STAT1006 Assignment

Lisa Luff - 16167920

Contents

Introduction

This assignment aims to explore and analyse how various factors can predict life expectancy, and how this varies between countries, and for gender. This is evaluated using data from Day, A. (ed.) (1992), _The Annual Register 1992_, 234, London: Longmans, and U.N.E.S.C.O. 1990 Demographic Year Book_ (1990), New York: United Nations. It was designed for use by first year statistics students.

The data is collected from 97 different countries and is broken up into sub-groups:

- Eastern Europe
- South America and Mexico
- Western Europe, North America, Japan, Australia, and New Zealand
- Middle East
- Asia
- Africa

The average life expectancy is also split into sub-groups by gender.

To achieve the aim of the assignment exploratory analysis is conducted to find which of the given factors have the greatest influence on average life expectancy, and how that varies amongst the sub-groups, then use that to find the best predictive model for average life expectancy.

If average life expectancy is able to be predicted accurately, this can identify important factors that influence this. It could also identify countries of concern that data is not currently held for. Or predict countries who will need greater care as their situation changes over time, allowing preventive action to be taken. Ultimately the aim is to improve average life expectancy.

Exploration of Data Patterns

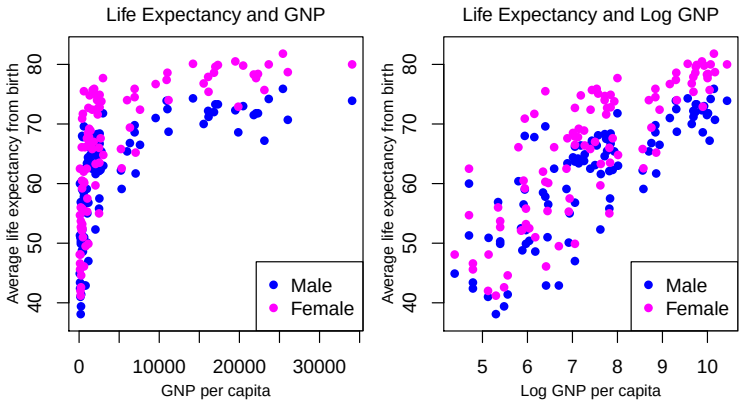
The variables being analysed to create a predictive model of average life expectancy are:

- Average male life expectancy from birth
- Average female life expectancy from birth
- Live birth rate per 1000 people in the population
- Death rate per 1000 people in the population
- Infant deaths (infants are those under 1 year of age) per 1000 infants in the population
- Gross National Product per capita in US dollars

Before anything else, an assessment of the data patterns is to be completed using a scatterplot of each explanatory variable against male and female life expectancy to assess the form, direction, strength, and variance of the data.

There appears to be a strong negative linear relationship between Life expectancy for birth rate, and death rate. There is clustering on the low end of infant death rates, which is otherwise negatively linear. GNP has a positive exponential relationship.

As the data is to be used in a linear model, GNP needs to be transformed from exponential to linear. This is done by using the log of all GNP values. The resulting plots are shown on the following page.

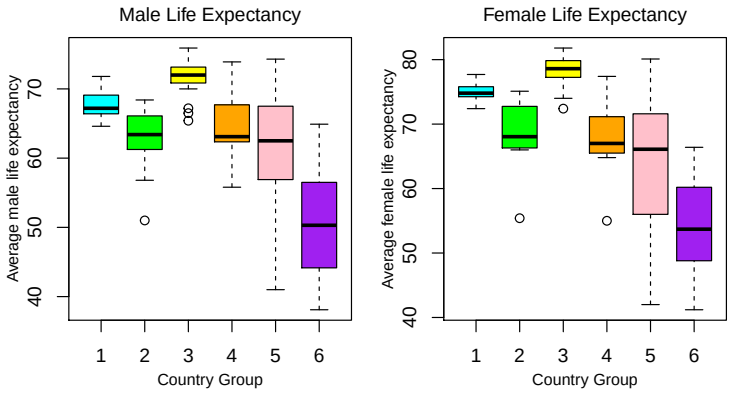


The plot now shows a strong positive linear relationship. This is confirmed by an improvement in correlation. Originally correlation was 0.632 for males, and 0.639 for females, and the log correlation is now 0.802 for males, and 0.816 for females. Increased by 0.183 and 0.177 respectively.

Exploring Data by Sub-Group

As there are so many possible factors that could possibly affect the average life expectancy of a country, it is important to explore how culture might have an impact on the data being assessed.

The affect of gender and how that differs amongst the country sub-groups also needs to be explored due to gender differences, but also how genders life experiences can be varied depending on culture. This is done in the below boxplots:



Where; 1 = Eastern Europe, 2 = South America and Mexico, 3 = Western Europe, North America, Japan, Australia, and New Zealand, 4 = Middle East, 5 = Asia, 6 = Africa

There is is a wide variety of medians, and variance, with several data sets containing outliers. The life expectancy for males and females does appear to be similar across groups, however a paired t-test of all life expectancy, and grouped life expectancy differences between males and females found there is a significant gender difference across the board.

Exploration of correlation plots for each explanatory variable for each gender, and boxplots exploring the explanatory variables by country group shows large variation in both correlation, and medians and variance across country groups across the board.

Simple Linear Models

To find the best predictive model, the best simple linear model and multiple linear regression model will be assessed. To do this a linear model for each explanatory variable is created and assessed, then comparitive anaylsis done between these models to find the most appropriate model.

Due to the large differences amongst subgroups, there might not be one best model for predictive analysis. So the best model will be assessed per country subgroup focussing on female life expectancy. The correlation graph matrices showed the data is not as Normally distributed for all Explanatory variables, for all groups, so the models will be limited to two which have relatively Normal data, but are also very different in terms of the data itself.

Model assessments:

- Western Europe, North America, Japan, Australia, and New Zealand

Test	Birth Rate	Death Rate	Infant Deaths	Log GNP
Random	✓	✓	curved	✓
Equal variance	skewed right	✓	✓	✓
Normality	tails wander	bottom tail wanders	✓	✓
Outliers	3	3	3	3
Leverage points	2 borderline	1 borderline	0	1
Test statistic	0.463	-1.422	-6.145	4.280
P-Value	0.649	0.173	1.08e-05	0.000507
Decision	Do not reject null	Do not reject null	Reject null	Reject null
R^2	0.01245	0.1063	0.6895	0.5186

As shown in the above table, for the first group, infant death rate is by far the most significant, and has the best R^2 value. The residuals do appear to be somewhat non-linear, however not quite enough to perform a transformation. Therefore, the chosen linear model is:
Female average life expectancy = 86.1828 - 1.0186*Infant deaths per 1000 infants in the population

$\beta_0 = 86.1828$ is not interpretable as it is outside of the data sets range, and is not a practically possible.

However, a β_1 of -1.0186 means that for each additional infant death per 1000 infants in the population, female average life expectancy is reduced by 1.0186 years.

- Africa

Test	Birth Rate	Death Rate	Infant Deaths	Log GNP
Random	negative linear	✓	✓	✓
Equal variance	decreasing	✓	✓	✓
Normality	✓	bottom tail wanders	s curve	✓
Outliers	3	3	3	3
Leverage points	2 borderline	1 borderline	1 borderline	0
Test statistic	-4.676	-17.13	-8.374	5.111
P-Value	8.62e-05	2.51e-13	1.01e-08	2.8e-05
Decision	Reject null	Reject null	Reject null	Reject null
R^2	0.4666	0.9215	0.7372	0.511

As shown in the above table, for the second group, all of the explanatory variables are significant, however death rates have a very high R^2 value. There are a few points that break away from the QQline, however as they are not leverage points, I will not be removing them. Therefore, the chosen linear model is:
Female average life expectancy = 74.71114 - 1.40689*Death rate per 1000 people in the population

$\beta_0 = 74.71114$ is not interpretable as it is outside of the data sets range, and is not a practically possible.

However, a β_1 of -1.40689 means that for each increase by 1 in death rate per 1000 people in the population, female average life expectancy is reduced by 1.40689 years.

Multiple Linear Regression

To choose the best possible multiple linear regression model to compare against the simple regression model, four different methods of choosing the best combination of variables are used, and then those models analysed and compared.

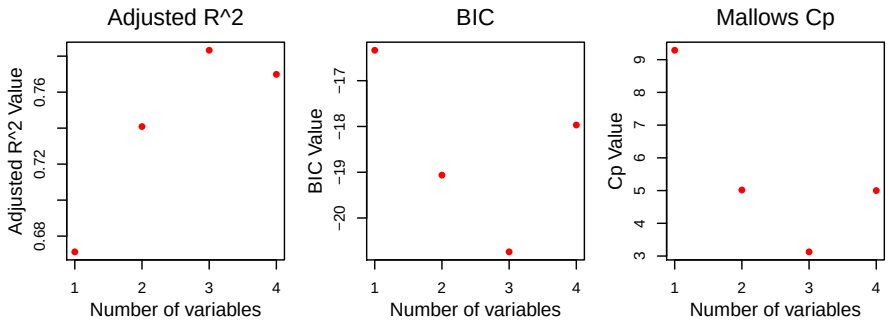
The methods being used are brute force, forward, backward and stepwise selection.

Again this will be done to find the best model for female life expectancy in the same two subgroups.

- Western Europe, North America, Japan, Australia and New Zealand
  - Brute force best model for each number of variables:

	births	deaths	infant	log.GNP
1 (1)			*	
2 (1)		*	*	
3 (1)		*	*	*
4 (1)	*	*	*	*

- Then compare those models with each other:



- Choosing the best model:
  - R^2_{adj} - Want the highest value
    - 3 variable model
  - BIC - Want the lowest possible
    - 3 variable model
  - C_p - Want the value to be approximately the same as the total number of parameters
    - 3 variable model

According to the brute force model, the 3 variable model is the best model, which uses the death rate, infant death rate and log GNP.

- Forward best model:

```
## (Intercept)      infant      deaths     log.GNP
##  71.3192606  -0.6894930  -0.5110037   1.7526050
```

According to the forward model, the same 3 variable model is the best model.

- Backward best model:

```
## (Intercept)      deaths      infant     log.GNP
##  71.3192606  -0.5110037  -0.6894930   1.7526050
```

According to the backward model, the same 3 variable model is the best model.

- Stepwise best model:

```
## (Intercept)      deaths      infant     log.GNP
##  71.3192606  -0.5110037  -0.6894930   1.7526050
```

According to the backward model, the same 3 variable model is the best model.

All four methods agree that the three variable model with death rate, infant death rate and log GNP is the best model. So only that multiple linear regression model will be assessed for validity.

The three variable model is:
Average female life expectancy = 71.3193 - 0.5110*Deaths per 1000 people in the population - 0.6895*Infant deaths per 1000 infants in the population + 1.7526*Log GNP per capita in US dollars

Diagnostics:
The assessment done on the individual variables previously carries through to the multiple linear model. The diagnostics show that the multiple linear model is valid based on all the assumptions for a linear model being met.

Despite all three methods agreeing, log.GNP is not significant, and a test of a 2 variable model comparing to the 3 variable model shows there is not a significant difference, so the model used will be the 2 variable model:
Average female life expectancy = 90.1516 - 0.4359*Deaths per 1000 people in the population - 1.0004*Infant deaths per 1000 infants in the population

Next is assessing if the relationship is significant, by testing the hypothesis that there is no relationship based on a significance level of 0.05. Hypothesis testing:

1. Hypotheses:

- $H_0 : \beta_1 = \beta_2 = 0$
- H_A : at least one $\beta_i \neq 0$

2. Test statistic:

- $F =$ 26.73

3. Sampling distribution:

- $F \sim f_{(2,16)}$

4. P-Value

- $P(|f_{(2,16)}| > 26.73) = 7.922e-06$

5. Decision

- Based on the p-value being so small compared to the significance level of 0.05, there is strong evidence to reject the null hypothesis

6. Conclusion

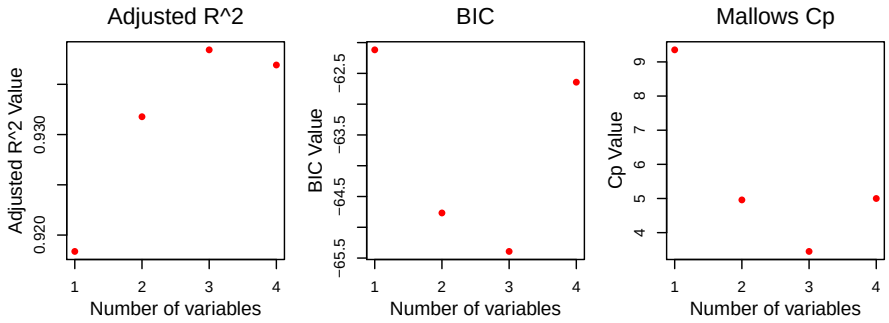
- There is strong evidence to suggest that there is a relationship between average female life expectancy and at least one of the explanatory variables.

How well the data is explained by the model is assessed using R^2_{adj} which takes into account that there are multiple variables. In this case $R^2_{adj} = 0.7409$, which means that 74.09% of the data is explained by the model. Which is a moderate proportion of the data, suggesting the model is a good fit.

- Africa
  - Brute force best model for each number of variables:

	births	deaths	infant	log.GNP
1 (1)		*		
2 (1)	*	*		
3 (1)	*	*	*	
4 (1)	*	*	*	*

- Then compare those models with each other:



- Choosing the best model:
  - R^2_{adj} - Want the highest value
    - 3 variable model
  - BIC - Want the lowest possible
    - 3 variable model
  - C_p - Want the value to be approximately the same as the total number of parameters
    - 3 variable model

According to the brute force model, the 3 variable model is the best model, which uses the birth rate, death rate, infant death rate.

- Forward best model:

```
## (Intercept)      deaths      births      infant
## 82.02826424 -1.05961680 -0.19054197 -0.03919124
```

According to the forward model, the same 3 variable model is the best model.

- Backward best model:

```
## (Intercept)      births      deaths      infant
## 82.02826424 -0.19054197 -1.05961680 -0.03919124
```

According to the backward model, the same 3 variable model is the best model.

- Stepwise best model:

```
## (Intercept)      births      deaths      infant
## 82.02826424 -0.19054197 -1.05961680 -0.03919124
```

According to the backward model, the same 3 variable model is the best model.

All four methods agree that the three variable model with birth rate, death rate, and infant death rate is the best model. So only that multiple linear regression model will be assessed for validity.

The three variable model is:
Average female life expectancy = 82.02826 - 0.19054*Births per 1000 people in the population - 1.05962*Deaths per 1000 people in the population - 0.03919*Infant deaths per 1000 infants in the population

Diagnostics:
The assessment done on the individual variables previously carries through to the multiple linear model. The diagnostics show that the multiple linear model is valid based on all the assumptions for a linear model being met.

Despite all three methods agreeing, infant birth rate is not significant, and a test of a 2 variable model comparing to the 3 variable model shows there is not a significant difference, so the model used will be the 2 variable model:
Average female life expectancy = 81.3112 - 0.19421*Births per 1000 people in the population - 1.26687*Deaths per 1000 people in the population

Next is assessing if the relationship is significant, by testing the hypothesis that there is no relationship based on a significance level of 0.05. Hypothesis testing:

1. Hypotheses:

- $H_0 : \beta_1 = \beta_2 = 0$
- H_A : at least one $\beta_i \neq 0$

2. Test statistic:

- $F =$ 178.5

3. Sampling distribution:

- $F \sim f_{(2,24)}$

4. P-Value

- $P(|f_{(2,24)}| > 178.5) = 3.892e - 15$

5. Decision

- Based on the p-value being so small compared to the significance level of 0.05, there is strong evidence to reject the null hypothesis

6. Conclusion

- There is strong evidence to suggest that there is a relationship between average female life expectancy and at least one of the explanatory variables.

In this case $R^2_{adj} = 0.9318$. Which is a high proportion of the data, suggesting the model is a good fit.

Best Predictive Model

The goal is to create the best predictive model, so deciding whether the simple linear model, or multiple linear model is the best model will be decided based on which is better at predicting. As the final model is to be the best model, a test group of data was not removed during analysis, therefore the predicitve analysis to be used is the predicted residual sum of squares method (PRESS).

Western Europe, North America, Japan, Australia and New Zealand - The two best models were:

- Simple linear model
  - Female average life expectancy = 86.1828 - 1.0186*Infant deaths per 1000 infants in the population
- Multi-variate model
  - Average female life expectancy = 90.1516 - 0.4359*Deaths per 1000 people in the population - 1.0004*Infant deaths per 1000 infants in the population

With the PRESS method, the model with the lowest PRESS is considered to be the best model. The results are as follows:

- Simple linear model: 34.59387
- Multi-variate model: 28.89231

The preferred model is the one with the lowest PRESS score, which in this case is the multi-variate model.Therefore, the multi-variate model is the model best suited to give accurate predictions about the average female life expentancy of a country within this group.

Africa - The two best models were:

- Simple linear model
  - Female average life expectancy = 74.71114 - 1.40689*Death rate per 1000 people in the population
- Multi-variate model
  - Average female life expectancy = 81.3112 - 0.19421*Births per 1000 people in the population - 1.26687*Deaths per 1000 people in the population

With the PRESS method, the model with the lowest PRESS is considered to be the best model. The results are as follows:

- Simple linear model: 114.0283
- Multi-variate model: 97.59061

The preferred model is the one with the lowest PRESS score, which in this case is the multi-variate model.Therefore, the multi-variate model is the model best suited to give accurate predictions about the average female life expentancy of a country within this group.

Discussion

As shown exploring the data, there are large differences between the various groups, so any predictive analysis should be done based on groups, and consideration should be given to how the countries are grouped.

The final results were “good” fits, however based on there being evidence of some pattern to the residuals of the final models, I believe with more advanced data transformation techiniques, or modelling techniques better fits can be found and would be more suited to use for prediction.

The limitations of this study are having so few explanatory variables. A persons life expectancy depends on so many more variables than just the four assessed, and I don’t feel those assessed provide the full story. Things such as medical care access, quality of medical care, access to clean water, and enough food, etc. Beyond that though, a better indicator of how wealth affects life expectancy is likely to be median personal as opposed to GNP.

Conclusion

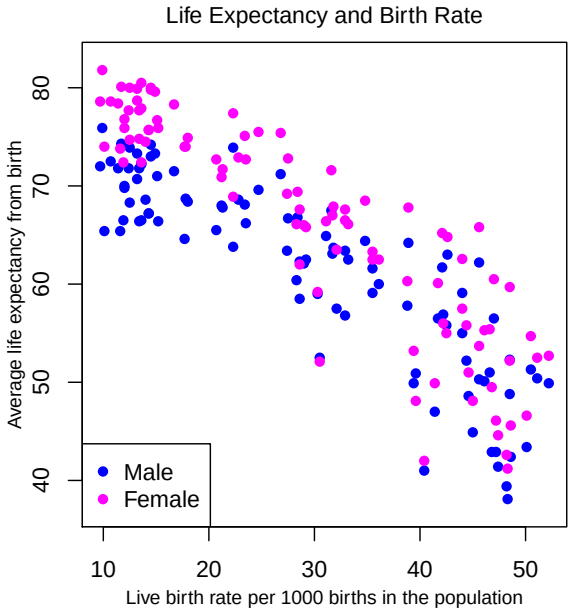
In conclusion, different variables will be more indicative of average life expectancy in females in different countries, and more studies should be done to find out why that is, so better predictive models can be made.

Appendix:

Exploratory scatterplots and correlations

- Birth Rate:

```
# Male and female life expectancy ~ birth rate plot
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
plot(Male.Expectancy ~ Birth.Rate, data = LifeExpectancy,
     ylim = c(min(Female.Expectancy, Male.Expectancy) - 1, max(Female.Expectancy, Male.Expectancy) + 1),
     pch = 16,
     col = 'blue',
     ann = FALSE)
mtext(side = 3, line = 0.5,
      'Life Expectancy and Birth Rate',
      cex = 1)
mtext(side = 1, line = 2,
      'Live birth rate per 1000 births in the population',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average life expectancy from birth',
      cex = 0.8)
points(Female.Expectancy ~ Birth.Rate, data = LifeExpectancy,
       pch = 16,
       col = 'magenta')
legend('bottomleft', legend = c('Male', 'Female'),
       col = c('blue', 'magenta'),
       pch = 16)
```




```
# Male and female life expectancy ~ birth rate analysis
cor(LifeExpectancy$Birth.Rate, LifeExpectancy$Male.Expectancy, method = 'pearson')
```

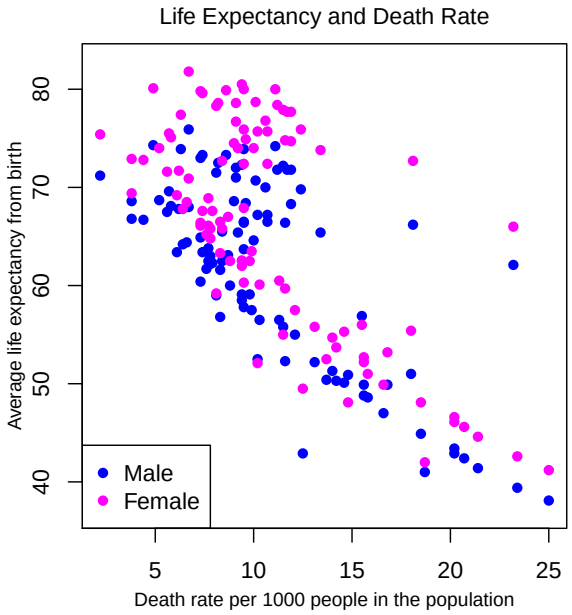
```
## [1] -0.8665189
```

```
cor(LifeExpectancy$Birth.Rate, LifeExpectancy$Female.Expectancy, method = 'pearson')
```

```
## [1] -0.894414
```

- Death Rate:

```
# Male and female life expectancy ~ death rate plot
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
plot(Male.Expectancy ~ Death.rate, data = LifeExpectancy,
     ylim = c(min(Female.Expectancy, Male.Expectancy) - 1, max(Female.Expectancy, Male.Expectancy) + 1),
     pch = 16,
     col = 'blue',
     ann = FALSE)
mtext(side = 3, line = 0.5,
      'Life Expectancy and Death Rate',
      cex = 1)
mtext(side = 1, line = 2,
      'Death rate per 1000 people in the population',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average life expectancy from birth',
      cex = 0.8)
points(Female.Expectancy ~ Death.rate, data = LifeExpectancy,
       pch = 16,
       col = 'magenta')
legend('bottomleft', legend = c('Male', 'Female'),
       col = c('blue', 'magenta'),
       pch = 16)
```




```
# Male and female life expectancy ~ death rate analysis
cor(LifeExpectancy$Death.rate, LifeExpectancy$Male.Expectancy, method = 'pearson')
```

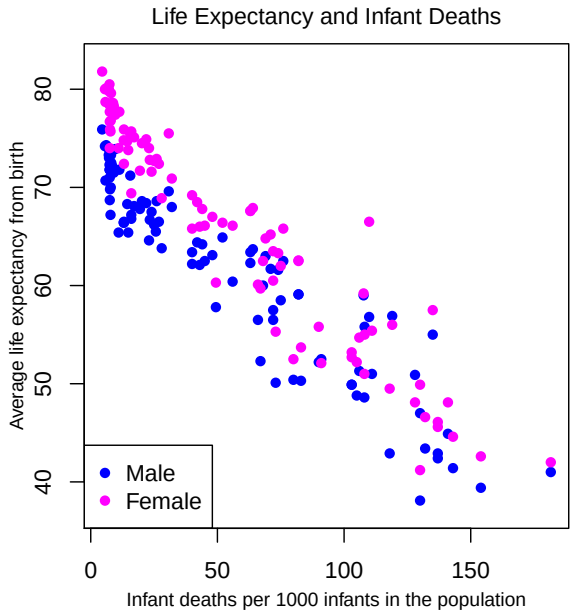
```
## [1] -0.7334666
```

```
cor(LifeExpectancy$Death.rate, LifeExpectancy$Female.Expectancy, method = 'pearson')
```

```
## [1] -0.6930331
```

- Infant Deaths:

```
# Male and female life expectancy ~ infant deaths plot
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
plot(Male.Expectancy ~ Infant.Deaths, data = LifeExpectancy,
     ylim = c(min(Female.Expectancy, Male.Expectancy) - 1, max(Female.Expectancy, Male.Expectancy) + 1),
     pch = 16,
     col = 'blue',
     ann = FALSE)
mtext(side = 3, line = 0.5,
      'Life Expectancy and Infant Deaths',
      cex = 1)
mtext(side = 1, line = 2,
      'Infant deaths per 1000 infants in the population',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average life expectancy from birth',
      cex = 0.8)
points(Female.Expectancy ~ Infant.Deaths, data = LifeExpectancy,
       pch = 16,
       col = 'magenta')
legend('bottomleft', legend = c('Male', 'Female'),
       col = c('blue', 'magenta'),
       pch = 16)
```




```
# Male and female life expectancy ~ infant deaths analysis
cor(LifeExpectancy$Infant.Deaths, LifeExpectancy$Male.Expectancy, method = 'pearson')
```

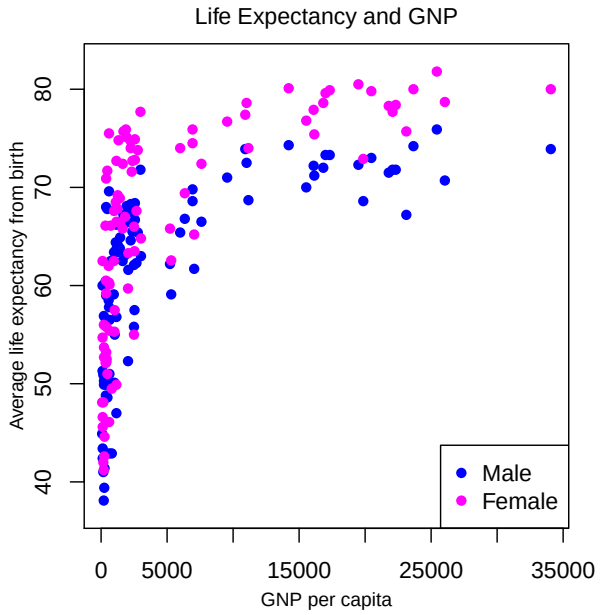
```
## [1] -0.9368384
```

```
cor(LifeExpectancy$Infant.Deaths, LifeExpectancy$Female.Expectancy, method = 'pearson')
```

```
## [1] -0.9553516
```

- GNP:

```
# Male and female life expectancy ~ GNP plot
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
plot(Male.Expectancy ~ GNP, data = LifeExpectancy,
     ylim = c(min(Female.Expectancy, Male.Expectancy) - 1, max(Female.Expectancy, Male.Expectancy) + 1),
     pch = 16,
     col = 'blue',
     ann = FALSE)
mtext(side = 3, line = 0.5,
      'Life Expectancy and GNP',
      cex = 1)
mtext(side = 1, line = 2,
      'GNP per capita',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average life expectancy from birth',
      cex = 0.8)
points(Female.Expectancy ~ GNP, data = LifeExpectancy,
       pch = 16,
       col = 'magenta')
legend('bottomright', legend = c('Male', 'Female'),
       col = c('blue', 'magenta'),
       pch = 16)
```




```
# Male and female life expectancy ~ GNP analysis
cor(LifeExpectancy$GNP, LifeExpectancy$Male.Expectancy, method = 'pearson')
```

```
## [1] 0.6320091
```

```
cor(LifeExpectancy$GNP, LifeExpectancy$Female.Expectancy, method = 'pearson')
```

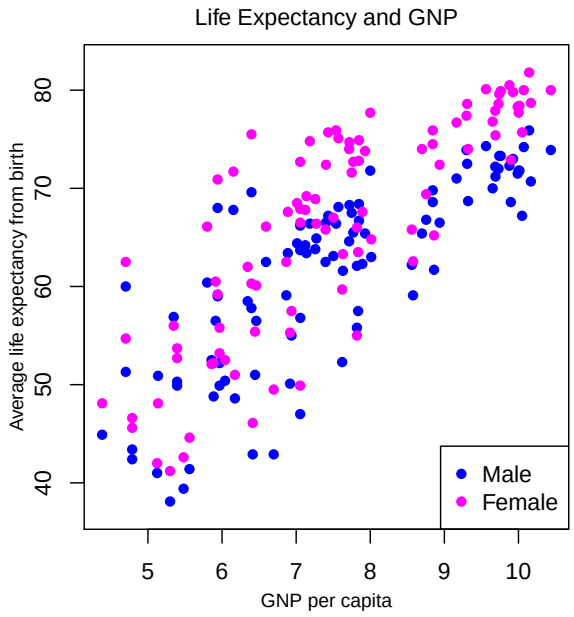
```
## [1] 0.6389077
```

- Log GNP:
  - Creating the log data

```
log.GNP <- log(LifeExpectancy$GNP)
```

- Scatterplot and correlation

```
# Male and female life expectancy ~ log GNP plot
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
plot(LifeExpectancy$Male.Expectancy ~ log.GNP,
     ylim = c(min(LifeExpectancy$Female.Expectancy, LifeExpectancy$Male.Expectancy) - 1, max(LifeExpectancy$Femal
     pch = 16,
     col = 'blue',
     ann = FALSE)
mtext(side = 3, line = 0.5,
      'Life Expectancy and GNP',
      cex = 1)
mtext(side = 1, line = 2,
      'GNP per capita',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average life expectancy from birth',
      cex = 0.8)
points(LifeExpectancy$Female.Expectancy ~ log.GNP,
       pch = 16,
       col = 'magenta')
legend('bottomright', legend = c('Male', 'Female'),
       col = c('blue', 'magenta'),
       pch = 16)
```




```
# Male and female life expectancy ~ log GNP analysis
cor(log.GNP, LifeExpectancy$Male.Expectancy, method = 'pearson')
```

```
## [1] 0.8011146
```

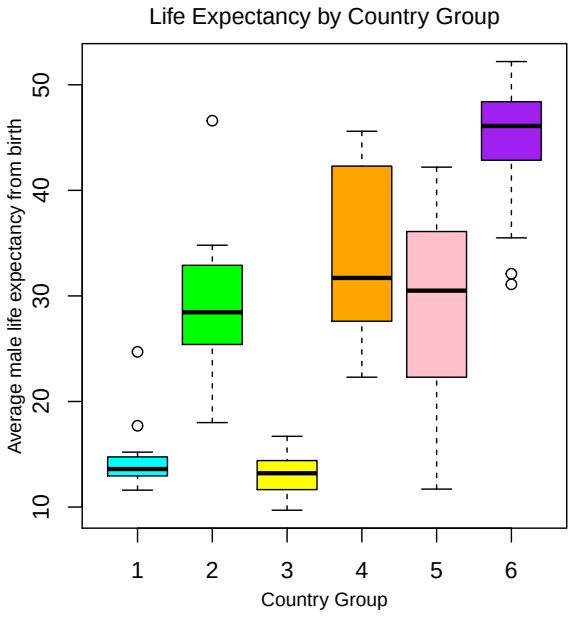
```
cor(log.GNP, LifeExpectancy$Female.Expectancy, method = 'pearson')
```

```
## [1] 0.8158444
```

Exploration by Sub-Group

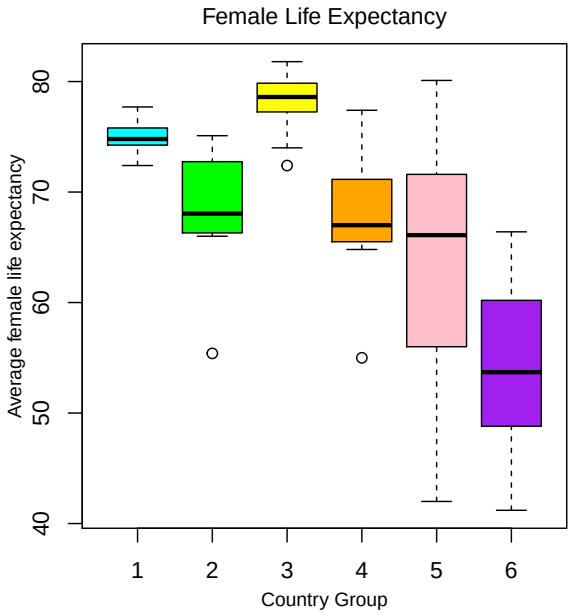
- Boxplots
  - Male

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(LifeExpectancy$Birth.Rate ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Life Expectancy by Country Group',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average male life expectancy from birth',
      cex = 0.8)
```



- Female

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(LifeExpectancy$Female.Expectancy ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Female Life Expectancy',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Average female life expectancy',
      cex = 0.8)
```



- Set up sub-groups

```
maleexpectancy <- LifeExpectancy$Male.Expectancy
femaleexpectancy <- LifeExpectancy$Female.Expectancy
births <- LifeExpectancy$Birth.Rate
deaths <- LifeExpectancy$Death.rate
infant <- LifeExpectancy$Infant.Deaths
groups <- LifeExpectancy$Country.Group

maledata <- cbind.data.frame(maleexpectancy, births, deaths, infant, log.GNP, groups)
femaledata <- cbind.data.frame(femaleexpectancy, births, deaths, infant, log.GNP, groups)

library(dplyr)
sub1M <- filter(maledata[1:5], groups == 1)
sub2M <- filter(maledata[1:5], groups == 2)
sub3M <- filter(maledata[1:5], groups == 3)
sub4M <- filter(maledata[1:5], groups == 4)
sub5M <- filter(maledata[1:5], groups == 5)
sub6M <- filter(maledata[1:5], groups == 6)

sub1F <- filter(femaledata[1:5], groups == 1)
sub2F <- filter(femaledata[1:5], groups == 2)
sub3F <- filter(femaledata[1:5], groups == 3)
sub4F <- filter(femaledata[1:5], groups == 4)
sub5F <- filter(femaledata[1:5], groups == 5)
sub6F <- filter(femaledata[1:5], groups == 6)
```

- Test of significance, gender differences
  - Male vs Female all

```
t.test(maleexpectancy, femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  maleexpectancy and femaleexpectancy
## t = -19.379, df = 96, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -5.143454 -4.187680
## sample estimates:
## mean of the differences
##               -4.665567
```

- Sub group 1

```
t.test(sub1M$maleexpectancy , sub1F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub1M$maleexpectancy and sub1F$femaleexpectancy
## t = -15.896, df = 10, p-value = 2e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -8.323269 -6.276731
## sample estimates:
## mean of the differences
##                    -7.3
```

- Sub group 2

```
t.test(sub2M$maleexpectancy , sub2F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub2M$maleexpectancy and sub2F$femaleexpectancy
## t = -12.126, df = 11, p-value = 1.045e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -6.882336 -4.767664
## sample estimates:
## mean of the differences
##                  -5.825
```

- Sub group 3

```
t.test(sub3M$maleexpectancy , sub3F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub3M$maleexpectancy and sub3F$femaleexpectancy
## t = -30.452, df = 18, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -7.139735 -6.218160
## sample estimates:
## mean of the differences
##               -6.678947
```

- Sub group 4

```
t.test(sub4M$maleexpectancy , sub4F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub4M$maleexpectancy and sub4F$femaleexpectancy
## t = -6.9262, df = 10, p-value = 4.063e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.025174 -2.065735
## sample estimates:
## mean of the differences
##               -3.045455
```

- Sub group 5

```
t.test(sub5M$maleexpectancy , sub5F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub5M$maleexpectancy and sub5F$femaleexpectancy
## t = -4.4745, df = 16, p-value = 0.0003832
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.109227 -1.467243
## sample estimates:
## mean of the differences
##               -2.788235
```

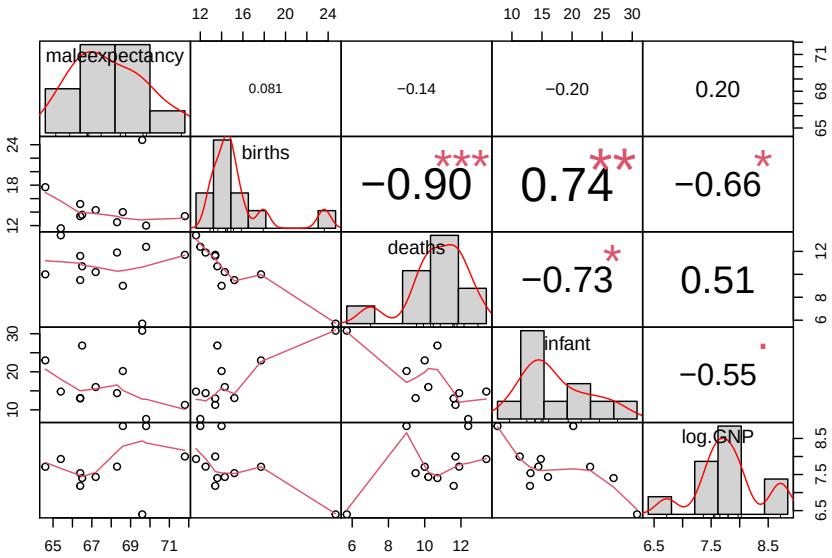
- Sub group 6

```
t.test(sub6M$maleexpectancy , sub6F$femaleexpectancy, mu = 0, paired = TRUE)
```

```
##
##  Paired t-test
##
## data:  sub6M$maleexpectancy and sub6F$femaleexpectancy
## t = -13.406, df = 26, p-value = 3.464e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.039223 -2.965221
## sample estimates:
## mean of the differences
##               -3.502222
```

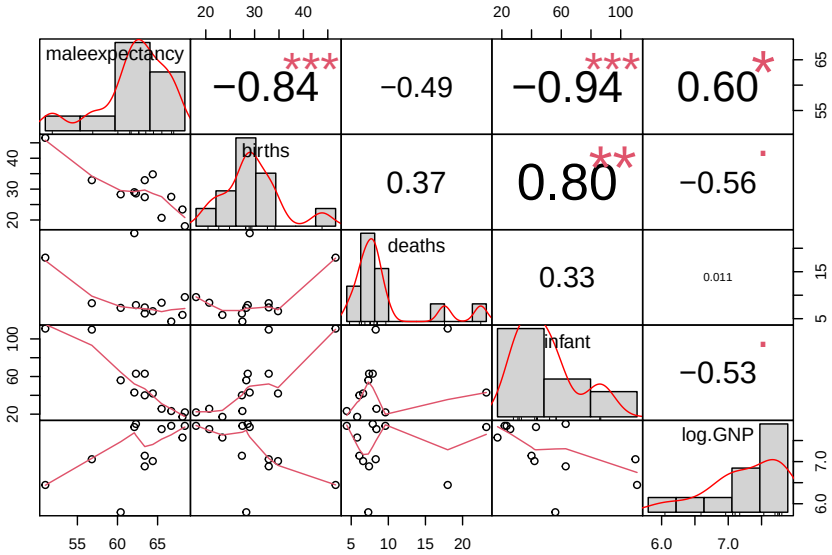
- Correlations
  - Male sub group 1

```
library(PerformanceAnalytics)
chart.Correlation(sub1M)
```



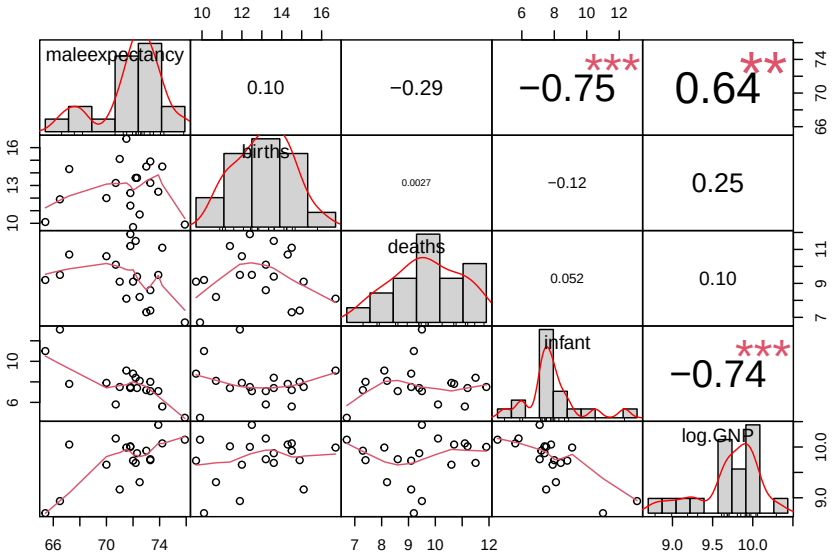
- Male sub group 2

```
chart.Correlation(sub2M)
```



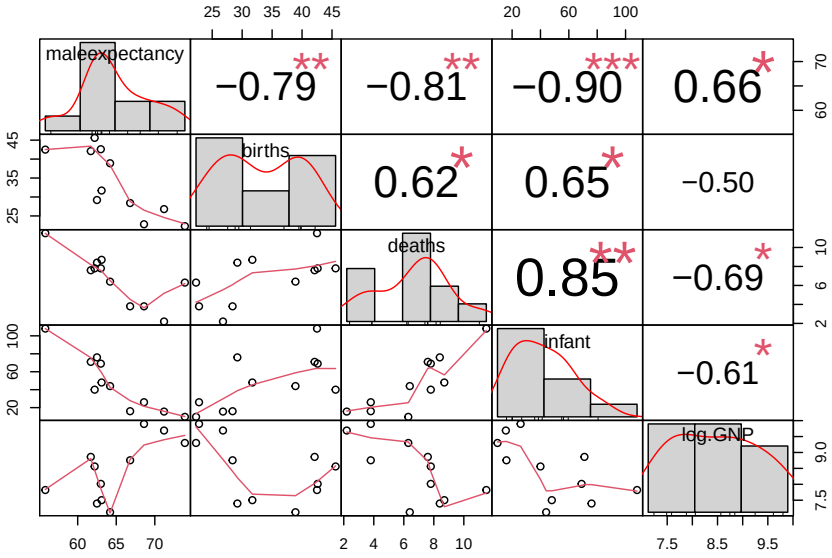
- Male sub group 3

```
chart.Correlation(sub3M)
```



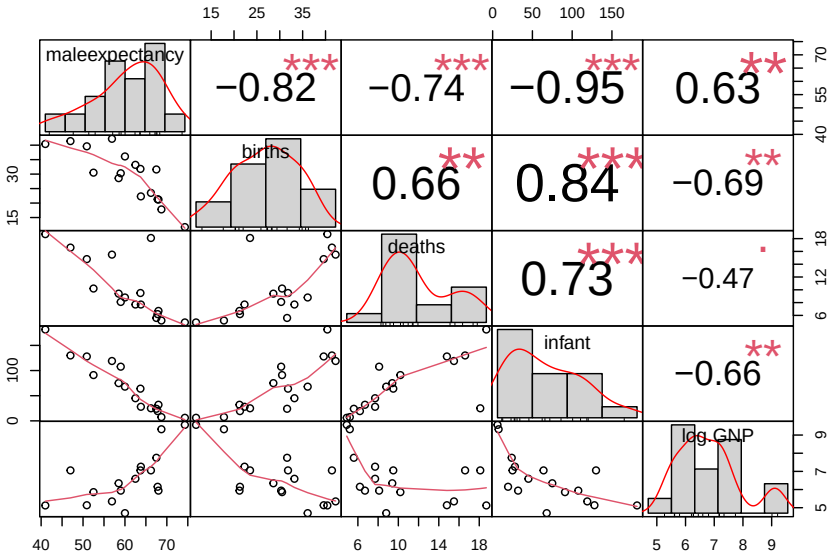
- Male sub group 4

```
chart.Correlation(sub4M)
```



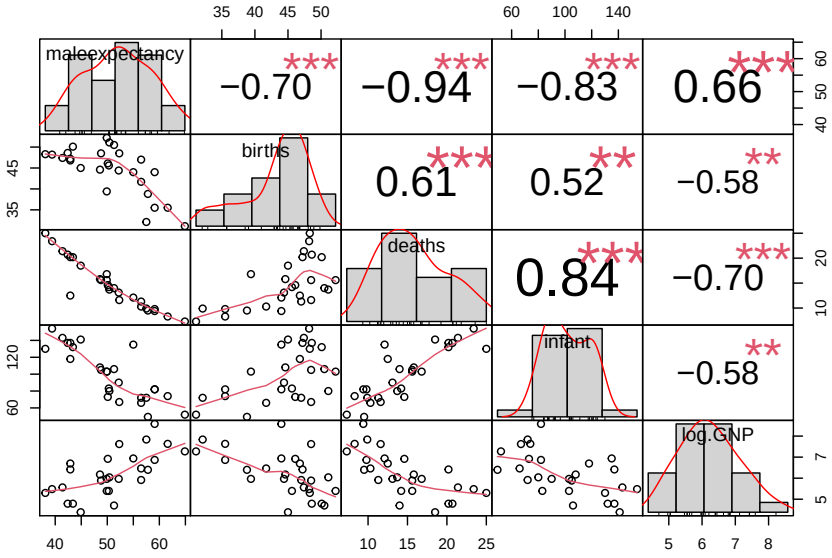
- Male sub group 5

```
chart.Correlation(sub5M)
```



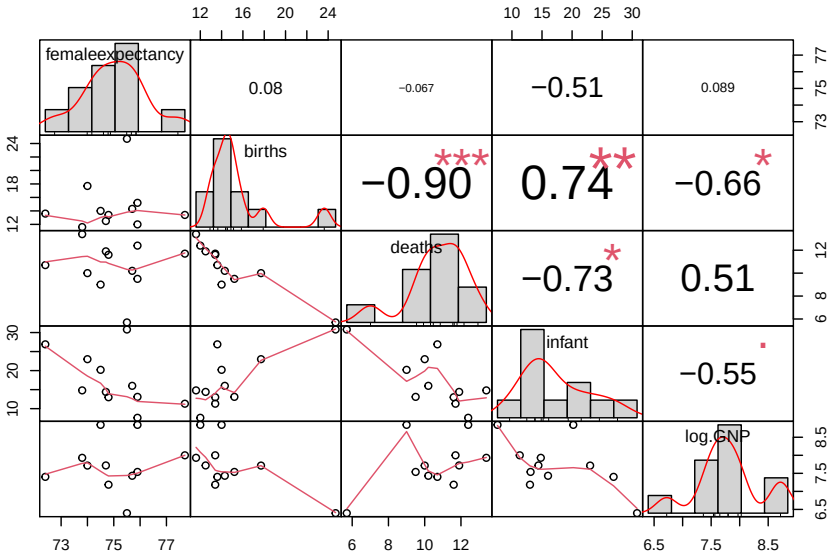
- Male sub group 6

```
chart.Correlation(sub6M)
```



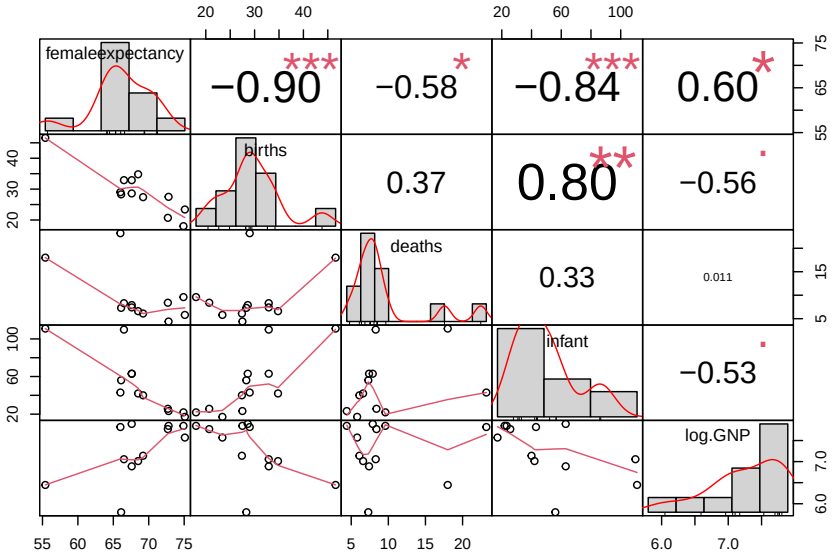
- Female sub group 1

```
chart.Correlation(sub1F)
```



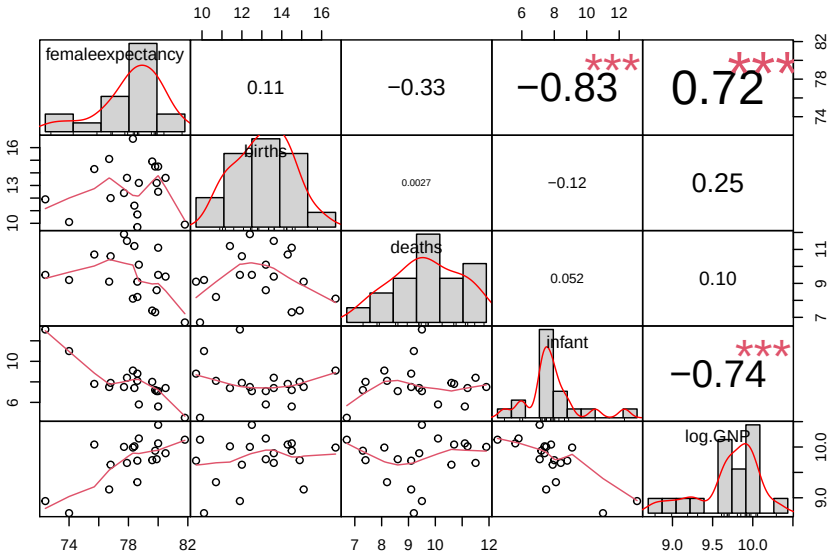
- Female sub group 2

```
chart.Correlation(sub2F)
```



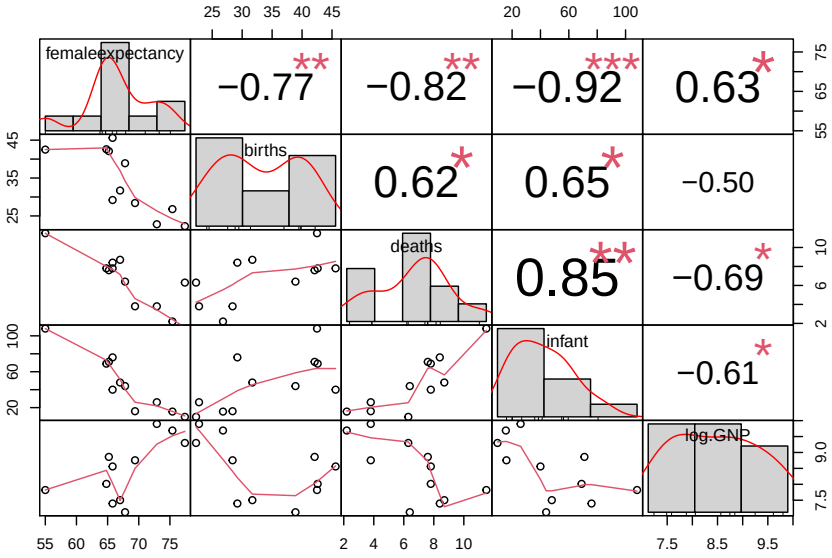
- Female sub group 3

```
chart.Correlation(sub3F)
```



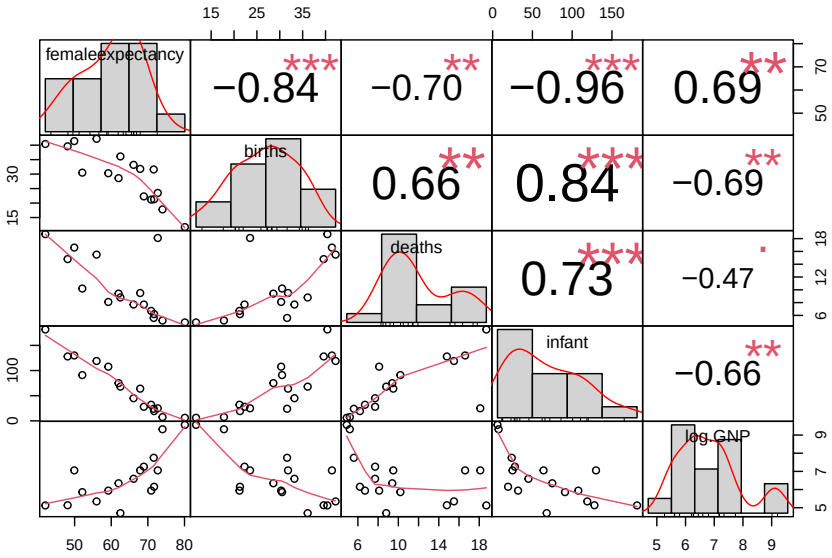
- Female sub group 4

```
chart.Correlation(sub4F)
```



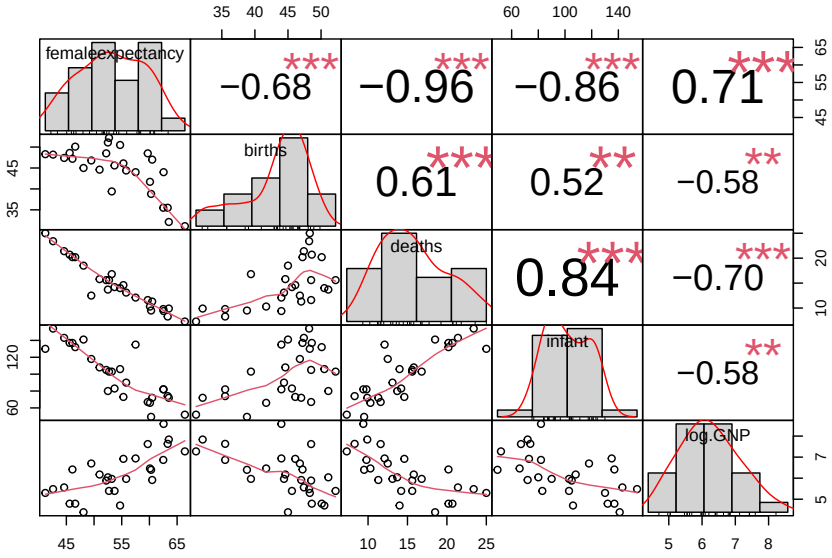
- Female sub group 5

```
chart.Correlation(sub5F)
```



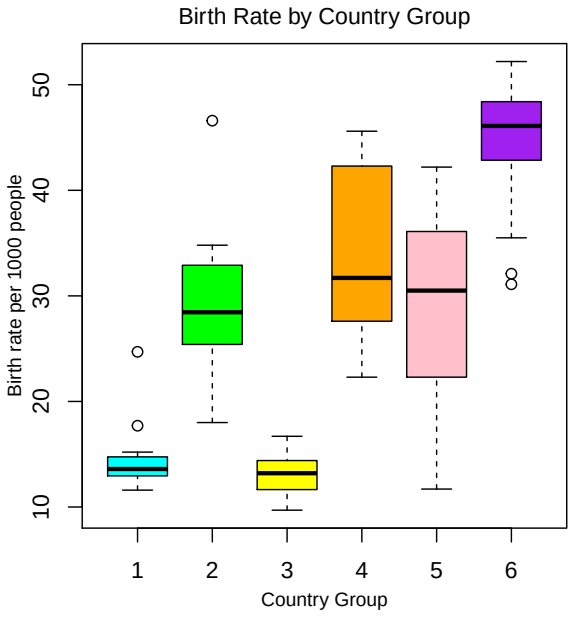
- Female sub group 6

```
chart.Correlation(sub6F)
```



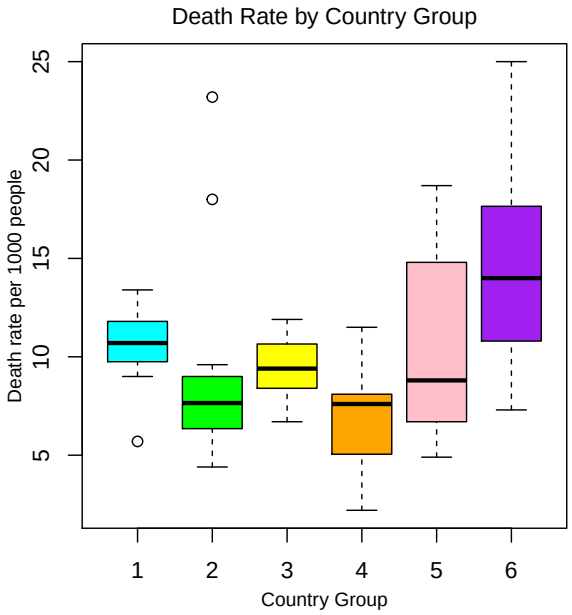
- Other explanatory variable boxplots
  - Birth rate

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(LifeExpectancy$Birth.Rate ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Birth Rate by Country Group',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Birth rate per 1000 people',
      cex = 0.8)
```



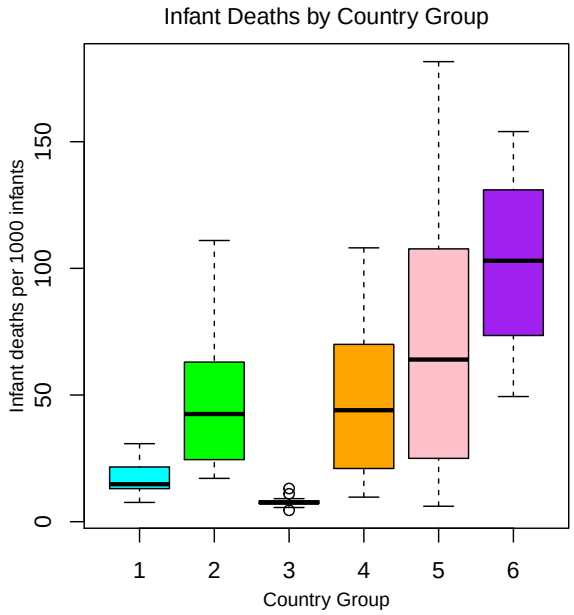
- Death rate

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(LifeExpectancy$Death.rate ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Death Rate by Country Group',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Death rate per 1000 people',
      cex = 0.8)
```



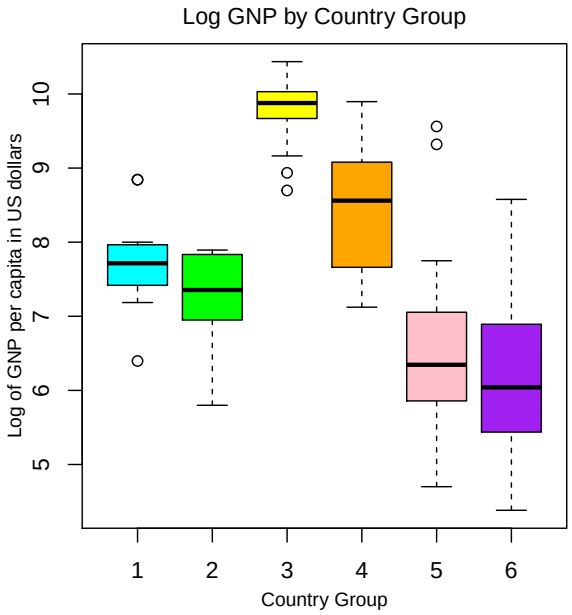
- Infant death

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(LifeExpectancy$Infant.Deaths ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Infant Deaths by Country Group',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Infant deaths per 1000 infants',
      cex = 0.8)
```



- Log GNP

```
par(pty = 's',
    mar = c(3, 1, 2, 0.5))
boxplot(log.GNP ~ LifeExpectancy$Country.Group,
     ann = FALSE,
     col = c('cyan', 'green', 'yellow', 'orange', 'pink', 'purple'))
mtext(side = 3, line = 0.5,
      'Log GNP by Country Group',
      cex = 1)
mtext(side = 1, line = 2,
      'Country Group',
      cex = 0.8)
mtext(side = 2, line = 2,
      'Log of GNP per capita in US dollars',
      cex = 0.8)
```



Simple Linear Models

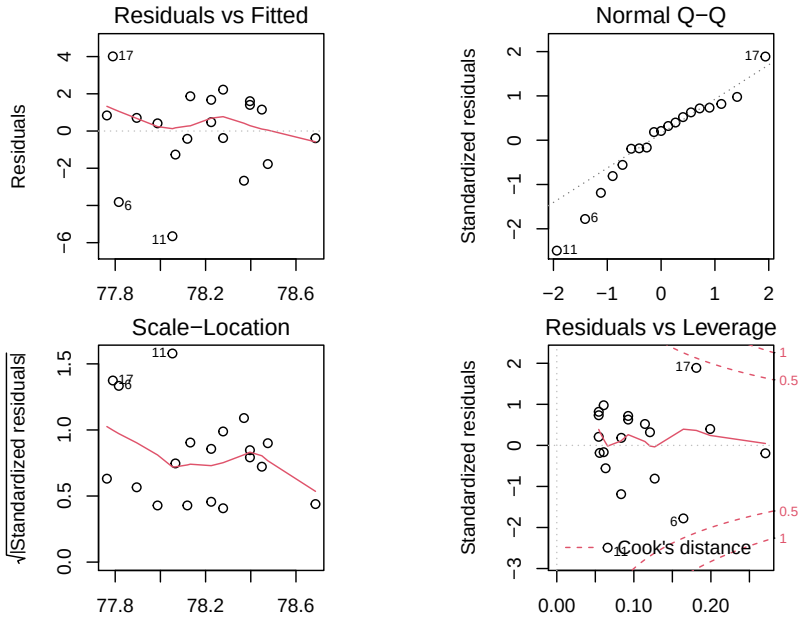
- Sub group 3
  - Birth rates:
    * Creating the linear model and printing the summary

```
birth3.lm <- lm(femaleexpectancy ~ births, data = sub3F)
summary(birth3.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ births, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.6533 -0.8429  0.4752  1.5038  4.0105
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  76.4838     3.7012  20.665 1.75e-13 ***
## births        0.1319     0.2849   0.463    0.649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.347 on 17 degrees of freedom
## Multiple R-squared:  0.01245,    Adjusted R-squared:  -0.04564
## F-statistic: 0.2143 on 1 and 17 DF,  p-value: 0.6493
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(birth3.lm)
```



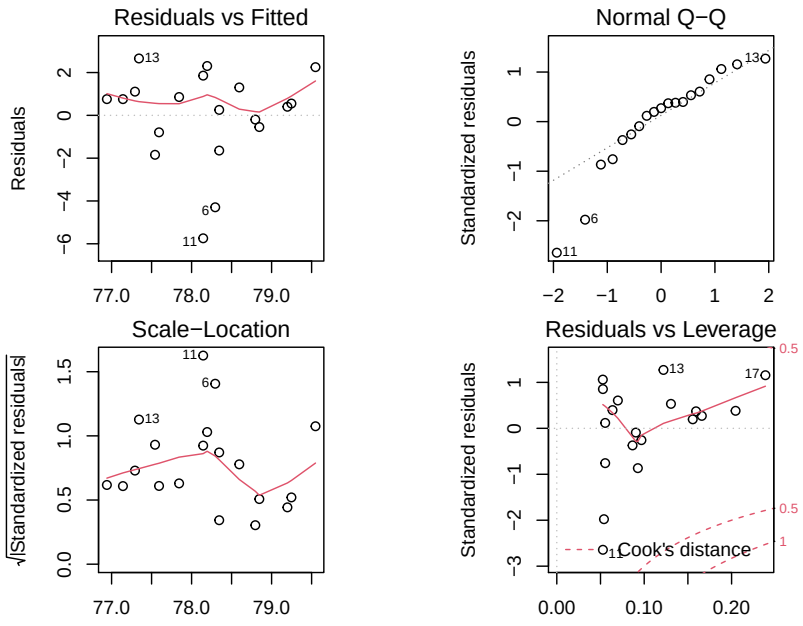
- Death rates:
  - Creating the linear model and printing the summary

```
death3.lm <- lm(femaleexpectancy ~ deaths, data = sub3F)
summary(death3.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ deaths, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7446 -0.6699  0.5522  1.2059  2.6576
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  82.9081     3.3647  24.641 9.65e-15 ***
## deaths       -0.5014     0.3526  -1.422    0.173
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.233 on 17 degrees of freedom
## Multiple R-squared:  0.1063, Adjusted R-squared:  0.05375
## F-statistic: 2.022 on 1 and 17 DF,  p-value: 0.1731
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(death3.lm)
```



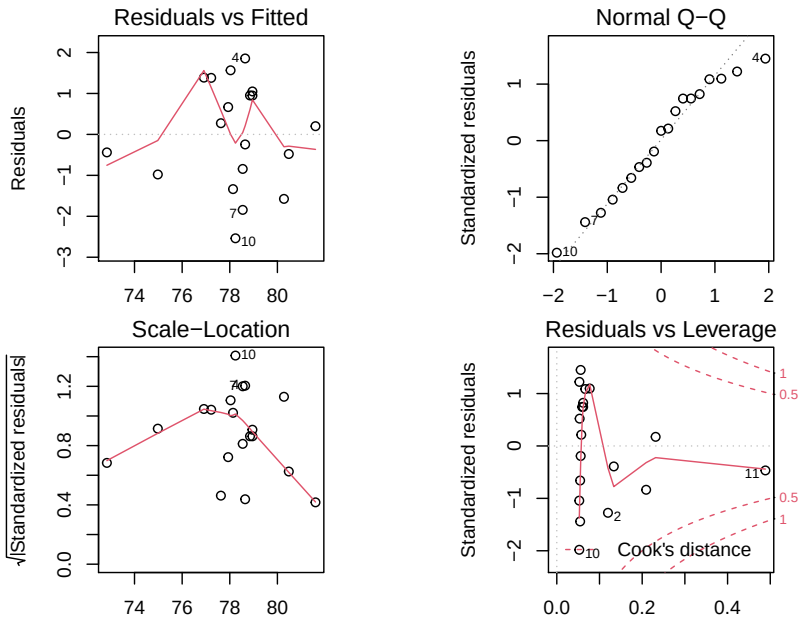
- Infant deaths:
  - Creating the linear model and printing the summary

```
infant3.lm <- lm(femaleexpectancy ~ infant, data = sub3F)
summary(infant3.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ infant, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.5379 -0.9110  0.2008  1.0000  1.8546
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  86.1828     1.3371  64.454  < 2e-16 ***
## infant       -1.0186     0.1658  -6.145 1.08e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.316 on 17 degrees of freedom
## Multiple R-squared:  0.6895, Adjusted R-squared:  0.6713
## F-statistic: 37.76 on 1 and 17 DF,  p-value: 1.08e-05
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(infant3.lm)
```



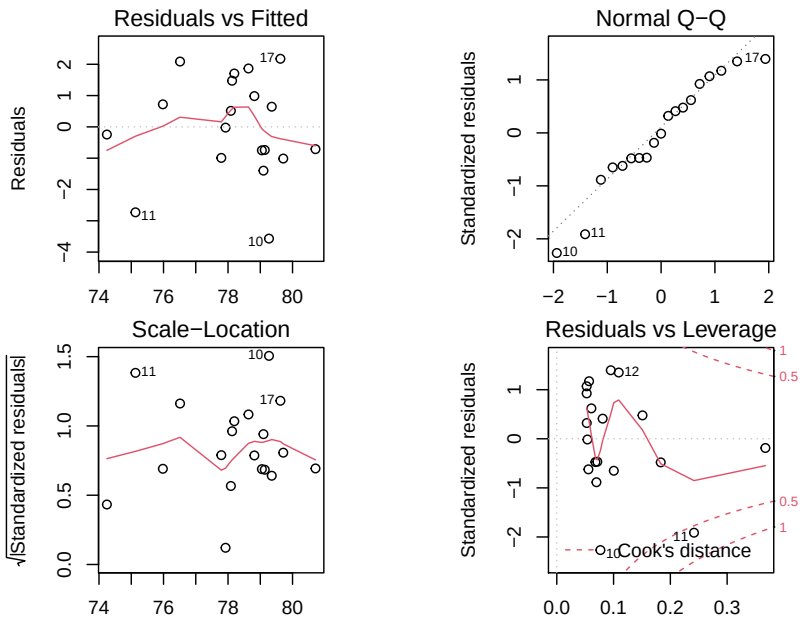
- Log GNP:
  - Creating the linear model and printing the summary

```
logGNP3.lm <- lm(femaleexpectancy ~ log.GNP, data = sub3F)
summary(logGNP3.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ log.GNP, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5696 -0.8703 -0.0232  1.2289  2.1760
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.8813     8.4898   4.933 0.000126 ***
## log.GNP       3.7208     0.8694   4.280 0.000507 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.639 on 17 degrees of freedom
## Multiple R-squared:  0.5186, Adjusted R-squared:  0.4903
## F-statistic: 18.32 on 1 and 17 DF,  p-value: 0.0005067
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(logGNP3.lm)
```



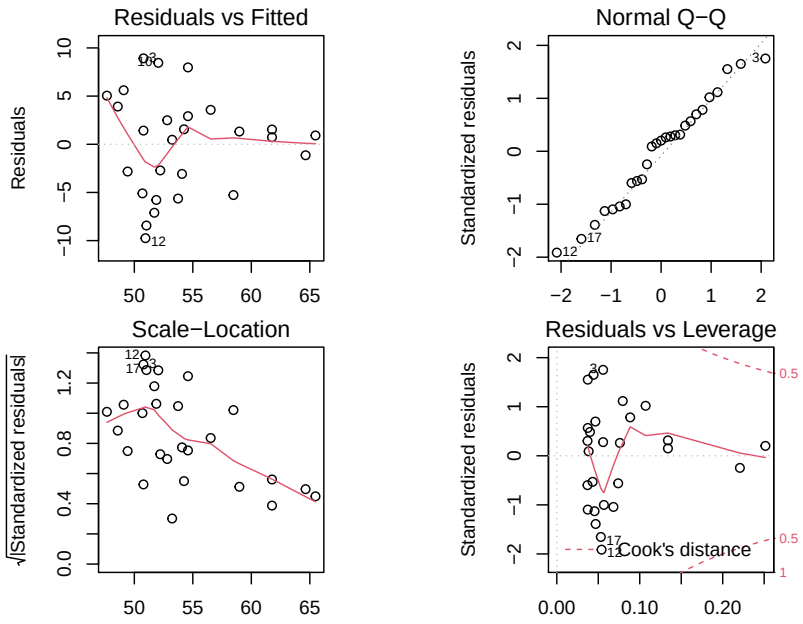
- Sub group 6
  - Birth rates:
    * Creating the linear model and printing the summary

```
birth6.lm <- lm(femaleexpectancy ~ births, data = sub6F)
summary(birth6.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ births, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.7497 -4.0864  0.9143  3.2444  8.9193
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  91.7687     8.1097  11.316 2.50e-11 ***
## births       -0.8451     0.1807  -4.676 8.62e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.24 on 25 degrees of freedom
## Multiple R-squared:  0.4666, Adjusted R-squared:  0.4453
## F-statistic: 21.87 on 1 and 25 DF,  p-value: 8.621e-05
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(birth6.lm)
```



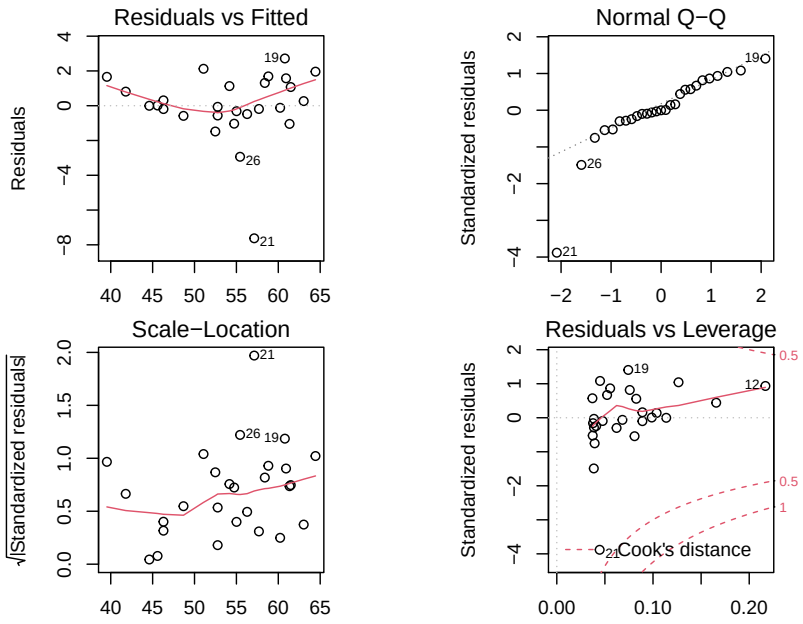
- Death rates:
  - Creating the linear model and printing the summary

```
death6.lm <- lm(femaleexpectancy ~ deaths, data = sub6F)
summary(death6.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ deaths, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6250 -0.5222 -0.0037  1.2191  2.7171
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 74.71114    1.26153   59.22  < 2e-16 ***
## deaths      -1.40689    0.08212  -17.13 2.51e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.01 on 25 degrees of freedom
## Multiple R-squared:  0.9215, Adjusted R-squared:  0.9184
## F-statistic: 293.5 on 1 and 25 DF,  p-value: 2.513e-15
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(death6.lm)
```



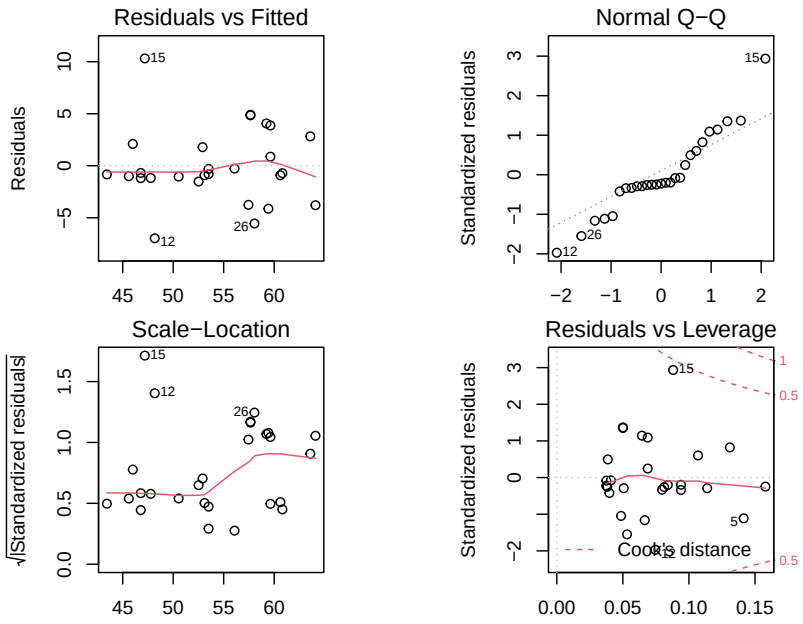
- Infant deaths:
  - Creating the linear model and printing the summary

```
infant6.lm <- lm(femaleexpectancy ~ infant, data = sub6F)
summary(infant6.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ infant, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.9734 -1.1846 -0.8058  1.9429 10.3141
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 73.84803    2.45758  30.049  < 2e-16 ***
## infant      -0.19750    0.02358  -8.374 1.01e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.678 on 25 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7267
## F-statistic: 70.13 on 1 and 25 DF,  p-value: 1.01e-08
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(infant6.lm)
```



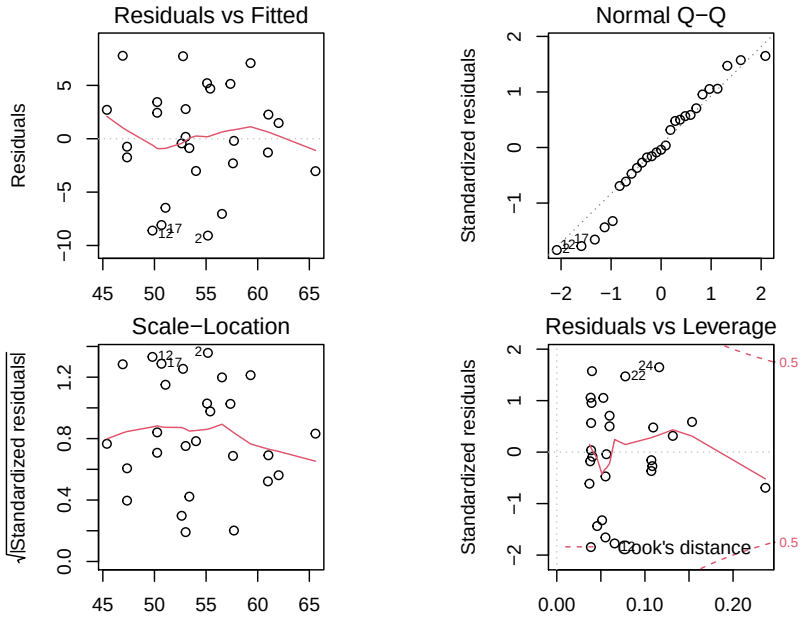
- Log GNP:
  - Creating the linear model and printing the summary

```
logGNP6.lm <- lm(femaleexpectancy ~ log.GNP, data = sub6F)
summary(logGNP6.lm)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ log.GNP, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.0748 -2.6619 -0.1977  3.1081  7.7750
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  24.2873     5.9200   4.103 0.000381 ***
## log.GNP       4.8160     0.9423   5.111  2.8e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.017 on 25 degrees of freedom
## Multiple R-squared:  0.511,  Adjusted R-squared:  0.4914
## F-statistic: 26.12 on 1 and 25 DF,  p-value: 2.797e-05
```

* Diagnostics

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(logGNP6.lm)
```



Multiple Linear Regression

- Sub group 3
  - Brute force
    - Set up

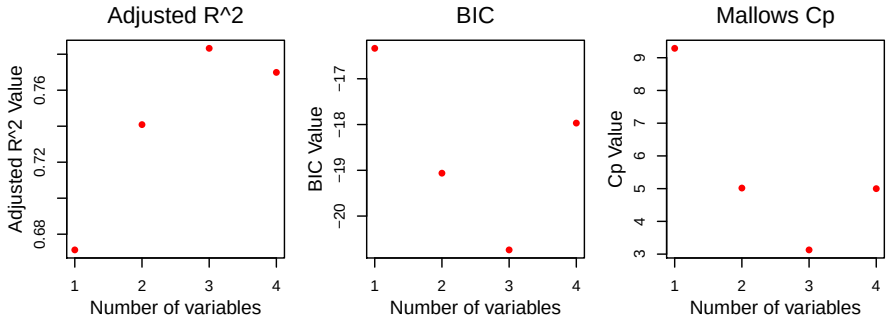
```
library(leaps)
library(knitr)

# Collect best models for each possible number of variables
brute3.models <- regsubsets(femaleexpectancy ~ ., data = sub3F,
                            nbest = 1,
                            nvmax = 4)

# Display the best models
brute3models.summary <- summary(brute3.models)
kable(brute3models.summary$outmat)
```

	births	deaths	infant	log.GNP
1 (1)			*	
2 (1)		*	*	
3 (1)		*	*	*
4 (1)	*	*	*	*

* Choice

```
par(mfrow = c(1, 3),
    mar = c(0.5, 4, 0.5, 0.5),
    pty = 's')
plot(1:4, brute3models.summary$adjr2,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'Adjusted R^2',
      cex = 1)
mtext(side = 1, line = 2.25,
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'Adjusted R^2 Value',
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
plot(1:4, brute3models.summary$bic,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'BIC',
      cex = 1)
mtext(side = 1, line = 2.25,
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'BIC Value',
```

```
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
plot(1:4, brute3models.summary$cp,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'Mallows Cp',
      cex = 1)
mtext(side = 1, line = 2.25,
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'Cp Value',
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
```



- Forward selection
  - Set up

```
# Minimal model
base3.model <- lm(femaleexpectancy ~ 1, data = sub3F)
```

```
* Choice
```

```
forward3.model <- step(base3.model,
                       scope = formula(sub3F),
                       direction = 'forward',
                       trace = 0)
forward3.model$coefficients
```

```
## (Intercept)      infant      deaths     log.GNP
##  71.3192606  -0.6894930  -0.5110037   1.7526050
```

- Backward selection
  - Set up

```
# Set up full model
full3.model <- lm(femaleexpectancy ~ ., data = sub3F)
```

* Choice

```
backward3.model <- step(full3.model,
                        direction = 'backward',
                        trace = 0)
backward3.model$coefficients
```

```
## (Intercept)      deaths      infant     log.GNP
##  71.3192606  -0.5110037  -0.6894930   1.7526050
```

- Stepwise selection
  - Choice

```
stepwise3.model <- step(full3.model,
                        trace = 0)
stepwise3.model$coefficients
```

```
## (Intercept)      deaths      infant     log.GNP
##  71.3192606  -0.5110037  -0.6894930   1.7526050
```

- Summary of chosen model

```
summary(stepwise3.model)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ deaths + infant + log.GNP, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3844 -0.6319  0.1026  0.6531  1.7748
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  71.3193     9.4564   7.542 1.77e-06 ***
## deaths       -0.5110     0.1730  -2.954  0.00985 **
## infant       -0.6895     0.2039  -3.382  0.00411 **
## log.GNP       1.7526     0.8624   2.032  0.06024 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.069 on 15 degrees of freedom
## Multiple R-squared:  0.8194, Adjusted R-squared:  0.7833
## F-statistic: 22.68 on 3 and 15 DF,  p-value: 7.918e-06
```

- Diagnostics
  - Multi-collinearity

```
library(car)
vif(stepwise3.model)
```

```
##   deaths   infant  log.GNP
## 1.050671 2.294835 2.314046
```

* 2 vs 3 model

```
model32var <- lm(femaleexpectancy ~ deaths + infant, data = sub3F)
anova(model32var, stepwise3.model)
```

```
## Analysis of Variance Table
##
## Model 1: femaleexpectancy ~ deaths + infant
## Model 2: femaleexpectancy ~ deaths + infant + log.GNP
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     16 21.843
## 2     15 17.127  1    4.7155 4.1298 0.06024 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

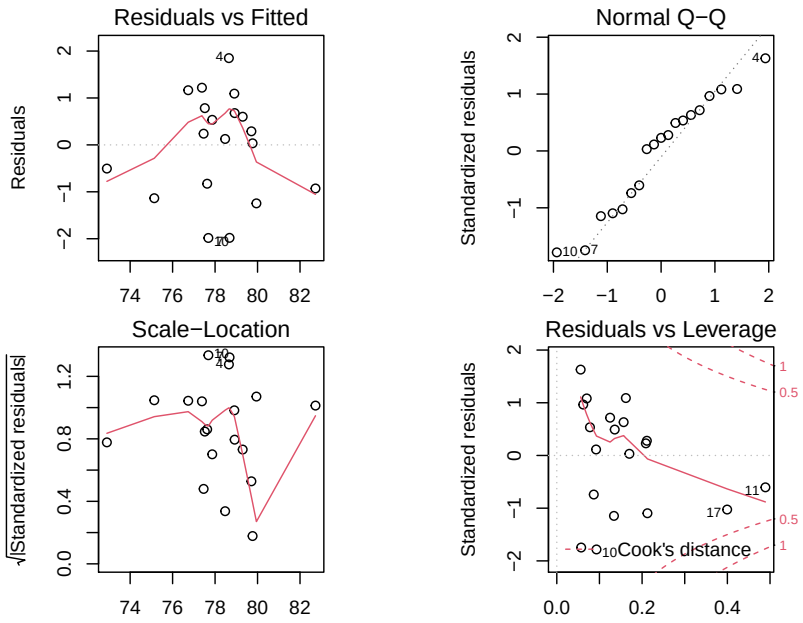
* Two variable summary

```
summary(model32var)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ deaths + infant, data = sub3F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9839 -0.8782  0.2391  0.7304  1.8492
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  90.1516     2.0588  43.788  < 2e-16 ***
## deaths       -0.4359     0.1848  -2.359   0.0313 *
## infant       -1.0004     0.1474  -6.788 4.36e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.168 on 16 degrees of freedom
## Multiple R-squared:  0.7697, Adjusted R-squared:  0.7409
## F-statistic: 26.73 on 2 and 16 DF,  p-value: 7.922e-06
```

* Diagnostic plots

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(model32var)
```



- Sub group 6
  - Brute force
    - Set up

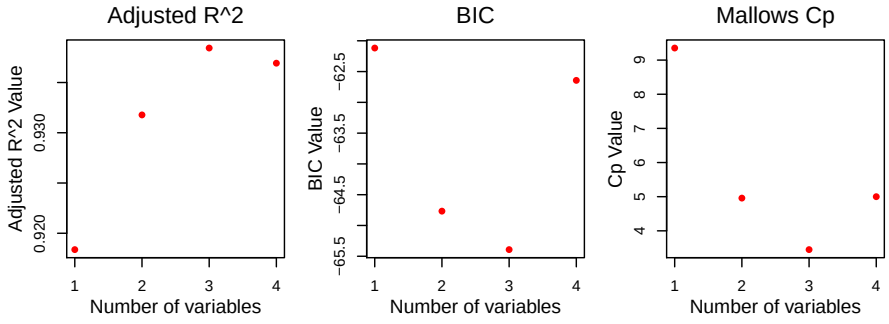
```
library(leaps)
library(knitr)

# Collect best models for each possible number of variables
brute6.models <- regsubsets(femaleexpectancy ~ ., data = sub6F,
                            nbest = 1,
                            nvmax = 4)

# Display the best models
brute6models.summary <- summary(brute6.models)
kable(brute6models.summary$outmat)
```

	births	deaths	infant	log.GNP
1 (1)		*		
2 (1)	*	*		
3 (1)	*	*	*	
4 (1)	*	*	*	*

* Choice

```
par(mfrow = c(1, 3),
    mar = c(0.5, 4, 0.5, 0.5),
    pty = 's')
plot(1:4, brute6models.summary$adjr2,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'Adjusted R^2',
      cex = 1)
mtext(side = 1, line = 2.25,
```

```
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'Adjusted R^2 Value',
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
plot(1:4, brute6models.summary$bic,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'BIC',
      cex = 1)
mtext(side = 1, line = 2.25,
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'BIC Value',
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
plot(1:4, brute6models.summary$cp,
     col = 'red',
     pch = 16,
     ann = FALSE,
     xaxt = 'n')
mtext(side = 3, line = 0.75,
      'Mallows Cp',
      cex = 1)
mtext(side = 1, line = 2.25,
      'Number of variables',
      cex = 0.8)
mtext(side = 2, line = 2.25,
      'Cp Value',
      cex = 0.8)
axis(1, labels = c("1", "2", "3", "4"), seq(1, 4, 1))
```



- Forward selection
  - Set up

```
# Minimal model
base6.model <- lm(femaleexpectancy ~ 1, data = sub6F)
```

```
* Choice
```

```
forward6.model <- step(base6.model,
                       scope = formula(sub6F),
                       direction = 'forward',
                       trace = 0)
forward6.model$coefficients
```

```
## (Intercept)      deaths      births      infant
## 82.02826424 -1.05961680 -0.19054197 -0.03919124
```

- Backward selection
  - Set up

```
# Set up full model
full6.model <- lm(femaleexpectancy ~ ., data = sub6F)
```

```
* Choice
```

```
backward6.model <- step(full6.model,
                        direction = 'backward',
                        trace = 0)
backward6.model$coefficients
```

```
## (Intercept)      births      deaths      infant
## 82.02826424 -0.19054197 -1.05961680 -0.03919124
```

- Stepwise selection
  - Choice

```
stepwise6.model <- step(full6.model,
                        trace = 0)
stepwise6.model$coefficients
```

```
## (Intercept)      births      deaths      infant
## 82.02826424 -0.19054197 -1.05961680 -0.03919124
```

- Summary of chosen model

```
summary(stepwise6.model)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ births + deaths + infant, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.7411 -0.3145  0.0707  0.8670  2.2226
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.02826    2.82713  29.015  < 2e-16 ***
## births      -0.19054    0.07590  -2.510   0.0195 *
## deaths      -1.05962    0.14152  -7.487 1.31e-07 ***
## infant      -0.03919    0.02067  -1.896   0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.746 on 23 degrees of freedom
## Multiple R-squared:  0.9455, Adjusted R-squared:  0.9384
## F-statistic: 133.1 on 3 and 23 DF,  p-value: 1.124e-14
```

- Diagnostics
  - Multi-collinearity

```
library(car)
vif(stepwise6.model)
```

```
##   births   deaths   infant
## 1.589392 3.937813 3.410428
```

* 2 vs 3 model

```
model62var <- lm(femaleexpectancy ~ births + deaths, data = sub6F)
anova(model62var, stepwise6.model)
```

```
## Analysis of Variance Table
##
## Model 1: femaleexpectancy ~ births + deaths
## Model 2: femaleexpectancy ~ births + deaths + infant
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     24 81.030
## 2     23 70.078  1    10.952 3.5946 0.0706 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

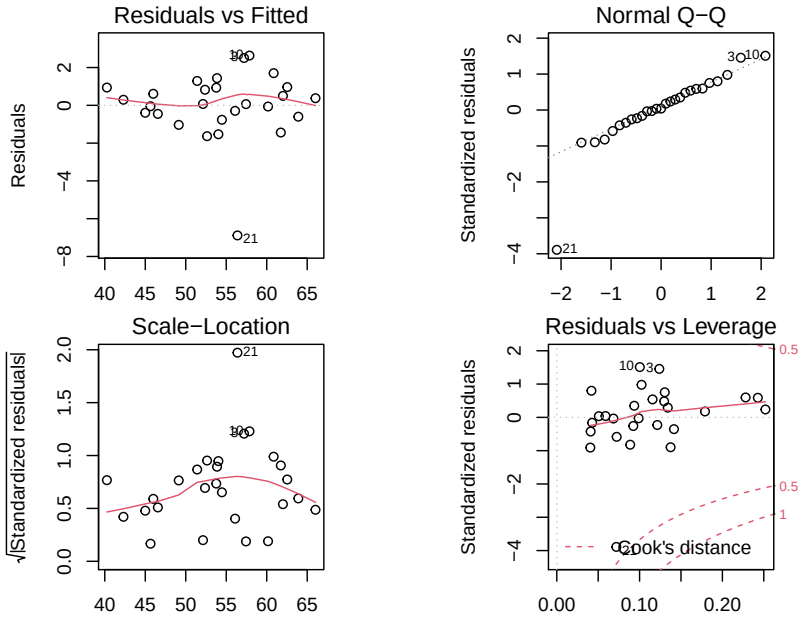
- Two variable summary

```
summary(model62var)
```

```
##
## Call:
## lm(formula = femaleexpectancy ~ births + deaths, data = sub6F)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8862 -0.5276  0.0713  0.9369  2.6324
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 81.31112    2.94927  27.570  < 2e-16 ***
## births      -0.19421    0.07987  -2.431   0.0229 *
## deaths      -1.26687    0.09462 -13.390 1.25e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.837 on 24 degrees of freedom
## Multiple R-squared:  0.937,  Adjusted R-squared:  0.9318
## F-statistic: 178.5 on 2 and 24 DF,  p-value: 3.892e-15
```

* Diagnostic plots

```
par(mfrow = c(2, 2),
    mar = c(2, 0.5, 1.75, 0.1),
    pty = 's')
plot(model62var)
```



Predictive Comparision

- Sub group 3
  - PRESS

```
library(DAAG)
press(death6.lm)
```

```
## [1] 114.0283
```

```
press(model32var)
```

```
## [1] 28.89231
```

- Sub group 3
  - PRESS

```
library(DAAG)
press(death6.lm)
```

```
## [1] 114.0283
```

```
press(model62var)
```

```
## [1] 97.59061
```