Assignment 2

Group 3: *Sophie Engels, Boyan Mihaylov and Jonas Schäfer*
Experimental Design and Data Analysis, VU Amsterdam (2024)

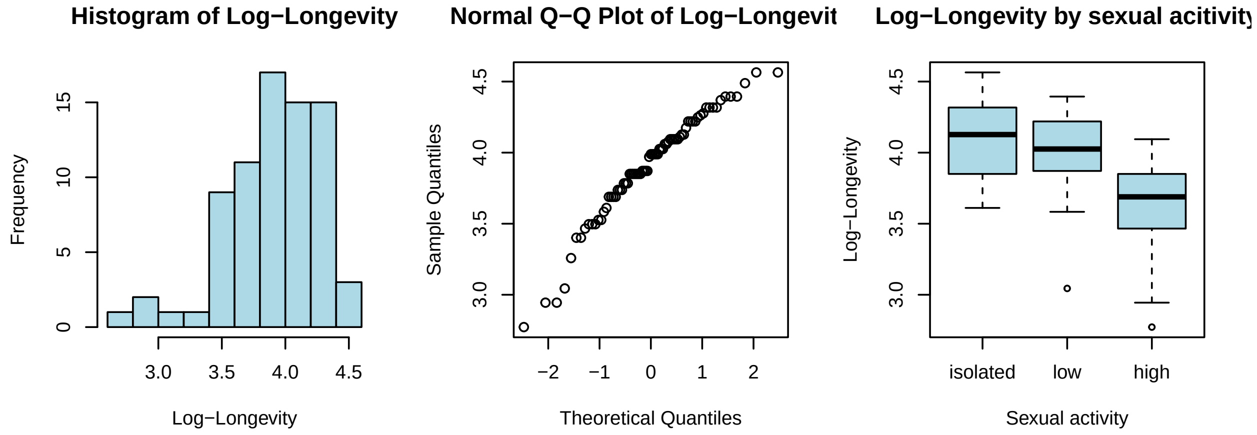
Exercise 1. Fruit Flies

In this section, we investigate the effect of sexual activity on longevity of fruit flies. First, we add a column called `loglongevity` to the data frame, containing the logarithm of the number of days until death (`longevity` column). We will use this as our response variable and further discuss this decision later in the report. We also use the *isolated* group as the baseline for the `activity` factor.

```
ff <- read.table("data/fruitflies.txt", header=TRUE)
ff$loglongevity <- log(ff$longevity, base=exp(1))
ff$activity <- factor(ff$activity, levels=c("isolated", "low", "high"))
summary(ff)
```

a) We first investigate the data through plots, checking for normality and gathering information whether sexual activity may influences longevity. We also perform a Levene test to check for homogeneity of variances between the three sexual activity groups.

```
par(mfrow=c(1,3))
hist(ff$loglongevity, main="Histogram of Log-Longevity", xlab="Log-Longevity",
     col="lightblue")
qqnorm(ff$loglongevity, main="Normal Q-Q Plot of Log-Longevity")
boxplot(loglongevity ~ activity, data=ff, main="Log-Longevity by sexual acitivity",
        col="lightblue", xlab="Sexual activity", ylab="Log-Longevity")
```



```
p_levene <- leveneTest(loglongevity ~ activity, data=ff)[1,3]
```

By inspecting the histogram, we notice that `loglongevity` is slightly right-shifted due to the logarithmic mapping. The Q-Q plot however still shows approximately normal distribution. As the Levene test returns a p-value of $0.4462 > 0.05$ we assume homogeneous variances and conclude that we are able to perform ANOVA to further assess whether sexual `activity` influences `loglongevity`. We first do this without taking the thorax (length) into account. The box plot provides first hints that the loglongevity range and its mean are higher for the *isolated* group and seem to decrease increasingly for the *low* and *high* sexual activity groups.

```r
ff_aov <- lm(loglongevity ~ activity, data=ff)
anova(ff_aov); summary(ff_aov)
p_activity <- anova(ff_aov)$'Pr(>F)'[1]
loglongev_by_activity <- tapply(ff$loglongevity, ff$activity, mean)
p_baseline_low <- summary(ff_aov)$coefficients[2,4]
```

The ANOVA strongly suggests that there is a statistically significant difference in log-longevity between the three activity levels *isolated*, *low* and *high* with $p=0 < 0.05$, confirming our first intuition. While deviations between our baseline and the *low* activity group aren't significant with $p = 0.1733 > 0.05$, the *high* activity group has substantial deviations from the baseline mean. The estimated mean loglongevities for the three sexual activity groups in order are 4.1193, 3.9998 and 3.6021.

b) We continue our analysis by including the numerical explanatory variable thorax length by extending our prior analysis using ANCOVA.

```r
ff_aocv2 <- lm(loglongevity ~ thorax + activity, data=ff)
anova(ff_aocv2); summary(ff_aocv2)
p_activity_2 <- anova(ff_aocv2)$'Pr(>F)'[2]
mean_thorax <- mean(ff$thorax)
mean_thorax_ff <- data.frame(thorax=mean_thorax, activity=levels(ff$activity))
pred_loglongev <- predict(ff_aocv2, newdata=mean_thorax_ff)
```

Here, we confirm again that sexual activity is a significant factor with $p=0 < 0.05$. We can notice its effect of decreasing longevity with increasing activity based also on our predictions for the three activity groups when fixing thorax length of the data set to its mean. The estimated mean loglongevities for the three sexual activity groups in order are 4.0852 (*isolated*), 3.9609 (*low*) and 3.6752 (*high*).

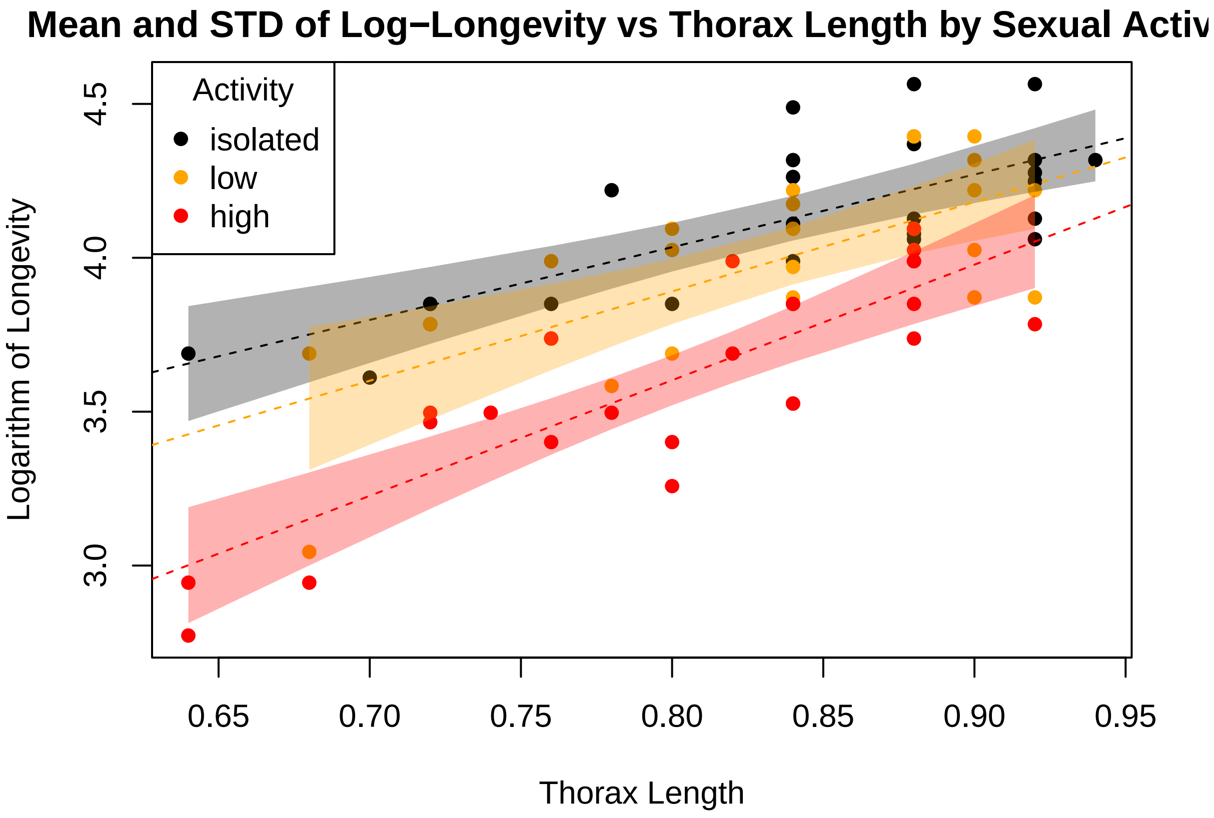
c) Following, we investigate the influence of thorax length on loglongevity. First, we perform a linear regression to visually inspect the potential relationship between the two variables, amongst others checking for differences in the slopes of the regression lines for the three sexual activity groups.

```r
par(mar=c(5, 5, 2, 2))
colors = c("black", "orange", "red")

# Color-coded scatter plot for activities
plot(ff$thorax, ff$loglongevity,
     xlab="Thorax Length", ylab="Logarithm of Longevity",
     main="Mean and STD of Log-Longevity vs Thorax Length by Sexual Activity",
     col=colors[as.numeric(factor(ff$activity))], pch=16)

# Legend
legend("topleft", legend=levels(ff$activity),
       col=colors, pch=16, title="Activity")

# Regression lines and confidence intervals
models <- list()
for (act in unique(ff$activity)) {
  subset_ff <- subset(ff, activity==act)
  model <- lm(loglongevity ~ thorax, data=subset_ff)
  models[[act]] <- model
  abline(model, col=colors[which(levels(ff$activity)==act)], lty=2)
  preds <- predict(model, interval="confidence", level=0.95)
  polygon(c(subset_ff$thorax, rev(subset_ff$thorax)),
          c(preds[, "lwr"], rev(preds[, "upr"])), border=NA,
        col=adjustcolor(colors[which(levels(ff$activity)==act)], alpha.f=0.3))
}
```



```
ancova_model <- lm(loglongevity ~ thorax * activity, data=ff)
# visreg(ancova_model, "thorax", "activity") # alternative one-liner
anova(ancova_model)
summary(ancova_model)
p_interaction <- anova(ancova_model)$'Pr(>F)'[3]
p_thorax <- summary(ancova_model)$coefficients[2,4]
```

By visually inspecting mean and standard deviation of the linear regression models, we assume that thorax length has an influence on loglongevity, although the difference between the *low* and *isolated* groups is small and perhaps negligible. Performing an ANCOVA, we find that the interaction between thorax length and activity are not quite significant enough with p=0.1536. ANCOVA results suggest that thorax length has a significant positive influence on loglongevity with $p = 0$ and this relationship appears to be consistent across different levels of sexual activity.

d) We prefer the analysis with thorax length, as it provides a more complete picture of the relationship between sexual activity, thorax length, and loglongevity. If one is solely interested in the influence of sexual activity on loglongevity or has prior information on variable relationships, then the analysis excluding thorax length is sufficient. However, it is incomplete and simplified which could cause issues if thorax length has a substantial hidden contribution due to interaction with sexual activity. Therefore, for a more comprehensive overview, the analysis including thorax length is preferred.

e) We now analyze the decision to use the logarithm of longevity as response variable by comparing ANCOVA results.

```r
ancova_model2 <- lm(longevity ~ thorax * activity, data=ff)
anova(ancova_model2); summary(ancova_model2)
leveneTest(longevity ~ activity, data=ff)
```

In the case of analyzing the true `longevity` value, no major changes in our interpretations of the results of ANCOVA are observed. We would not state that the baseline and *high* activity group have significant differences in `longevity`. However, the `thorax` length still has a significant positive influence on `longevity`. Levene tests and normality checks are still valid. Arguably, normality is clearer for the original `longevity` values.

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:

```r
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 predictors.

```r
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 examined for potential (leverage) points. 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 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.

```r
cdist_df <- data.frame()
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)
  cdist_df <- rbind(cdist_df, data.frame(Variable = names(birthweight1)[i],
                                         CooksDistance = max(cdist)))
}
print(t(cdist_df))
```

```
##                [,1]         [,2]         [,3]         [,4]         [,5]
## Variable       "Length"     "Headcirc"   "Gestation"  "mage"       "mnocig"
## CooksDistance "0.34953939" "0.13103662" "0.09450724" "0.69322117" "0.07594460"
##                [,6]         [,7]         [,8]         [,9]         [,10]
## Variable       "mheight"    "mppwt"      "fage"       "fedyrs"     "fnocig"
## CooksDistance "0.10932534" "0.13333161" "0.17291197" "0.26852864" "0.13214445"
##                [,11]
## Variable       "fheight"
## CooksDistance "0.33985493"
```

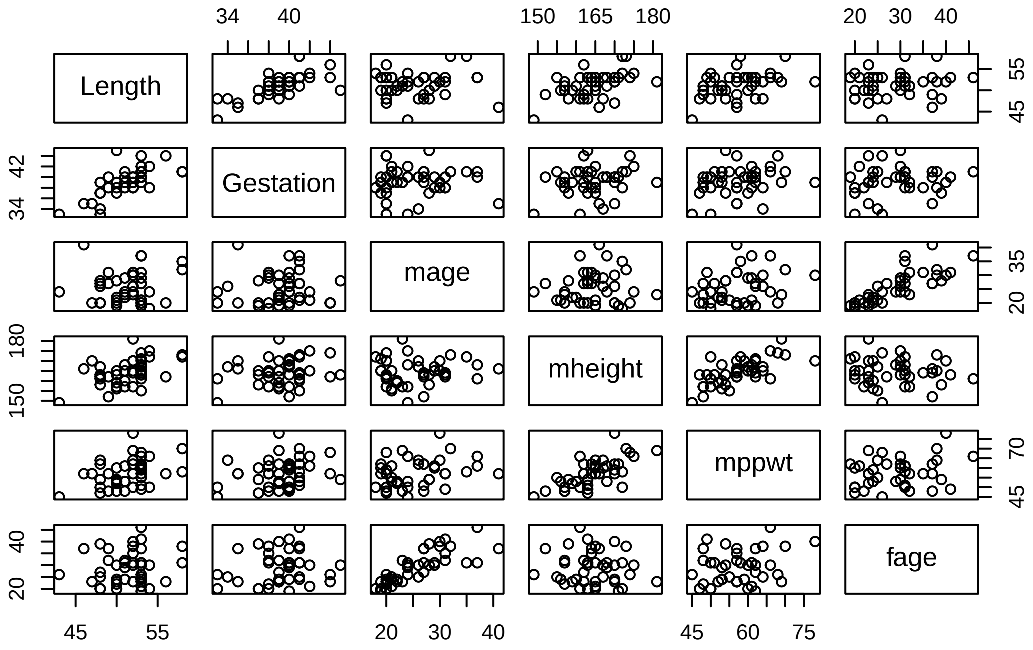
None of the Cook's distances are sufficiently large to conclude that the points are influence points. Another potential problem to be addressed is collinearity between the explanatory variables. A preliminary investigation can be conducted by analysing the variance inflation factors (VIF) of a model including the 11 predictors.

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

```
##    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, `mage` and `fage`, have VIF values close to 5. Since the above analysis does not provide information about the collinear groups that 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 examine their correlations visually, 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.

```r
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 the baby's growth in size prior to birth, the influence of the two variables seems logical.

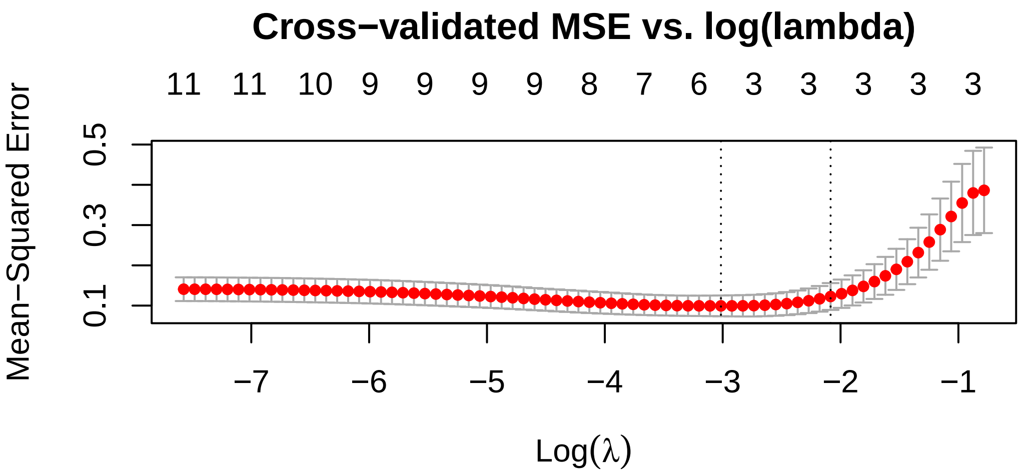
c) For the next exercise, the average of each predictor value from the reduced model is taken and used as a new observation, for which the 95% confidence and prediction intervals for the response variable `Birthweight` are calculated.

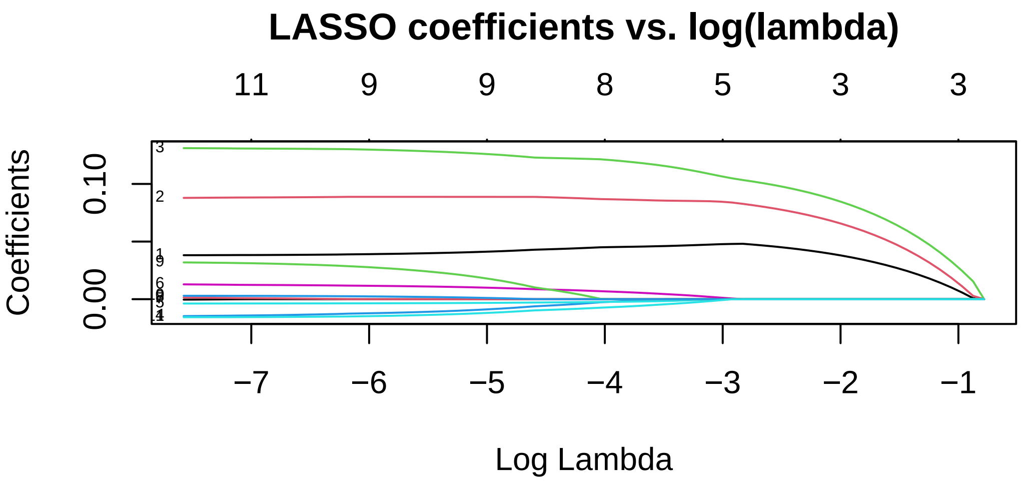
```r
averages <- data.frame(Headcirc=mean(birthweight2$Headcirc), Gestation=mean(birthweight2$Gestation))
stepdown_ci <- predict(bw_model, newdata=averages, interval="confidence", level=0.95)
stepdown_pi <- predict(bw_model, newdata=averages, interval="prediction", level=0.95)
```

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 is used to select the optimal value of the tuning parameter λ by cross-validation. Like in b), the reduction starts with the filtered data set `birthweight1`. 2/3 of the data points are randomly sampled for training and the remaining 1/3 is reserved for subsequent testing.

```r
library(glmnet); par(mfrow=c(2,1), mar=c(4,4,6,1)); set.seed(123)
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
```

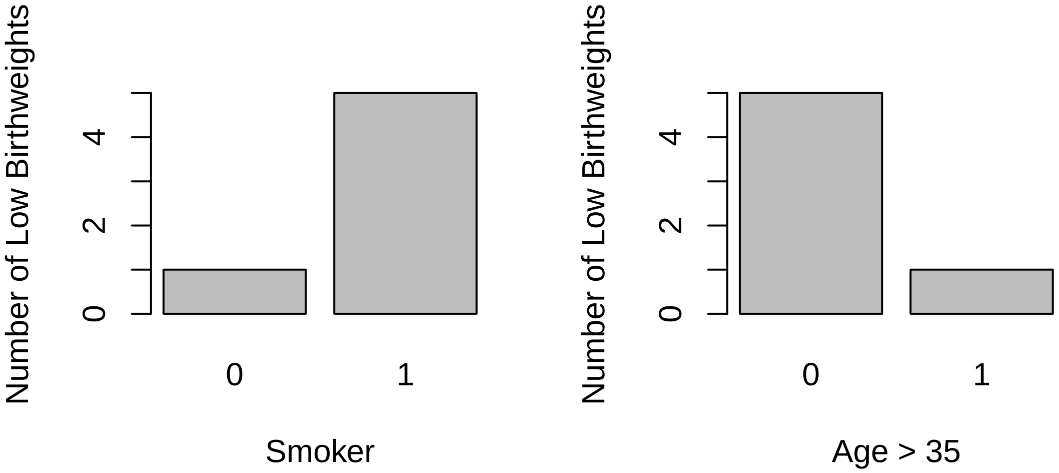
```
lasso_pred <- predict(lasso_model, s=lambda_min, newx=x_test)
mse_lasso <- mean((lasso_pred-y_test)^2)
stepdown_pred <- predict(bw_model, newdata=x_test_df)
mse_stepdown <- mean((stepdown_pred-y_test)^2)
```

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. Nonetheless, its merit may lie in its efficient handling of models of much higher complexity.

e) The next exercises consider a new subset of the data set: the binary variable `lowbwt`, indicating whether the newborn baby's weight is under 6 lbs, is used as a response variable, while the variables `Gestation` (gestation period), `smoker` (a binary variable indicating whether the mother is a smoker) and `mage35` (a binary variable indicating whether the mother's age is above 35) are used as predictors. A new subset of the data set is created with only the relevant contents. Two bar plots are created to illustrate the relationship between each of the binary predictors and the binary outcome.

```
library(dplyr); par(mfrow=c(1,2))
birthweight3 <- select(birthweight, lowbwt, Gestation, smoker, mage35)
totsmoker <- xtabs(~smoker, data=birthweight3)
barplot(xtabs(lowbwt~smoker, data=birthweight3), xlab="Smoker",
        ylab="Number of Low Birthweights")
totmage35 <- xtabs(~mage35, data=birthweight3)
barplot(xtabs(lowbwt~mage35, data=birthweight3), xlab="Age > 35",
        ylab="Number of Low Birthweights")
```



```r
tot <- xtabs(~smoker+mage35, data=birthweight3)
totlowbwt <- xtabs(lowbwt~smoker+mage35, data=birthweight3)
print(totsmoker); print(totmage35); print(round(totlowbwt/tot, 2))
```

```
## smoker
##  0  1
## 20 22
```

```
## mage35
##  0  1
## 38  4
```

```
##       mage35
## smoker    0    1
##      0 0.05 0.00
##      1 0.21 0.33
```

Among the observations, only a single non-smoking mother gave birth to an underweight baby. Looking at the age, only one mother older than 35 gave birth to an underweight baby. As the tabular summary of the variable `smoker` suggests, the distribution of smoking vs. non-smoking mothers is quite even, so the prevalence of smokers in low birth weights can be considered indicative of the claim that smoking mothers have lighter babies. On the other hand, there are only 4 35+ mothers in the data set, so the proportions of light babies per age category needs to be considered: 1 in 4 older mothers have a light baby, whereas this applies for only $5/38 \approx 0.13$ of the younger mothers. This may be taken as an indication that a low birth weight occurs in a higher percentage with older mothers. A table indicating the proportion of low birth weights for each combination of `smoker` and `mage35` suggests that the highest occurrence of low birth weights is among smoking mothers older than 35 - low birth weight applies to one third of this group.

f) The relationship between the response variable and the predictors is further investigated by fitting a logistic regression model without interactions to the data.

```r
bw_logmodel <- glm(lowbwt~Gestation+smoker+mage35, data=birthweight3, family="binomial")
summary(bw_logmodel)
```

```
##
## Call:
## glm(formula = lowbwt ~ Gestation + smoker + mage35, family = "binomial",
##     data = birthweight3)
```

```
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  48.9920    22.5659   2.171   0.0299 *
## Gestation    -1.4633     0.6700  -2.184   0.0290 *
## smoker        5.4495     3.3567   1.623   0.1045
## mage35        0.3223     4.3751   0.074   0.9413
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 34.450  on 41  degrees of freedom
## Residual deviance: 11.803  on 38  degrees of freedom
## AIC: 19.803
##
## Number of Fisher Scoring iterations: 8
```

The summary of the model provides the coefficients needed to estimate the odds ratios for the predictors using the formula $\hat{o_k} = e^{\mu+c_1 g_k+c_2 s_k+c_3 m_k}$, where g_k, s_k and m_k are the k-th observation of `Gestation`, `smoker` and `mage35`, respectively, and c_1, c_2 and c_3 are their coefficients. The coefficient for `Gestation` is negative, which means that the odds of a low birth weight decrease with increasing gestation period, namely by a factor $e^{-1.4633} \approx 0.231$. The coefficients for `smoker` and `mage35` are positive, which means that the odds of a low birth weight increase for smoking mothers and mothers older than 35. This happens by a factor $e^{5.4495} \approx 232.64$ for smoking mothers and by a factor $e^{0.3223} \approx 1.38$ for older mothers. This confirms the previous observations that older age and, more primarily, smoking are associated with a higher risk of low birth weight

g) Here, we follow a similar procedure to that in f), but we include the respective interactions in the models. Then we can run an ANOVA to determine if the interaction term is significant.

```
bw_logmodel_int1 <- glm(lowbwt~Gestation+smoker+mage35+Gestation:smoker,
                        data=birthweight3, family="binomial")
bw_logmodel_int2 <- glm(lowbwt~Gestation+smoker+mage35+Gestation:mage35,
                        data=birthweight3, family="binomial")
anova(bw_logmodel_int1, test="Chisq"); anova(bw_logmodel_int2, test="Chisq")
```

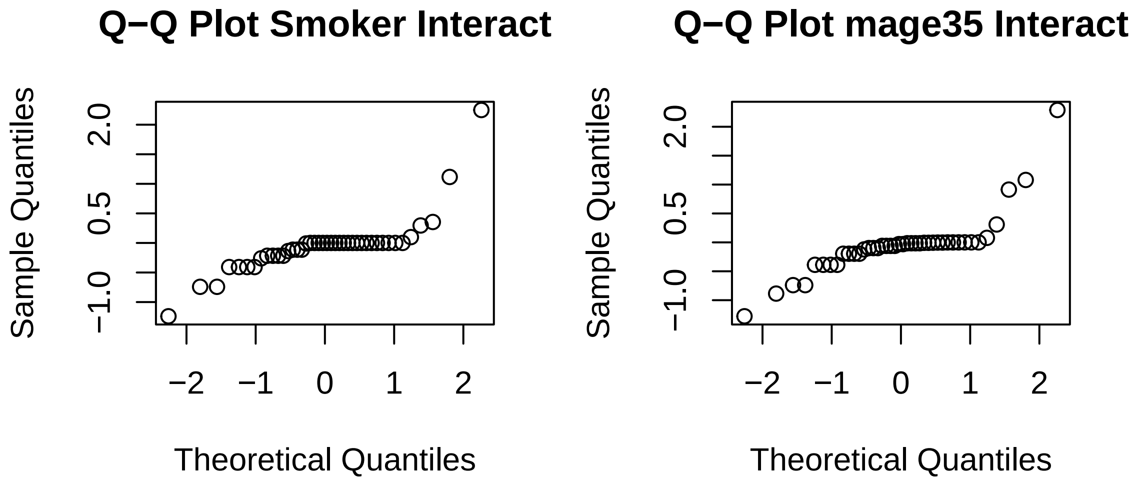
```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: lowbwt
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                41     34.450
## Gestation         1  16.4247        40     18.025 5.062e-05 ***
## smoker            1   6.2162        39     11.809   0.01266 *
## mage35            1   0.0054        38     11.803   0.94138
## Gestation:smoker  1   1.6620        37     10.141   0.19733
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
```

```
##
## Model: binomial, link: logit
##
## Response: lowbwt
##
## Terms added sequentially (first to last)
##
##
##                  Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                               41     34.450
## Gestation         1  16.4247        40     18.025 5.062e-05 ***
## smoker            1   6.2162        39     11.809   0.01266 *
## mage35            1   0.0054        38     11.803   0.94138
## Gestation:mage35  1   0.1240        37     11.680   0.72476
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We notice that, for both models where the interaction term was tested individually, neither interaction term was significant according to our ANOVA. However, both the variables `Gestation` and `smoker` are significant, and should be included in the model despite not having an interaction. In this case, we would prefer to use the model from f) and not g).

```r
par(mfrow=c(1,2))
qqnorm(residuals(bw_logmodel_int1), main="Q-Q Plot Smoker Interact")
qqnorm(residuals(bw_logmodel_int2), main="Q-Q Plot mage35 Interact")
```



It is also a good idea to check our assumptions for the ANOVA test as well. We see that our model residuals are not quite linear in the Q-Q Plot, so normality is doubtful. This is further proof that using the model with interactions not well founded, and should proceed with our model from part f) with no interactions.

h) Our resulting model from f) is used below to determine the probability of low birth weight for the combinations of factors with a gestation length of 40 weeks.

```r
combo1 = data.frame(Gestation=40, smoker=0, mage35=0)
combo2 = data.frame(Gestation=40, smoker=1, mage35=0)
combo3 = data.frame(Gestation=40, smoker=0, mage35=1)
combo4 = data.frame(Gestation=40, smoker=1, mage35=1)
```

```r
plbw_combo1 = predict(bw_logmodel,combo1,type="response")
plbw_combo2 = predict(bw_logmodel,combo2,type="response")
plbw_combo3 = predict(bw_logmodel,combo3,type="response")
plbw_combo4 = predict(bw_logmodel,combo4,type="response")
summary_table <- matrix(c(plbw_combo1, plbw_combo3, plbw_combo2, plbw_combo4),
                        byrow = TRUE, ncol = 2,
                        dimnames = list(c("Non-smoker", "Smoker"),
                                        c("Under 35", "Over 35")))
print(summary_table)
```

```
##                Under 35      Over 35
## Non-smoker 7.199185e-05 9.936346e-05
## Smoker     1.647403e-02 2.259660e-02
```

The results from the probability estimation are summarized in the above table. We notice once again that the probability of having a baby with a low birth weight is highest for a mother that is over 35 and a smoker, and lowest for a mother that is below 35 and not a smoker, which follows our intuition.

i) In order to perform a contingency table test, the dataframe needs to be converted into a matrix.

```r
#Creating the matrix for the smoker boolean low birthweight
grouped_df_smoker <- with(birthweight3, table(lowbwt, smoker))
dims <- dim(grouped_df_smoker)
flat_table <- as.vector(grouped_df_smoker)
matrix_data_smoker <- matrix(flat_table, nrow = prod(dims[-1]), ncol = dims[1])
rownames(matrix_data_smoker) <- c("lowbwt=0", "lowbwt=1")
colnames(matrix_data_smoker) <- c("smoker=0", "smoker=1")
#Creating the matrix for the mage35 boolean low birthweight
grouped_df_mage35 <- with(birthweight3, table(lowbwt, mage35))
dims <- dim(grouped_df_mage35)
flat_table <- as.vector(grouped_df_mage35)
matrix_data_mage35 <- matrix(flat_table, nrow = prod(dims[-1]), ncol = dims[1])
rownames(matrix_data_mage35) <- c("lowbwt=0", "lowbwt=1")
colnames(matrix_data_mage35) <- c("mage35=0", "mage35=1")
#Conducting the two Fisher tests since we have two 2x2 matrices
pval_smoker <- fisher.test(matrix_data_smoker)[[1]]
pval_mage35 <-  fisher.test(matrix_data_mage35)[[1]]
```

Since both of our p-values, 0.1870341, 0.4737336 are above 0.05, we would not reject the null hypothesis that there is neither a dependence between smoking and low birth weight nor a dependence between `mage35` and low birth weight. Additionally, we do not have at least 5 observations for each combination, which makes the test unreliable. However, the approach in general is not incorrect. One advantage to using a chi-square test is that it simplifies the problem to mere dependence vs independence between variables. However, it does not give specific coefficient variables to delineate how the probabilities of events occurring change with respect to the categorical variables, nor does it allow for continuous variables such as `Gestation` to be taken into account.

Exercise 3. School awards

a)

```r
#Read in the data and extract the relevant column
awards=read.table("data/awards.txt",header=TRUE)
#Poisson regression with program as a factor
awards$prog=factor(awards$prog)
awardsglm1 = glm(num_awards~prog, family=poisson, data=awards)
```

```
drop1(awardsglm1, test="Chisq")
```

```
## Single term deletions
##
## Model:
## num_awards ~ prog
##        Df Deviance    AIC    LRT Pr(>Chi)
## <none>      216.10 512.42
## prog    2   228.83 521.15 12.733 0.001718 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(awardsglm1)
```

```
##
## Call:
## glm(formula = num_awards ~ prog, family = poisson, data = awards)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.5486     0.1961  -2.797  0.00515 **
## prog2         0.7068     0.2158   3.275  0.00106 **
## prog3         0.4432     0.2463   1.799  0.07199 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 216.10  on 197  degrees of freedom
## AIC: 512.42
##
## Number of Fisher Scoring iterations: 5
```

Upon inspection of the output from drop1, we notice that program has a significant impact on the number of awards that each student receives. By looking at the summary output of the model, we notice that program 2 (“general”) has the highest associated coefficient, i.e. the probability of a student in the general program of increasing is higher for that student than in the vocational or academic programs.

```
program1 = data.frame(prog="1")
program2 = data.frame(prog="2")
program3 = data.frame(prog="3")
estimated_num_awards_p1 = predict(awardsglm1,program1,type="response")
estimated_num_awards_p2 = predict(awardsglm1,program2,type="response")
estimated_num_awards_p3 = predict(awardsglm1,program3,type="response")
```

The expected number of awards for programs 1, 2, and 3 are 0.5778, 1.1714, and 0.9 respectively. We notice that students in program 2 will have a higher expected number of awards, which confirms out findings from the summary of the poisson regression.

b) Since the Kruskal-Wallis test is nonparametric, there is no assumption of normality that needs to be met for the data; hence, it is applicable in this situation. As a nonparametric alternative to ANOVA, we can apply the test to see if there is a significant difference between the distributions of awards among the different programs.

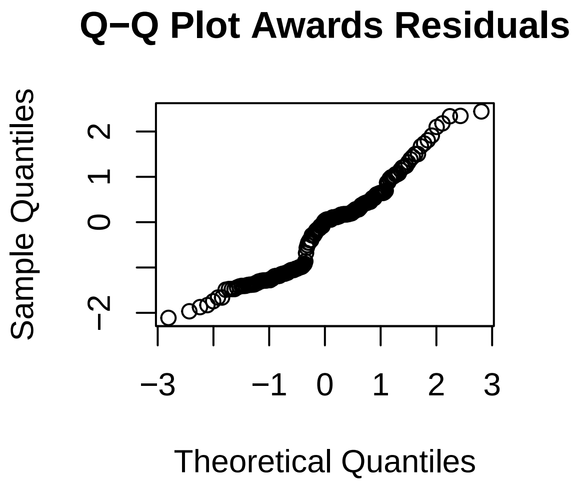
```
kruskal.test(awards$num_awards, awards$prog)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  awards$num_awards and awards$prog
## Kruskal-Wallis chi-squared = 10.755, df = 2, p-value = 0.00462
```

The rest results in a significant p-value, so we can reject the null hypothesis that all of the distributions for the numbers of awards for the different programs are the same. This aligns with our findings from the poisson regression above.

c) First, we conduct the Poisson regression with program type, math, and their interaction to see if their interaction is significant in the model. Then, we want to conduct an ANOVA to determine if an interaction exists. It is important to check the ANOVA model assumptions.

```
awardsglm2 = glm(num_awards~prog*math, family=poisson, data=awards)
qqnorm(residuals(awardsglm2), main="Q-Q Plot Awards Residuals")
```



```
anova(awardsglm2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: num_awards
##
## Terms added sequentially (first to last)
##
##
##           Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                        199     228.83
## prog       2  12.7334       197     216.10  0.001718 **
## math       1  18.0527       196     198.05 2.149e-05 ***
## prog:math  2   3.6911       194     194.35  0.157935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Q-Q plot looks relatively linear, so normality in the model residuals is a valid assumption. From ANOVA results, we see that the interaction between program type and math score is insignificant, so we do not need to include it in the model. However, both program and math are significant when determining the number of awards, so we will keep those variables in the model and leave out their interaction.

```
awardsglm3 = glm(num_awards~prog+math, family=poisson, data=awards)
summary(awardsglm3)
```

```
## 
## Call:
## glm(formula = num_awards ~ prog + math, family = poisson, data = awards)
## 
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)    
## (Intercept) -2.372577   0.475525  -4.989 6.06e-07 ***
## prog2        0.452621   0.224746   2.014   0.0440 *  
## prog3        0.561720   0.247482   2.270   0.0232 *  
## math         0.035779   0.008344   4.288 1.80e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for poisson family taken to be 1)
## 
##     Null deviance: 228.83  on 199  degrees of freedom
## Residual deviance: 198.05  on 196  degrees of freedom
## AIC: 496.36
## 
## Number of Fisher Scoring iterations: 5
```

From the summary of th results with math included as an explanatory variable, we notice that the third program ("academic") now has the highest coefficient, so it will increase the probability that a student has a higher average number of awards the most.

```
options(tinytex.verbose = TRUE)
program1 = data.frame(prog="1", math=56)
program2 = data.frame(prog="2", math=56)
program3 = data.frame(prog="3", math=56)
estimated_num_awards_p1 = predict(awardsglm3,program1,type="response")
estimated_num_awards_p2 = predict(awardsglm3,program2,type="response")
estimated_num_awards_p3 = predict(awardsglm3,program3,type="response")
```

The expected number of awards for programs 1, 2, and 3 are 0.6915, 1.0873, and 1.2126 respectively. Here, we notice once again that the students in the academic program have the highest expected number awards when we include math scores as an explanatory variable, which aligns with the results from the summary of our model.