Assignment 2 - Exercise 2

Exercise 2. Birthweights

This exercise explores the data set `Birthweight.csv` which contains information on new born babies and their parents. A first examination reveals the 16 variables with 42 observations:

```
birthweight <- read.csv("data/Birthweight.csv")
str(birthweight)
```

```
## 'data.frame':    42 obs. of  16 variables:
##  $ ID         : int  1360 1016 462 1187 553 1636 820 1191 1081 822 ...
##  $ Length     : int  56 53 58 53 54 51 52 53 54 50 ...
##  $ Birthweight: num  4.55 4.32 4.1 4.07 3.94 3.93 3.77 3.65 3.63 3.42 ...
##  $ Headcirc   : int  34 36 39 38 37 38 34 33 38 35 ...
##  $ Gestation  : int  44 40 41 44 42 38 40 42 38 38 ...
##  $ smoker     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mage       : int  20 19 35 20 24 29 24 21 18 20 ...
##  $ mnocig     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mheight    : int  162 171 172 174 175 165 157 165 172 157 ...
##  $ mppwt      : int  57 62 58 68 66 61 50 61 50 48 ...
##  $ fage       : int  23 19 31 26 30 31 31 21 20 22 ...
##  $ fedyrs     : int  10 12 16 14 12 16 16 10 12 14 ...
##  $ fnocig     : int  35 0 25 25 0 0 0 25 7 0 ...
##  $ fheight    : int  179 183 185 189 184 180 173 185 172 179 ...
##  $ lowbwt     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mage35     : int  0 0 1 0 0 0 0 0 0 0 ...
```

```
head(birthweight)
```

```
##      ID Length Birthweight Headcirc Gestation smoker mage mnocig mheight mppwt
## 1 1360     56        4.55       34        44      0   20      0     162    57
## 2 1016     53        4.32       36        40      0   19      0     171    62
## 3  462     58        4.10       39        41      0   35      0     172    58
## 4 1187     53        4.07       38        44      0   20      0     174    68
## 5  553     54        3.94       37        42      0   24      0     175    66
## 6 1636     51        3.93       38        38      0   29      0     165    61
##   fage fedyrs fnocig fheight lowbwt mage35
## 1   23     10     35     179      0      0
## 2   19     12      0     183      0      0
## 3   31     16     25     185      0      1
## 4   26     14     25     189      0      0
## 5   30     12      0     184      0      0
## 6   31     16      0     180      0      0
```

For the first part of the analysis, the variables `ID`, `smoker`, `lowbwt` and `mage35` are disregarded, the column `Birthweight` is selected as a response variable, while the other 11 variables are considered potential predictors.

```
birthweight1 <- birthweight
birthweight1$ID <- NULL; birthweight1$smoker <- NULL
birthweight1$lowbwt <- NULL; birthweight1$mage35 <- NULL
```

a) The explanatory variables `Length`, `Headcirk`, `Gestation`, `mage`, `mnosig`, `mheight`, `mppwt`, `fage`, `fedyrs`, `fnosig`, and `fheight` are to be examined for potential (leverage) points and, in case such are found, it is to be verified whether these are influence points by examining the effect of their removal. A series of box plots for all variables provides a first glance on the presence of such points.

```
par(mfrow=c(3, 4), mar=c(1,2,2,2))
for (i in 1:12) {
  boxplot(birthweight1[,i], main=names(birthweight1)[i])
}
```
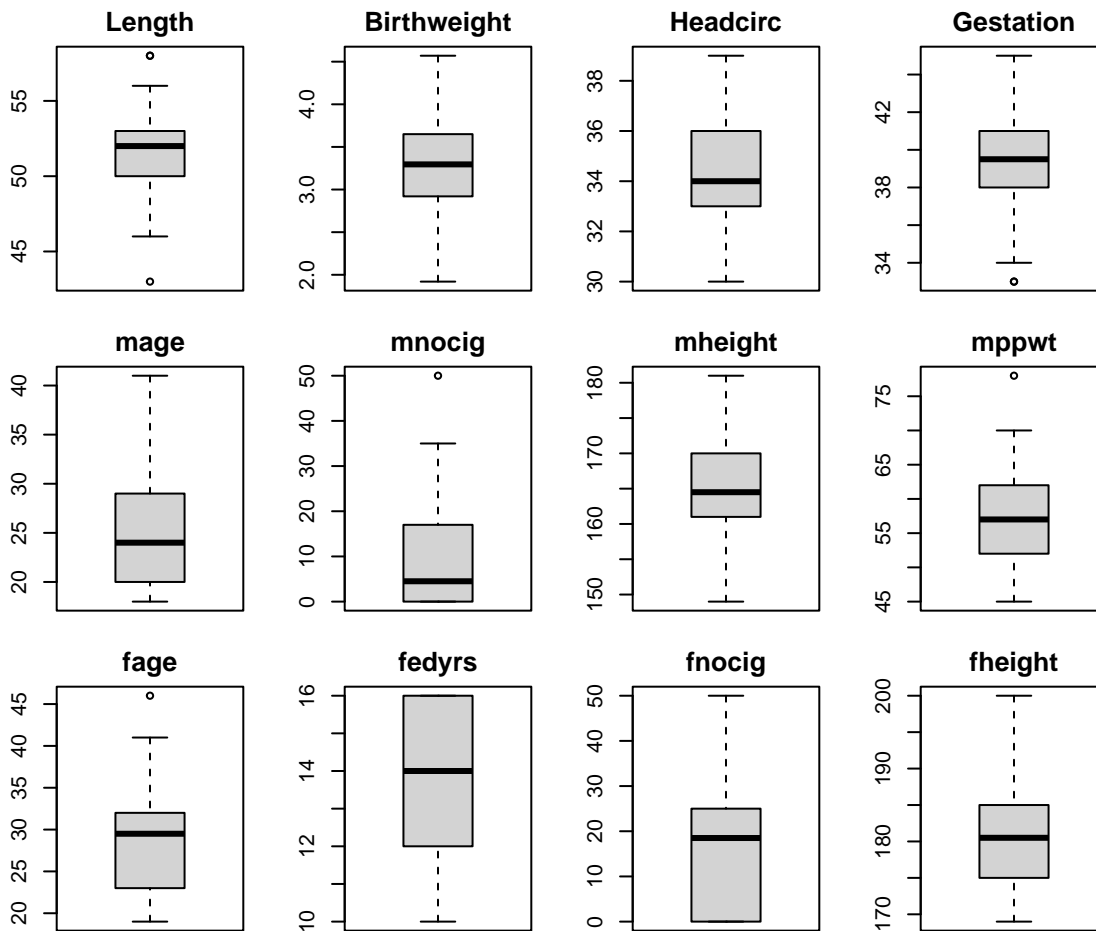

The box plots reveal visible outliers in the variables `Length`, `Gestation`, `mnocig`, `mppwt` and `fage`. These predictors are therefore selected for more in-depth analysis through a scatter plot of each against the response variable `Birthweight`.

```
predictors = c("Length", "Gestation", "mnocig", "mppwt", "fage")
par(mfrow=c(5, 2), mar=c(1.5,2,2,2))
for (i in 1:5) {
  plot(birthweight1[,predictors[i]], birthweight1$Birthweight,
       main=predictors[i], xlab=predictors[i], ylab="Birthweight")
  predlm <- lm(birthweight1$Birthweight ~ birthweight1[,predictors[i]])
  qqnorm(resid(predlm), main=predictors[i])
}
```