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 ...
```

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 explanatory variables.

```
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 qualitative investigation through box plots and scatter plots is possible, but a more compact and quantitative approach is to calculate Cook's distance for each observation within each explanatory variable. A Cook's distance of a potential point larger than one provides evidence that this point is in fact an influence point. The following code iterates through all predictors and prints the maximum Cook's distance among the observations.

```
for (i in 1:length(birthweight1)) {
  if (names(birthweight1)[i] == "Birthweight") next
  bw_model <- lm(Birthweight~birthweight1[,i], data=birthweight1)
  cdist <- cooks.distance(bw_model)
  print(paste(names(birthweight1)[i], max(cdist)))
}
```

```
## [1] "Length 0.349539386042009"
```

```
## [1] "Headcirc 0.131036624474646"
## [1] "Gestation 0.0945072417199716"
## [1] "mage 0.693221167012135"
## [1] "mnocig 0.0759446000944725"
## [1] "mheight 0.109325335314728"
## [1] "mppwt 0.133331605139495"
## [1] "fage 0.172911969691082"
## [1] "fedyrs 0.268528640018302"
## [1] "fnocig 0.132144446138219"
## [1] "fheight 0.339854927584761"
```

None of the resulting Cook's distances are sufficiently large to conclude that the points are influence points.

Another potential problem to be addressed is the presence of collinearity between the explanatory variables. A preliminary investigation can be conducted by analysing the variance inflation factors (VIF) of a model including the 11 predictors.

```
bw_model <- lm(Birthweight~., data=birthweight1)
library(car); vif(bw_model)
```

```
## Warning: package 'car' was built under R version 4.3.3
```

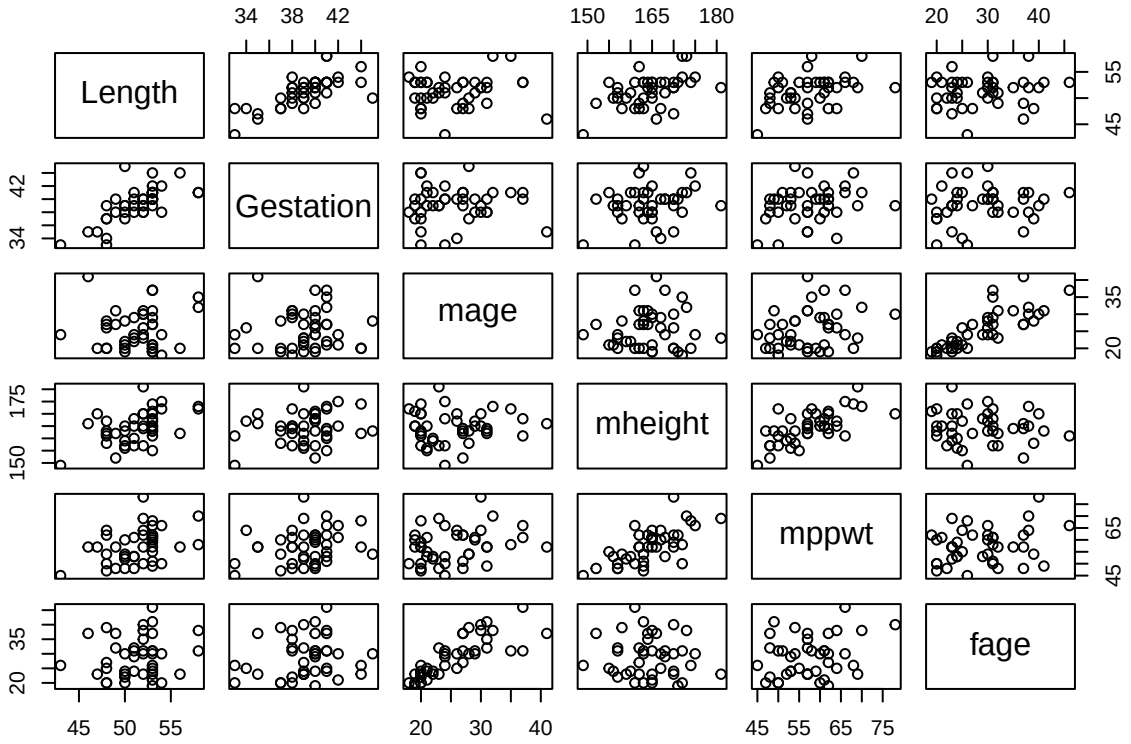
```
## Loading required package: carData
```

```
## Warning: package 'carData' was built under R version 4.3.3
```

```
##    Length  Headcirc Gestation      mage    mnocig   mheight     mppwt      fage
##  3.115295  1.835970  2.508132  4.028952  1.416515  3.110390  2.380129  4.517598
##    fedyrs    fnocig   fheight
##  1.614641  1.706549  1.619061
```

In general, the resulting VIF values are not large enough to indicate collinearity. Nonetheless, two variables, `mage` and `fage`, have VIF values close to 5. Since the above analysis does not provide information about the collinear groups of variables which these two belong to, a more detailed examination can be performed on a larger selection of variables, for instance the ones with a VIF larger than 2: `Length`, `Gestation`, `mage`, `mheight`, `mppwt` and `fage`. To illustrate visually in how far these two variables are correlated, a pairwise scatter plot is created.

```
pairs(birthweight1[,c("Length", "Gestation", "mage", "mheight", "mppwt", "fage")])
```



Linear correlations are observed in pairs like `Length` and `Gestation`, `mage` and `fage`, and `mheight` and `mppwt`. The first one is indicative of the fact that the length of the gestation period is correlated to the baby's growth. The second one is expected, as the mother's and father's age commonly do not differ a lot. The third one is also logical, as the mother's height and pre-pregnancy weight are likely to be correlated. The presence of collinearity is not a problem in this case, as the VIF values are not large enough to indicate that the estimates of the coefficients are unstable.

b) To reduce the number of explanatory variables, the step-down method is applied by iteratively analysing the significance of the influence of all independent variables on `Birthweight` and removing the least significant one, then repeating the process until all variables have a significant effect.

```
model_summary <- summary(bw_model)
p_values <- model_summary$coefficients[,4]
birthweight2 <- birthweight1
while (max(p_values) > 0.05) {
  max_p_name <- names(which.max(p_values))
  birthweight2 <- birthweight2[, names(birthweight2) != max_p_name]
  bw_model <- lm(Birthweight~., data=birthweight2)
  model_summary <- summary(bw_model)
  p_values <- model_summary$coefficients[,4]
}
model_summary
```

```
##
## Call:
## lm(formula = Birthweight ~ ., data = birthweight2)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -0.82889 -0.24763 -0.05136  0.25136  0.74352
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.44799    0.93936  -5.800 9.83e-07 ***
## Headcirc     0.11977    0.02449   4.891 1.77e-05 ***
## Gestation    0.11782    0.02223   5.299 4.85e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3441 on 39 degrees of freedom
## Multiple R-squared:  0.6911, Adjusted R-squared:  0.6753
## F-statistic: 43.63 on 2 and 39 DF,  p-value: 1.124e-10
```

The final model is reduced to only two explanatory variables: `Headcirc` and `Gestation`, both of which exhibit a p-value below 0.05. Since the head circumference can be considered indicative of the baby's size, and the length of the gestation period clearly determines how much the baby grows in size and weight prior to birth, the influence of the two variables seems logical.

c) For the next exercise, the average of each predictor value from the reduced model is taken and used as a new observation, for which the 95% confidence and prediction intervals for the response variable `Birthweight` are calculated.

```
averages <- data.frame(Headcirc=mean(birthweight2$Headcirc), Gestation=mean(birthweight2$Gestation))
averages
```

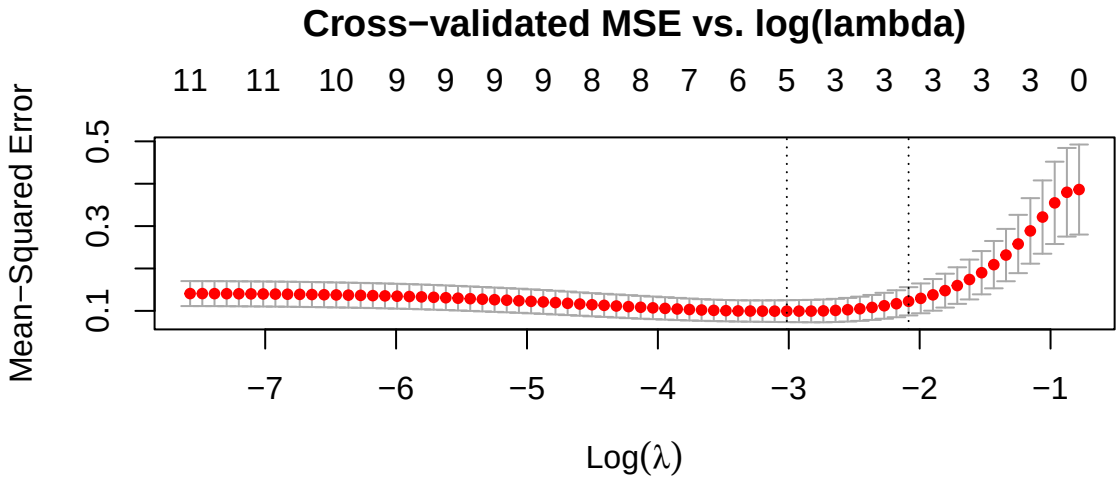
```
##   Headcirc Gestation
## 1 34.59524  39.19048
```

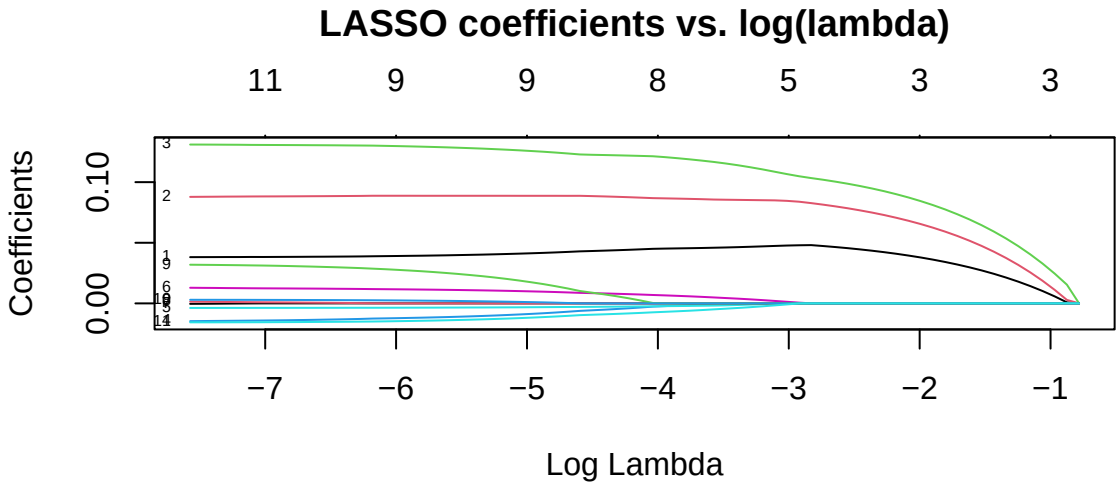
```
stepdown_ci <- predict(bw_model, newdata=averages, interval="confidence", level=0.95)
stepdown_pi <- predict(bw_model, newdata=averages, interval="prediction", level=0.95)
```

The resulting fitted value for `Birthweight` is 3.3128571. As expected, the prediction interval [2.6085627, 4.0171515] is wider than the confidence interval [3.2054533, 3.420261], as it encompasses individual observations instead of observation means and thus accounts for the error in these observations as well.

d) As an alternative to the step-down method, the LASSO method is applied to the original model to reduce the number of explanatory variables. The `cv.glmnet` function from the `glmnet` package is used to select the optimal value of the tuning parameter λ by cross-validation. Just like in b), the reduction starts with the filtered data set `birthweight1`. To train the model, two thirds of the data points are randomly sampled and one third is reserved for subsequent testing.

```
par(mfrow=c(2,1), mar=c(4,4,6,1))
library(glmnet)
set.seed(123) # For reproducibility
x_df <- birthweight1[, names(birthweight1) != 'Birthweight']
x <- as.matrix(x_df)
y <- as.double(birthweight1$Birthweight)
train <- sample(1:nrow(x), 0.67*nrow(x))
x_train <- x[train,]; y_train <- y[train]
x_test <- x[-train,]; y_test <- y[-train]; x_test_df <- x_df[-train,]
lasso_model <- glmnet(x_train, y_train, alpha=1)
cv_lasso <- cv.glmnet(x_train, y_train, alpha=1, type.measure="mse")
plot(cv_lasso, main="Cross-validated MSE vs. log(lambda)")
plot(cv_lasso$glmnet.fit, xvar="lambda", label=T, main="LASSO coefficients vs. log(lambda)")
```






```
lambda_min <- cv_lasso$lambda.min; lambda_1se <- cv_lasso$lambda.1se
coef(cv_lasso, s=cv_lasso$lambda.min)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                         s1
## (Intercept) -6.4208411623
## Length       0.0477099071
## Headcirc     0.0845937946
## Gestation    0.1066858728
## mage         .
## mnocig       .
## mheight      0.0013022375
## mppwt        .
## fage         .
## fedyrs       .
## fnocig       .
## fheight     -0.0008632859
```

```
lasso_pred <- predict(lasso_model, s=lambda_min, newx=x_test)
mse_lasso <- mean((lasso_pred-y_test)^2)
stepdown_pred <- predict(bw_model, newdata=x_test_df)
mse_stepdown <- mean((stepdown_pred-y_test)^2)
```

The plot of the cross-validated MSE shows for which value of the free parameter λ the penalty term $\lambda P(\beta)$ compensates the RSS term $\frac{1}{N}||Y - X\beta||^2$ the best. The left vertical line indicates the value of `lambda.min`, which is the value for which the cross-validated error is minimised. The right vertical line marks the value of `lambda.1se`, which is the largest value of λ such that the error is within one standard error of the minimum. The second plot depicts the shrinkage of the coefficients with increasing λ. Since this shrinkage means that the model is becoming simpler, `lambda.1se` is relevant for finding the optimal trade-off between a minimum MSE and a most simplified model.

When observing the coefficients of the simplified model, it is noticeable that `Headcirc` and `Gestation` are present, just like in the reduced model from the step-down method. The rest of the remaining coefficients in the LASSO model, `Length`, `mheight` and `fheight` are close to zero but still present. Both the LASSO-optimised model and the model from b) are tested with the remaining subset of one third of the data points. This results in a MSE of 0.2206134 for the LASSO method and a MSE of 0.1481082 for the step-down method. Surprisingly, the LASSO method does not seem to perform better than the step-down method in this case.