```r
ks.test(swiss$Fertility,"pnorm",mean(swiss$Fertility),sd(swiss$Fertility))
```

D	p-value
0.10	0.72

- The observed value of the Kolmogorov-Smirnov test statistic is: D = 0.10
- The exact probability of the observed value, D = 0.10, p-value = 0.72
- For the Fertility, p-value = 0.72, which is greater than .05.

Shapiro-Wilk normality test(Parametric test):

```r
shapiro.test(swiss$Fertility)
```

W	p-value
0.97	0.34

- The observed value of the Shapiro-Wilk statistic is: W = 0.97307
- The exact probability of the observed value, W = 0.97307, p-value = 0.3449
- For the Fertility, p = 0.3449, which is greater than .05.

The parent population is normally distributed.

For variable Agriculture:

Descriptive Statistics

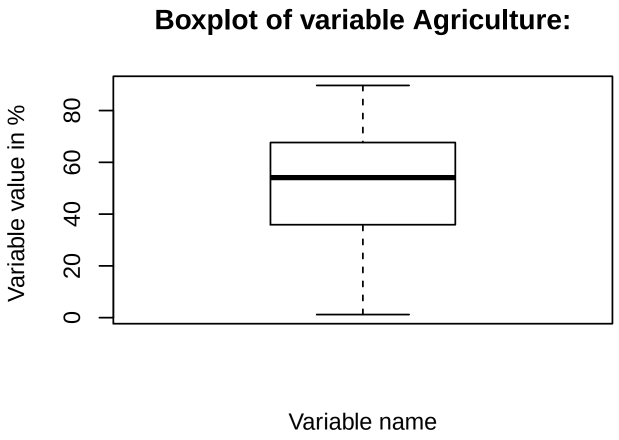
- Mean of Agriculture: 50.65
- SD of Fertility: 22.71
- Median of Agriculture: 54.1
- IQR of Agriculture: 31.75
- Maximum of Agriculture: 89.7
- Minimum of Agriculture: 1.2

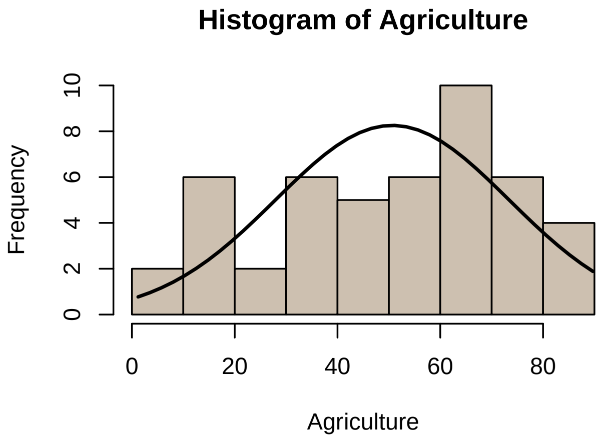
Stem-and-Leaf plot of Agriculture

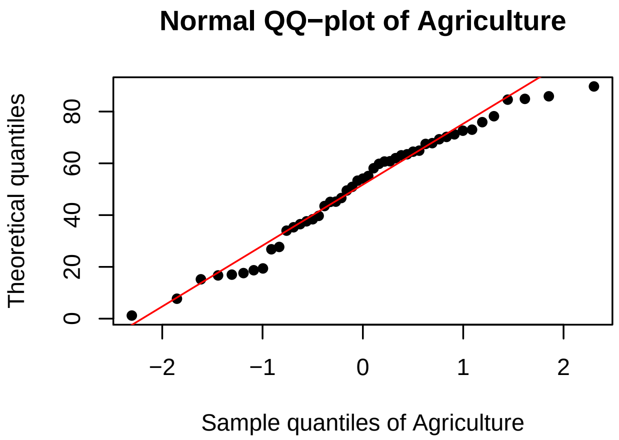
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 18
##   1 | 577899
##   2 | 78
##   3 | 45788
##   4 | 04557
##   5 | 013458
##   6 | 01123455889
##   7 | 013368
##   8 | 556
##   9 | 0
```

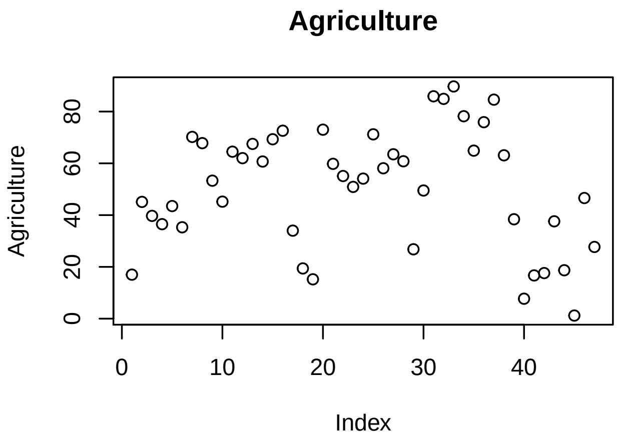
Thus we can observe that large amount of the values lies in the 60 -70 range . Also we can then infer that within the range 0-100 the maximum values lie towards the right but still since the data is highly random so we can infer that data comes from the normal population.

Visualization:









Histogram: The histogram with the density curve clearly shows that maximum frequency of the values lie far behind the density curve but the values follow a pattern like the normal variables i.e.first increasing and reaching the highest value and then descending and so the data is from normal population.

Boxplot: From this box plot we can interpret that the variable Agriculture is slightly left–skewed which we can comment as nearly symmetric, the values of variables are not that dispersed because the length of the two ends of the whiskers is almost equal and the median also lie almost in the center . And the data of Agriculture does not contain any outlier.

QQ Plot: The above plot clearly shows that most of the values lies along the normal line with two or three values away from the line . So we can interpret that the data is surely from a normal distribution.

Outliers: There is not outlier in the variable under consideration.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

D	p-value
0.10	0.66

- The observed value of the Kolmogorov-Smirnov test statistic is: D = 0.10314
- The exact probability of the observed value, D = 0.10314, p-value = 0.6613
- For the Agriculture, p-value = 0.6613, which is greater than .05.

Shapiro-Wilk normality test(Parametric test):

W	p-value
0.97	0.19

- The observed value of the Shapiro-Wilk statistic is: W = 0.96643
- The exact probability of the observed value, W = 0.96643, p-value = 0.193
- For the Agriculture, p = 0.193, which is greater than .05.

The parent population is normally distributed.

For variable Examination:

Descriptive Statistics

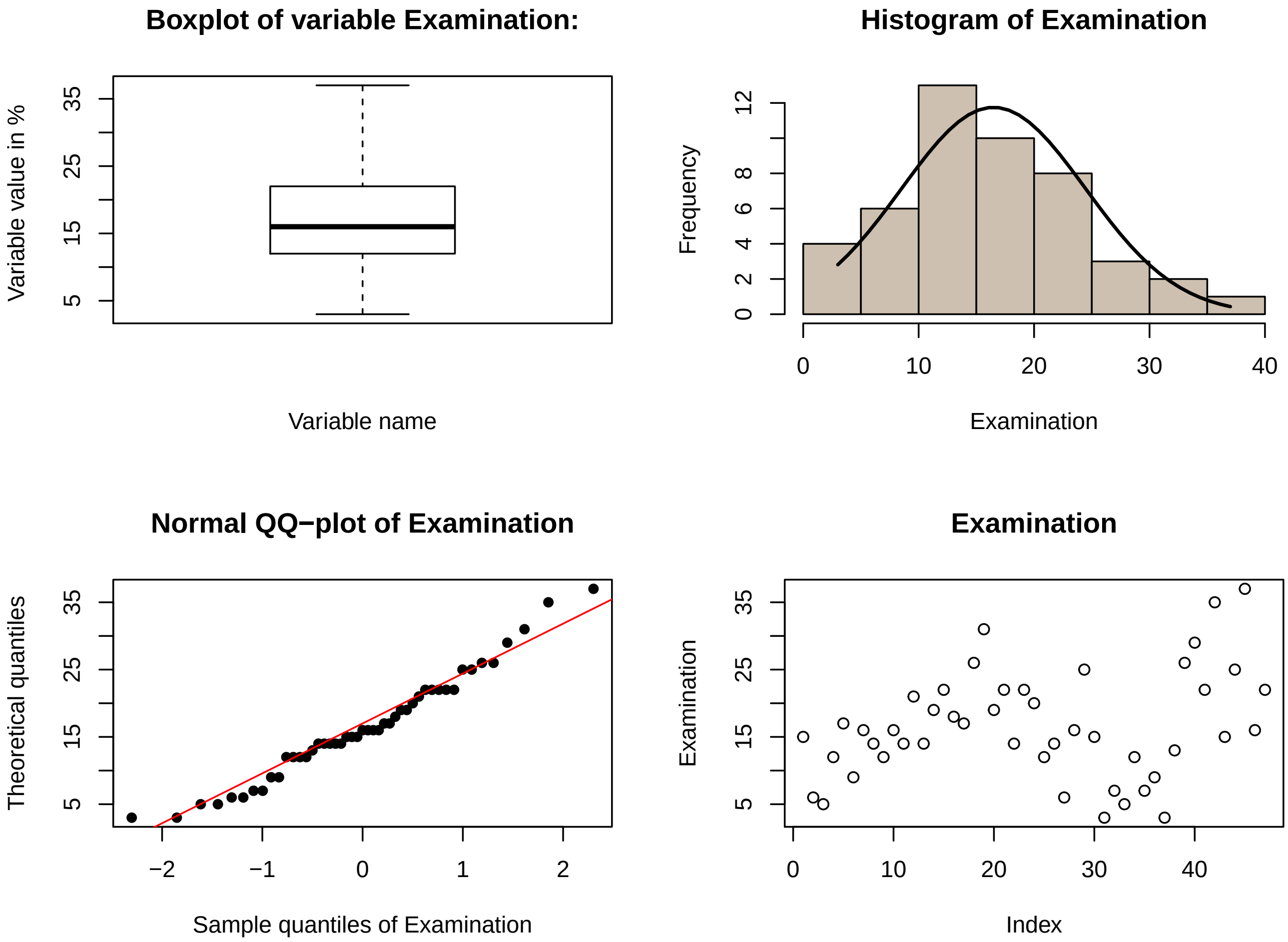
- Mean of Examination: 16.48
- Median of Examination: 16
- SD of Examination: 7.97
- IQR of Examination: 10
- Maximum of Examination: 37
- Minimum of Examination: 3

Stem-and-Leaf plot of Examination

```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 33
##   0 | 55667799
##   1 | 2222344444
##   1 | 555666677899
##   2 | 0122222
##   2 | 55669
##   3 | 1
##   3 | 57
```

We can observe that large amount of the values lies in the 10-20 range and the data is highly random so we can infer that data comes from the normal population.

Visualization:



Histogram: The histogram with the density curve clearly shows that maximum frequency of the values lie behind the density curve towards left but the values follow a pattern like the normal variables i.e.first increasing and reaching the highest value and then descending and so the data is from normal population.

Boxplot: From the box plot we can interpret that the variable Agriculture is slightly right–skewed which we can comment as nearly symmetric, the values of variables are not that dispersed because the length of the two ends of the whiskers is almost equal and the median also lie almost in the center . And the data of Examination do not contain any outlier.

QQ Plot: The plot clearly shows that most of the values lies along the normal line with two or three values away from the line . So we can interpret that the data is surely from a normal distribution.

Outliers: There is not outlier in the variable under consideration.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

D	p-value
0.10	0.75

- The observed value of the Shapiro-Wilk statistic is: D = 0.098924
- The exact probability of the observed value, D = 0.098924, p-value = 0.7472
- For the Examination, p-value = 0.7472, which is greater than .05.

Shapiro-Wilk normality test(Parametric test):

W	p-value
0.96	0.25

- The observed value of the Shapiro-Wilk statistic is: W = 0.96962
- The exact probability of the observed value, W = 0.96962, p-value = 0.2563
- For the Examination, p = 0.2563, which is greater than .05.

The parent population is normally distributed.

For variable Education:

Descriptive Statistics

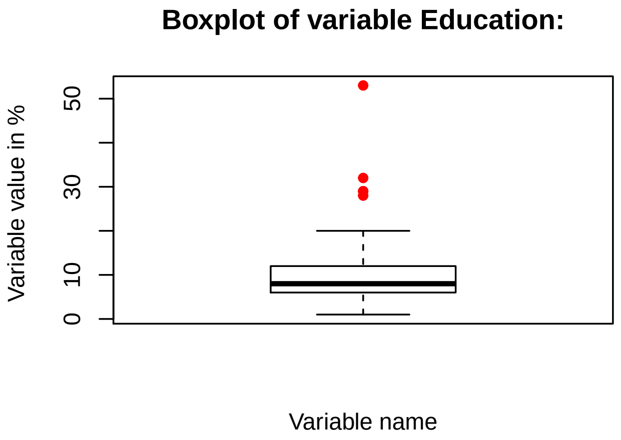
- Mean of Education: 10.97
- SD of Education: 9.61
- Median of Education: 8
- IQR of Education: 6
- Maximum of Education: 53
- Minimum of Education: 1
- Outliers in Education: 28, 32, 53, 29 and 29

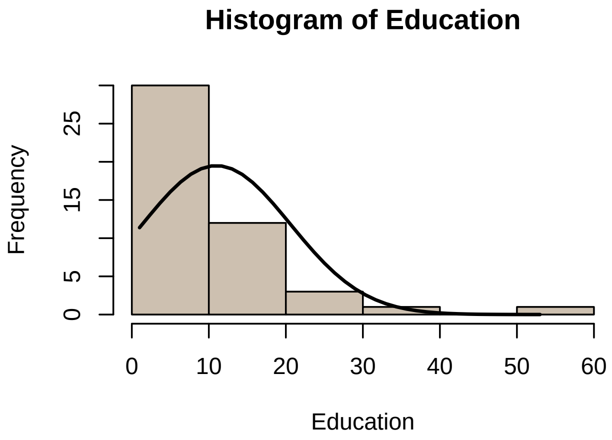
Stem-and-Leaf plot of Education

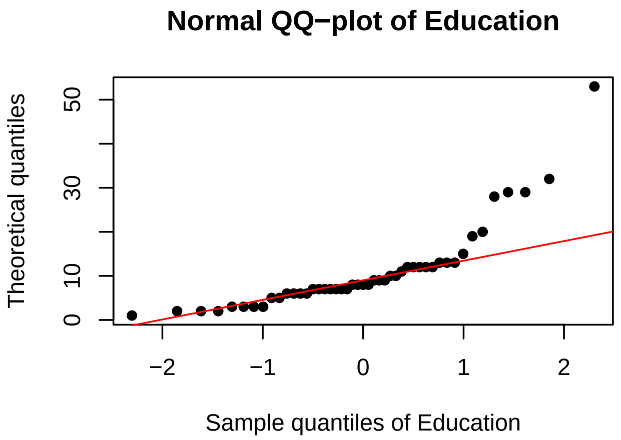
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##   0 | 1222333355666677777778888999
##   1 | 0012222233359
##   2 | 0899
##   3 | 2
##   4 |
##   5 | 3
```

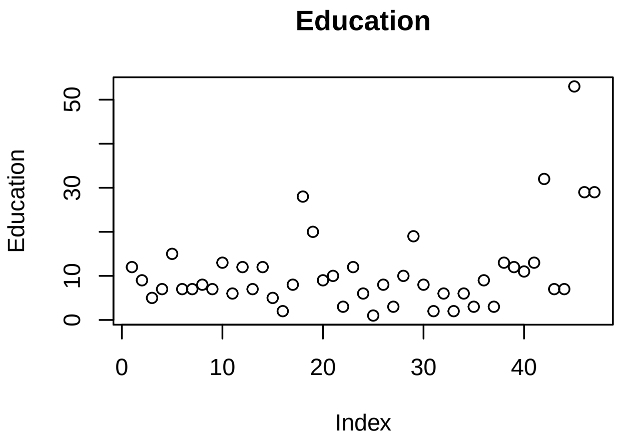
We can observe that large amount of the values lies in the 0-10 range and the data does not contains any value in 40-50 range and thus the variable is not continuous, so we can infer that data do not come from the normal population.

Visualization:









Histogram: The histogram with the density curve clearly shows that maximum frequency of the values lie towards extreme right i.e. the values of the variables are positively skewed and so the data is not from normal population.

Boxplot: From the box plot we can interpret that the variable Education is right–skewed which we can comment as positively skewed variable, the values of variables are not so dispersed because the length of the two ends of the whiskers is almost close to each other and the median also lie slightly towards right of the center . And the data of Education contains three outlier marked as red dots in the box plot.

QQ Plot: The plot clearly shows that most of the values lie away from the normal line. So we can interpret that the data is not from a normal distribution.

Outliers: Values 28, 32, 53, 29 and 29 are away from the data cloud (in this case extreme in y), so they will be considered as outliers.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

D	p-value
0.25	0.01

- The observed value of the Shapiro-Wilk statistic is: D = 0.24654
- The exact probability of the observed value, D = 0.24654, p-value = 0.006603
- For the Fertility, p-value = 0.006603, which is less than .05.

Shapiro-Wilk normality test(Parametric test):

W	p-value
0.75	0.00

- The observed value of the Shapiro-Wilk statistic is: W = 0.7482
- The exact probability of the observed value, W = 0.7482, p-value = 1.312e-07
- For the Education, p = 1.312e-07, which is less than .05.

The parent population is not normally distributed.

For variable Catholic:

Descriptive Statistics

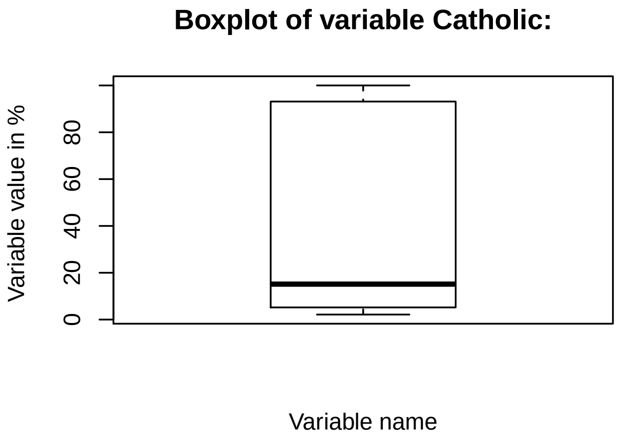
- Mean of Catholic: 41.14
- SD of Catholic: 41.70
- Median of Catholic: 15.14
- IQR of Catholic: 87.93
- Maximum of Catholic: 100
- Minimum of Catholic: 2.15

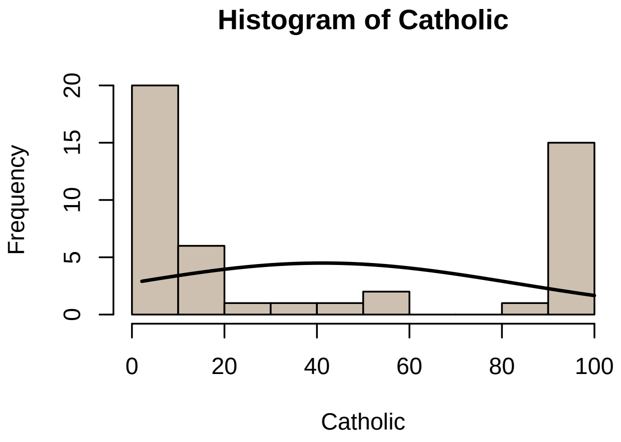
Stem-and-Leaf plot of Catholic

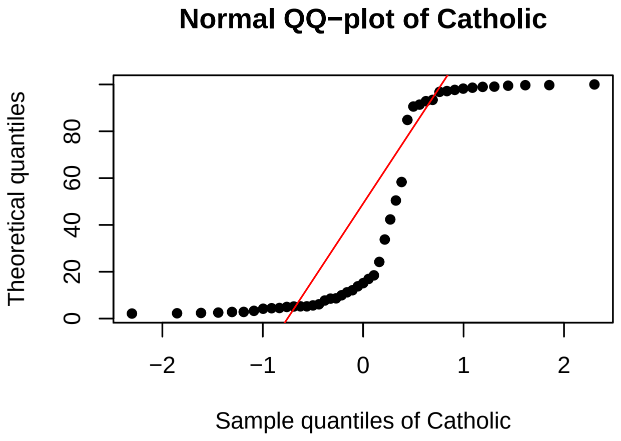
```
##
##   The decimal point is 1 digit(s) to the right of the |
##
##    0 | 2223333445555566899
##    1 | 0124578
##    2 | 4
##    3 | 4
##    4 | 2
##    5 | 08
##    6 |
##    7 |
##    8 | 5
##    9 | 113377889999
##   10 | 000
```

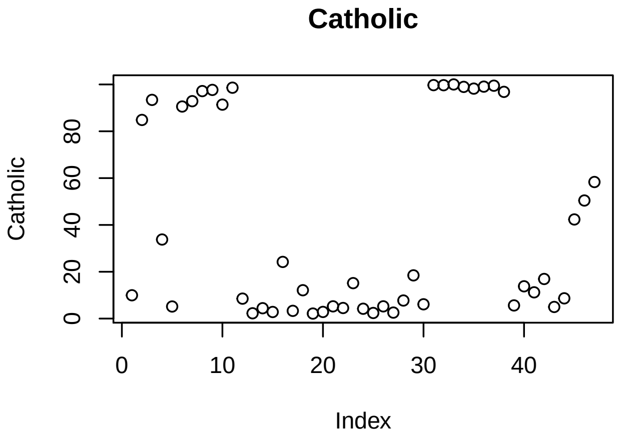
We can observe that large amount of the values lies in the 0-10 range and in the 90-100 range , the data do not contain any value in 60-70 range and thus the variable is not continuous, so we can infer that data do not come from the normal population.

Visualization:









Histogram: The histogram with the density curve clearly shows that maximum frequency of the values lie towards extreme right and left i.e. the values form a pattern of inverted U, and we also can observe that the density curve is almost a straight line . Hence , it is clearly seen that the data is not from normal population.

Boxplot: From the box plot we can interpret that the variable Catholic is right–skewed which we can comment as positively skewed variable, the values of variables are very much dispersed because the length of the two ends of the whiskers is almost far away from each other and the median also lie to the right of the center . And the data Catholic do not contain any outlier .

QQ Plot: The plot clearly shows that almost all the values lie away from the normal line except for two or three. So we can interpret that the data is not from a normal distribution.

Outliers: There is not outlier in the variable under consideration.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

D	p-value
0.26	0.003

- The observed value of the Kolmogorov-Smirnov statistic is: 0.25994
- The exact probability of the observed value, D = 0.25994, p-value = 0.003488
- For the Catholic, p = 0.003488, which is less than .05.

Shapiro-Wilk normality test(Parametric test):

W	p-value
0.74	0.00

- The observed value of the Shapiro-Wilk statistic is: W = 0.7463
- The exact probability of the observed value, W = 0.7463, p-value = 1.205e-07
- For the Catholic, p-value = 0.003488, which is less than .05.

The parent population is not normally distributed.

For variable Infant Mortality:

Descriptive Statistics

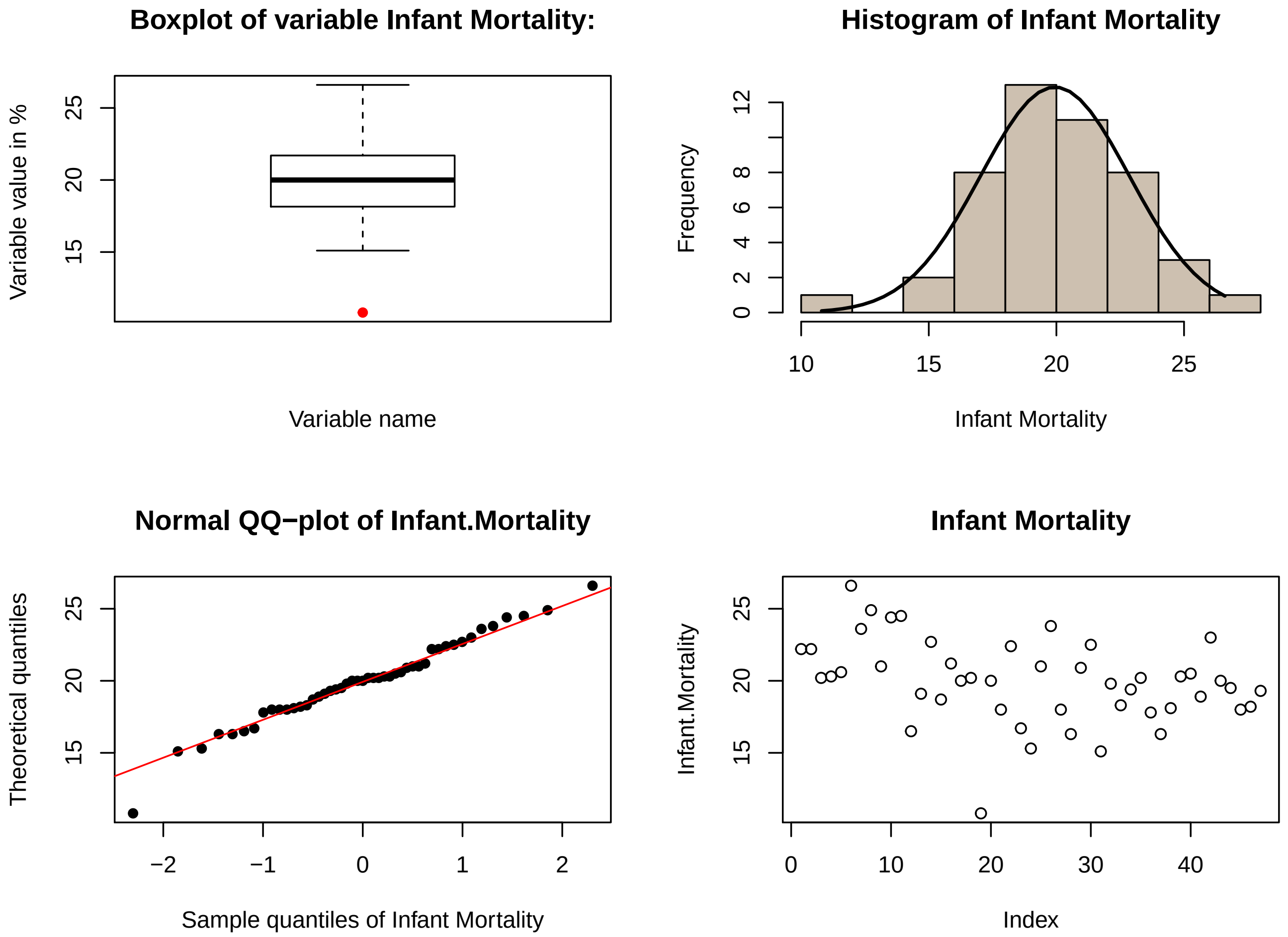
- Mean of Infant Mortality: 19.94
- SD of Infant Mortality: 2.91
- Median of Infant Mortality: 20
- IQR of Infant Mortality: 3.55
- Maximum of Infant Mortality: 26.6
- Minimum of Infant Mortality: 10.8
- Outliers in Infant.Mortality: 10.8

Stem-and-Leaf plot of Infant Mortality

```
##
##   The decimal point is at the |
##
##   10 | 8
##   12 |
##   14 | 13
##   16 | 33578
##   18 | 0001237913458
##   20 | 00022233569002
##   22 | 22457068
##   24 | 459
##   26 | 6
```

We can observe that large amount of the values lies in the 180-210 range and the data is randomly distributed in the range 100-270. So we can infer that data come from the normal population.

Visualization:



Histogram: The histogram with the density curve clearly shows that maximum frequency of the values lie in the center and all the bars are within the density curve. Hence , it is clearly seen that the data is from normal population.

Boxplot: From the box plot we can interpret that the variable Catholic is slightly left–skewed which we can comment as almost symmetric variable, the values of variables are not much dispersed because the length of the two ends of the whiskers is almost close to each other and the median also lie at the center . And the data Infant.Mortality contain only one outlier denoted with the red dots in the plot.

QQ Plot: The QQ-plot clearly shows that almost all the values lie along the normal line. So we can interpret that the data is not from a normal distribution.

Outliers: There a outlier in the variable under consideration and that is 10.8.

Hypothesis testing:

Kolmogorov-Smirnov test (Nonparametric test):

D	p-value
0.08	0.91

- The observed value of the Kolmogorov-Smirnov test statistic is: D = 0.082197
- The exact probability of the observed value, D = 0.082197, p-value = 0.9086
- For the Infant Mortality, p-value = 0.9086, which is greater than .05.

Shapiro-Wilk normality test(Parametric test):

W	p-value
0.97	0.49

- The observed value of the Shapiro-Wilk statistic is: W = 0.97762
- The exact probability of the observed value, W = 0.97762, p-value = 0.4978
- For the Infant Mortality, p = 0.4978, which is greater than .05.

The parent population is not normally distributed.

At the end of the series of conducted test and graphical inspection this can be concluded by both parametric and non-parametric test that **Fertility, Agriculture, Examination and Infant Mortality are from population with normally distribution** but **Education and Catholic are from population with any other distrubtion except normal distribution**.

Correlation analysis of the multivariate data:

```r
par(mfrow=c(1,1))
panel.cor <- function(x, y, digits=2, prefix="", cex.cor)
{
  usr <- par("usr"); on.exit(par(usr))
  par(usr = c(0, 1, 0, 1))
  r <- abs(cor(x, y))
  txt <- format(c(r, 0.123456789), digits=digits)[1]
  txt <- paste(prefix, txt, sep="")
  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)

  test <- cor.test(x,y)
  # borrowed from printCoefmat
  Signif <- symnum(test$p.value, corr = FALSE, na = FALSE,
                   cutpoints = c(0, 0.001, 0.01, 0.05, 0.1, 1),
                   symbols = c("***", "**", "*", ".", " "))

  text(0.5, 0.5, txt, cex = cex * r)
  text(.7, .7, Signif, cex=cex, col=2)
}

pairs(swiss, lower.panel=panel.smooth, upper.panel=panel.cor)
```

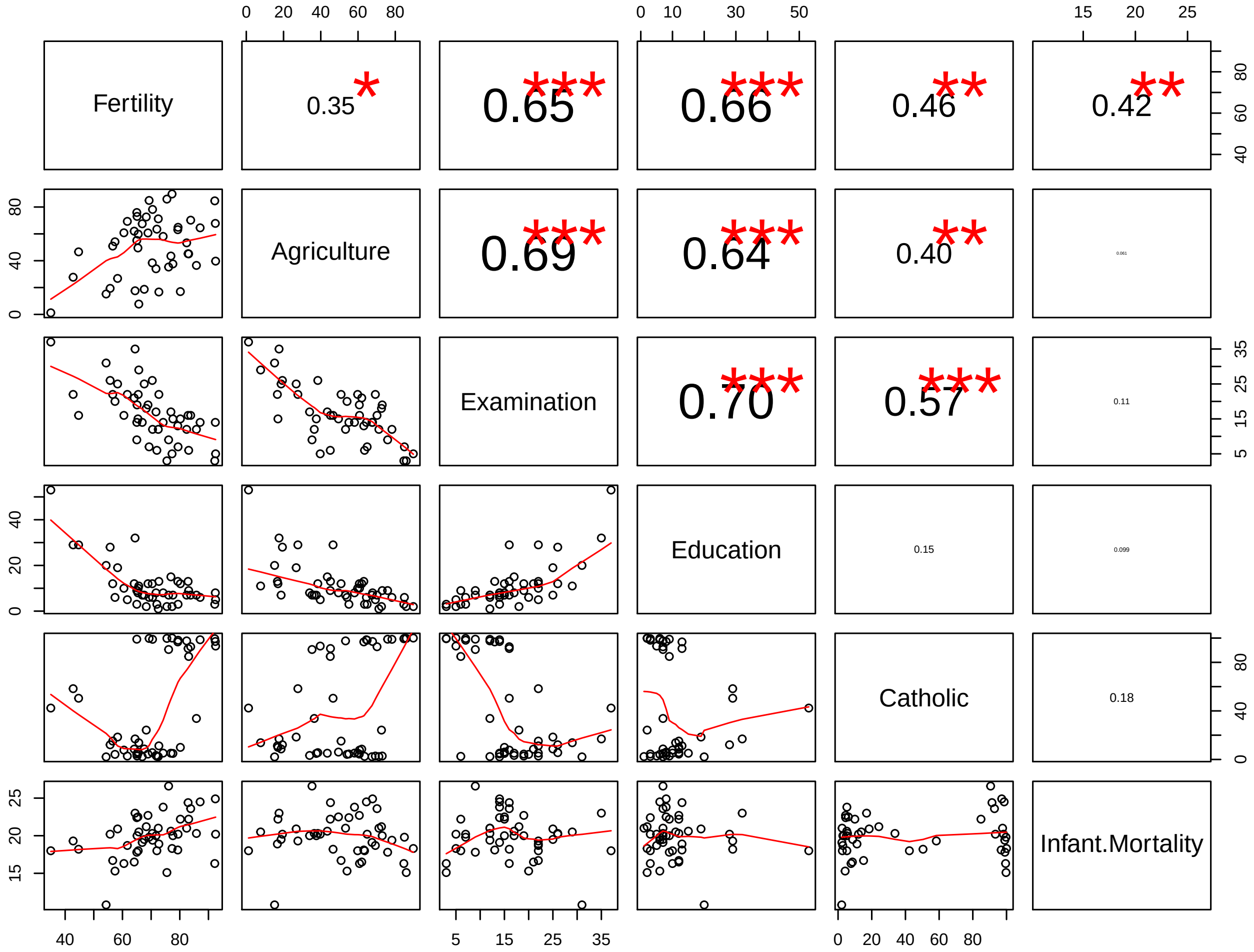


Table 1: Correlation matrix of swiss data set:

Fertility	Agriculture	Examination	Education	Catholic	Infant.Mortality
1.000	0.353	-0.646	-0.664	0.464	0.417
0.353	1.000	-0.687	-0.640	0.401	-0.061
-0.646	-0.687	1.000	0.698	-0.573	-0.114
-0.664	-0.640	0.698	1.000	-0.154	-0.099
0.464	0.401	-0.573	-0.154	1.000	0.175
0.417	-0.061	-0.114	-0.099	0.175	1.000

Pairwise comperison:

For Fertility and Agriculture :

Correlation Coefficien:

- Estimated Pearson's product-moment correlation is 0.3530792 and corresponding p-value is 0.01492 , So we reject the null hypothesis.

For Fertility and Examination :

- Estimated Pearson's product-moment correlation is -0.6458827 and corresponding p-value is 9.45e-07(<.01) , So we reject the null hypothesis.

For Fertility and Education :

- Estimated Pearson's product-moment correlation is -0.6637889 and corresponding p-value is 3.659e-07(<0.01) , So we reject the null hypothesis.

For Fertility and Catholic :

- Estimated Pearson's product-moment correlation is 0.4636847 and corresponding p-value is 0.001029 , So we reject the null hypothesis.

For Fertility and Infant Mortality :

- Estimated Pearson's product-moment correlation is 0.416556 and corresponding p-value is 0.003585 , So we reject the null hypothesis.

For Agriculture and Examination :

- Estimated Pearson's product-moment correlation is -0.6865422 and corresponding p-value is 9.952e-08 (<.01) , So we reject the null hypothesis.

For Agriculture and Education :

- Estimated Pearson's product-moment correlation is -0.6395225 and corresponding p-value is 1.305e-06 (<.01) , So we reject the null hypothesis.

For Agriculture and Catholic :

- Estimated Pearson's product-moment correlation is 0.4010951 and corresponding p-value is 0.005204 , So we reject the null hypothesis.

For Agriculture and Infant Mortality :

- Estimated Pearson's product-moment correlation is -0.06085861 and corresponding p-value is 0.6845 , So we accept the null hypothesis.

For Examination and Education :

- Estimated Pearson's product-moment correlation is 0.6984153 and corresponding p-value is 4.811e-08 (<.01) , So we reject the null hypothesis.

For Examination and Catholic :

- Estimated Pearson's product-moment correlation is -0.06085861 and corresponding p-value is 2.588e-05 (<.01) , So we reject the null hypothesis.

For Examination and Infant Mortality :

- Estimated Pearson's product-moment correlation is -0.1140216 and corresponding p-value is 0.4454 , So we accept the null hypothesis.

For Education and Catholic :

- Estimated Pearson's product-moment correlation is -0.1538589 and corresponding p-value is 0.3018 , So we accept the null hypothesis.

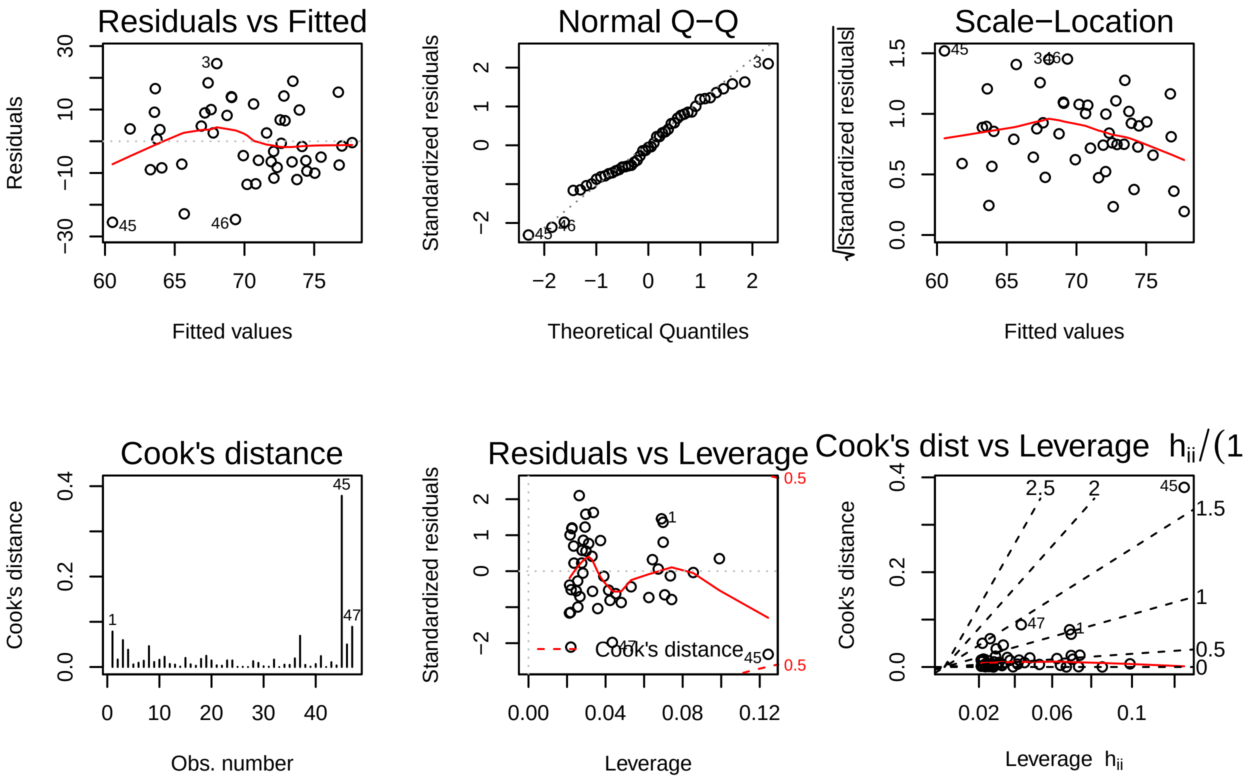
For Education and Infant Mortality :

- Estimated Pearson's product-moment correlation is -0.09932185 and corresponding p-value is 0.5065 , So we accept the null hypothesis.

For Catholic and Infant Mortality :

- Estimated Pearson's product-moment correlation is 0.1754959 and corresponding p-value is 0.238 , So we accept the null hypothesis.

Fitting OLS on Fertility variable by taking Agriculture as independent variable:



```
##
## Call:
## lm(formula = Fertility ~ Agriculture)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.5374  -7.8685  -0.6362   9.0464  24.4858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 60.30438    4.25126  14.185   <2e-16 ***
## Agriculture  0.19420    0.07671   2.532   0.0149 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.82 on 45 degrees of freedom
## Multiple R-squared:  0.1247, Adjusted R-squared:  0.1052
## F-statistic: 6.409 on 1 and 45 DF,  p-value: 0.01492
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.3044	4.2513	14.19	0.0000
Agriculture	0.1942	0.0767	2.53	0.0149

Multiple R-squared is the percentage of variance explained by the independent variable in the dependent variable. Adj Multiple R-squared is the adjusted value of the Multiple R-squared by the adjustment factor due to the estimation of the model parameter. Here Agriculture explains only the 10% variance in the Fertility data. Although from the table and summary statistics this can be shown that the intercept is not significant but the β_1 is significant. Over all the regression is significant shown the the ANOVA table.

The basic four assumption of regression analysis can be verified from the diagnostic plot of the regression analysis. The outliers and influential points can be diagnose from the diagnostic plot of residuals. The problem of less R-square can be overcomed by considering more variables under study.