Exercise 1

Adam Sulak, Onder Akacik, Arda Ergin

2024-02-22

```
options(digits=2)

library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)
library(car)
```

```
## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(glmnet)
```

```
## Loading required package: Matrix

## Loaded glmnet 4.1-8
```

```
color_choice <- "Dark2"
```

Exercise 2

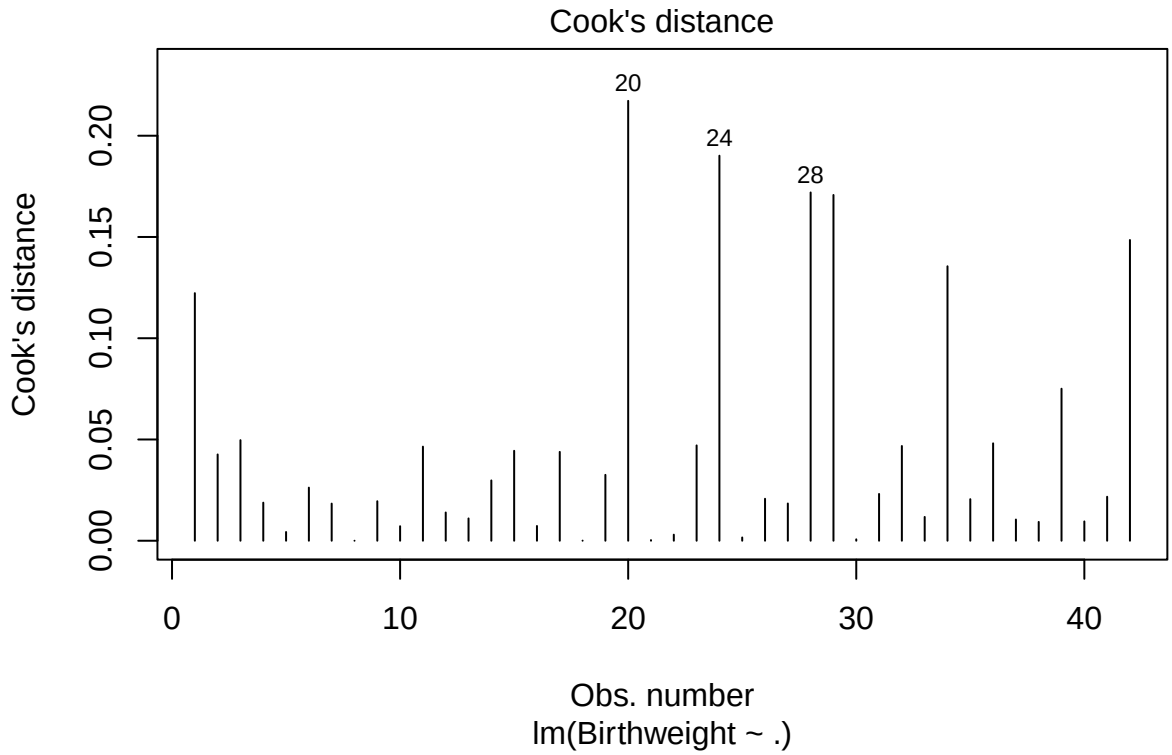
```
data <- read.csv("Birthweight.csv")
```

Question 2a)

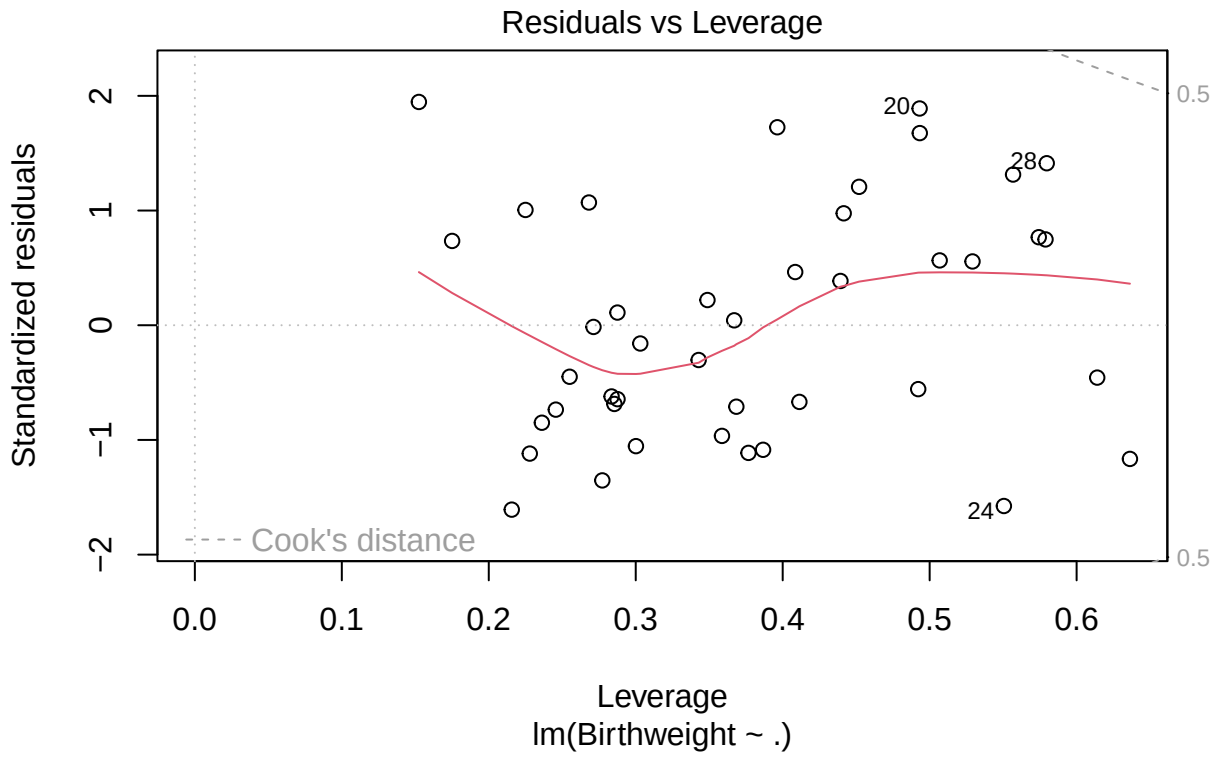
```
model_full <- lm(Birthweight ~ ., data = data)
```

In order to investigate "the problem of potential and influence points", we can first take a look at the diagnostic plots.

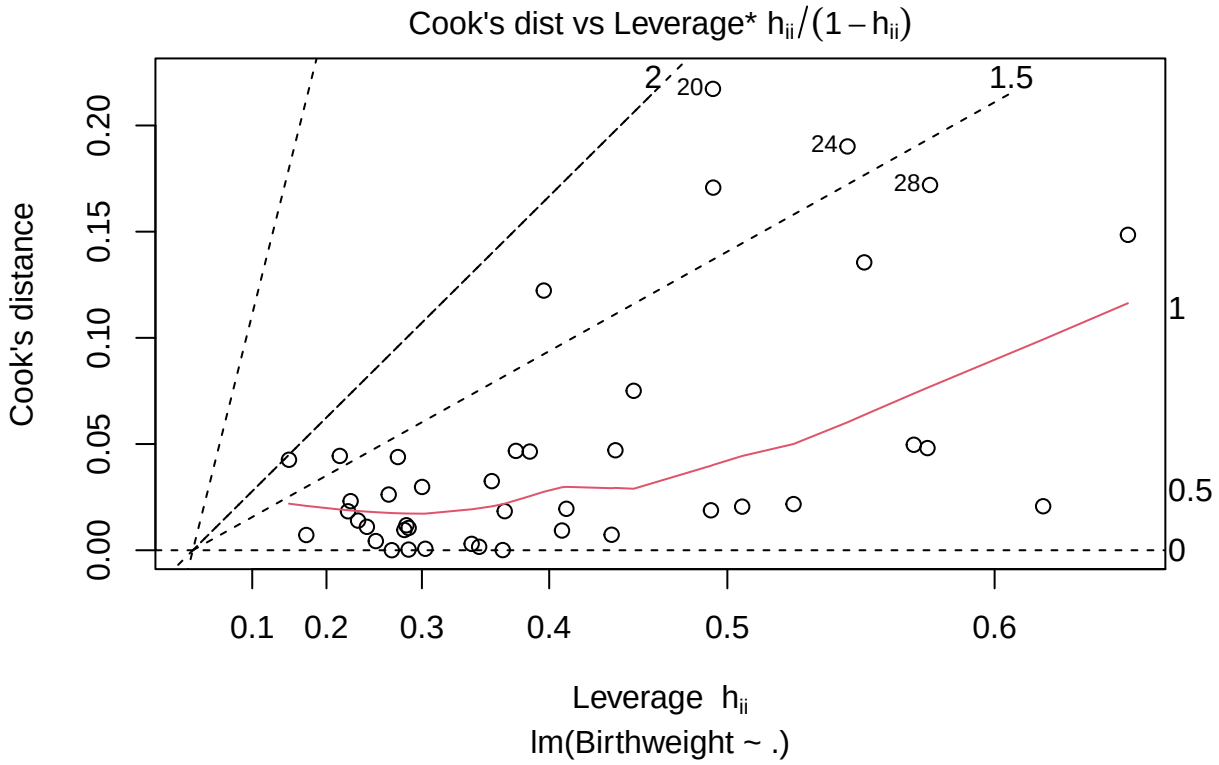
```
plot(model_full, which=4)
```




```
plot(model_full, which=5)
```

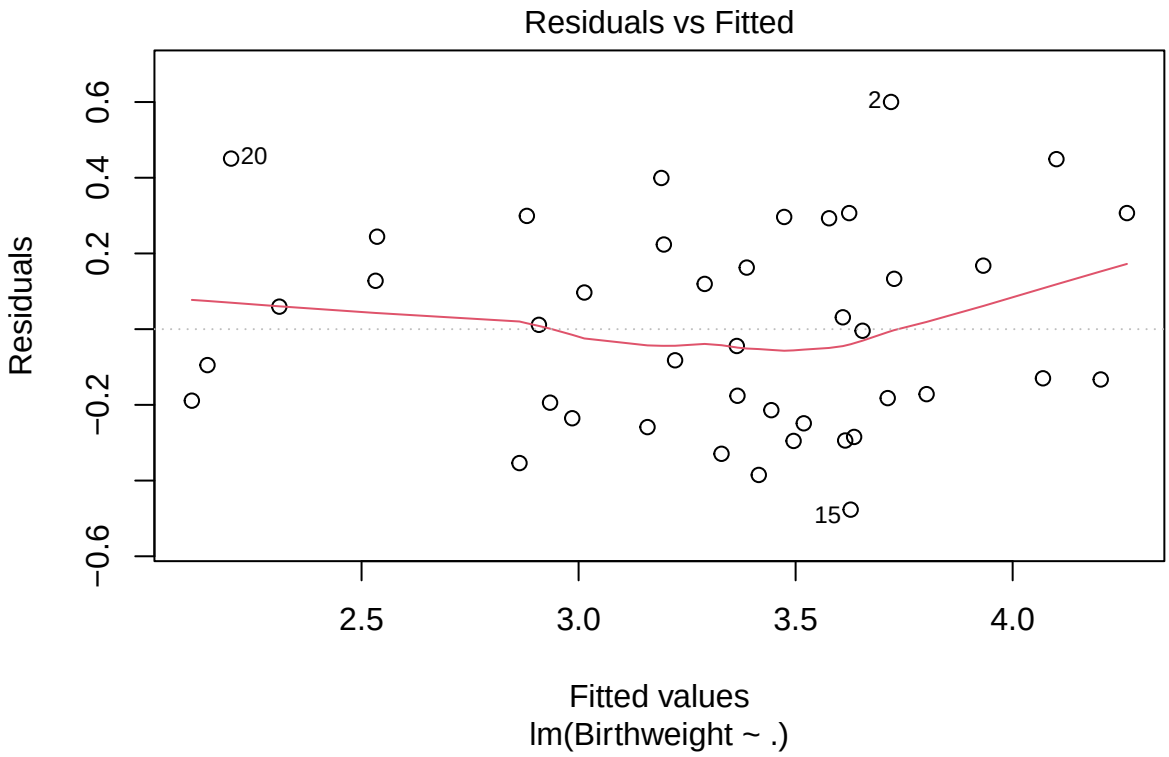


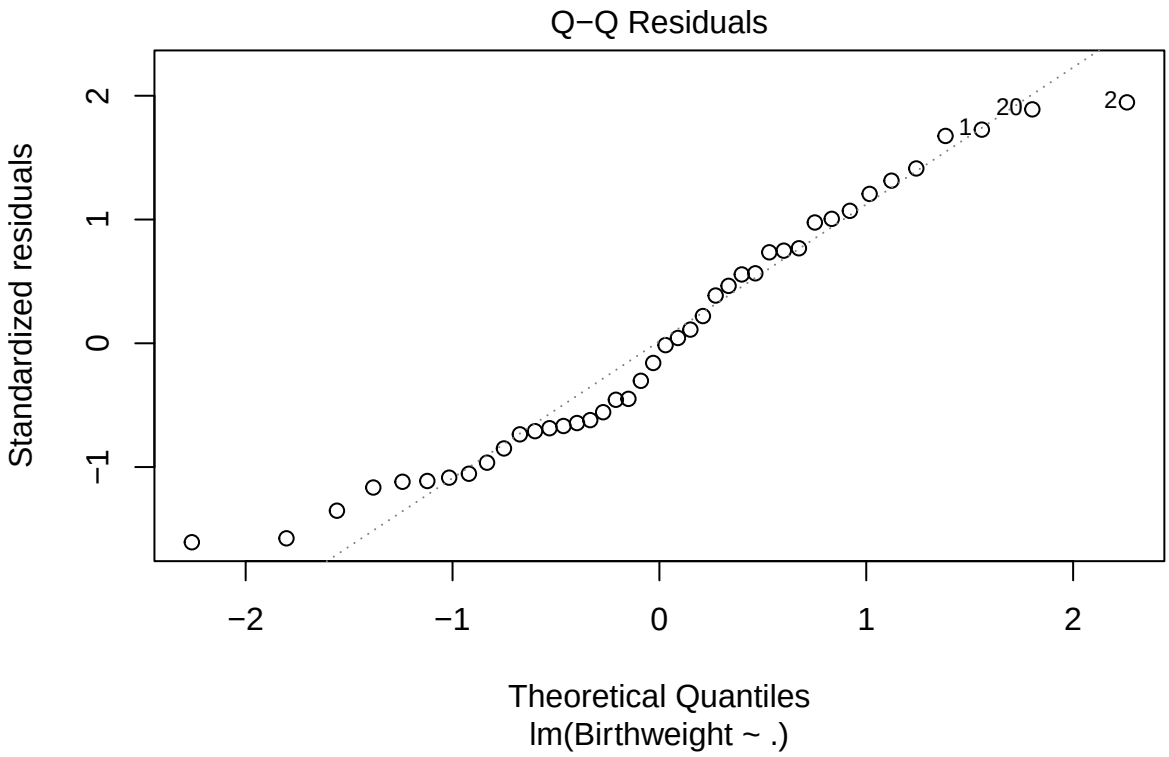

```
plot(model_full, which=6)
```

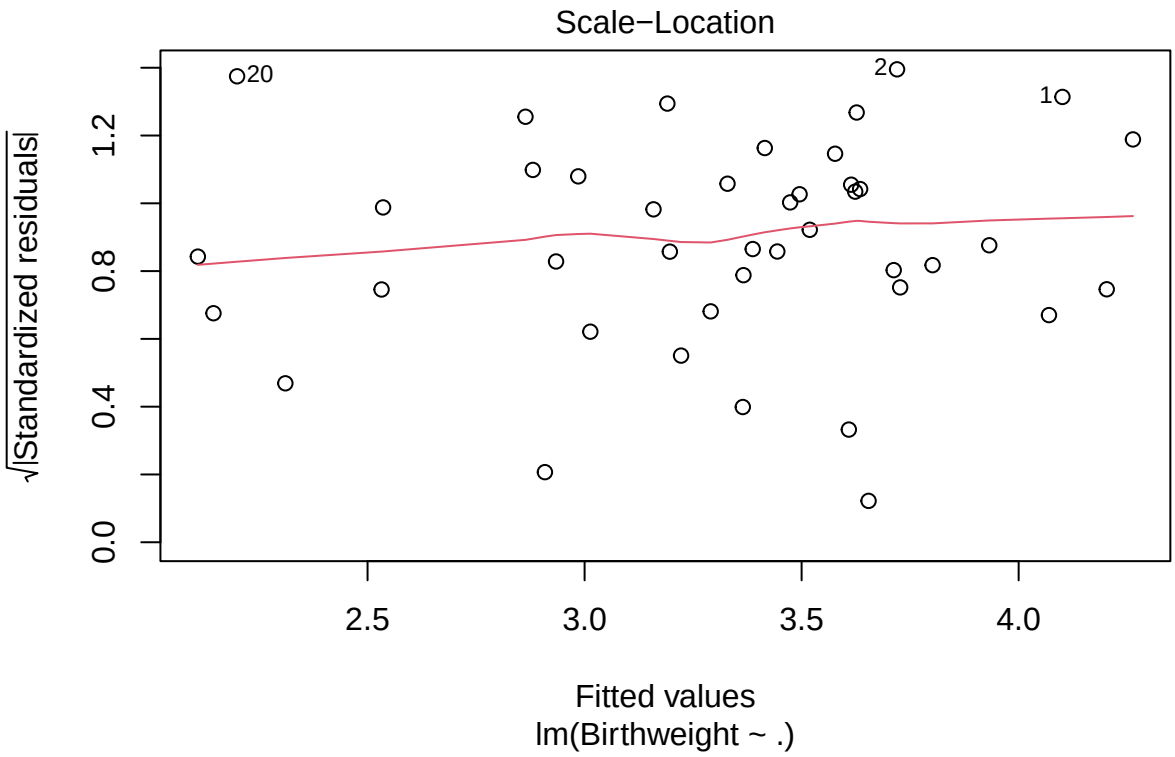


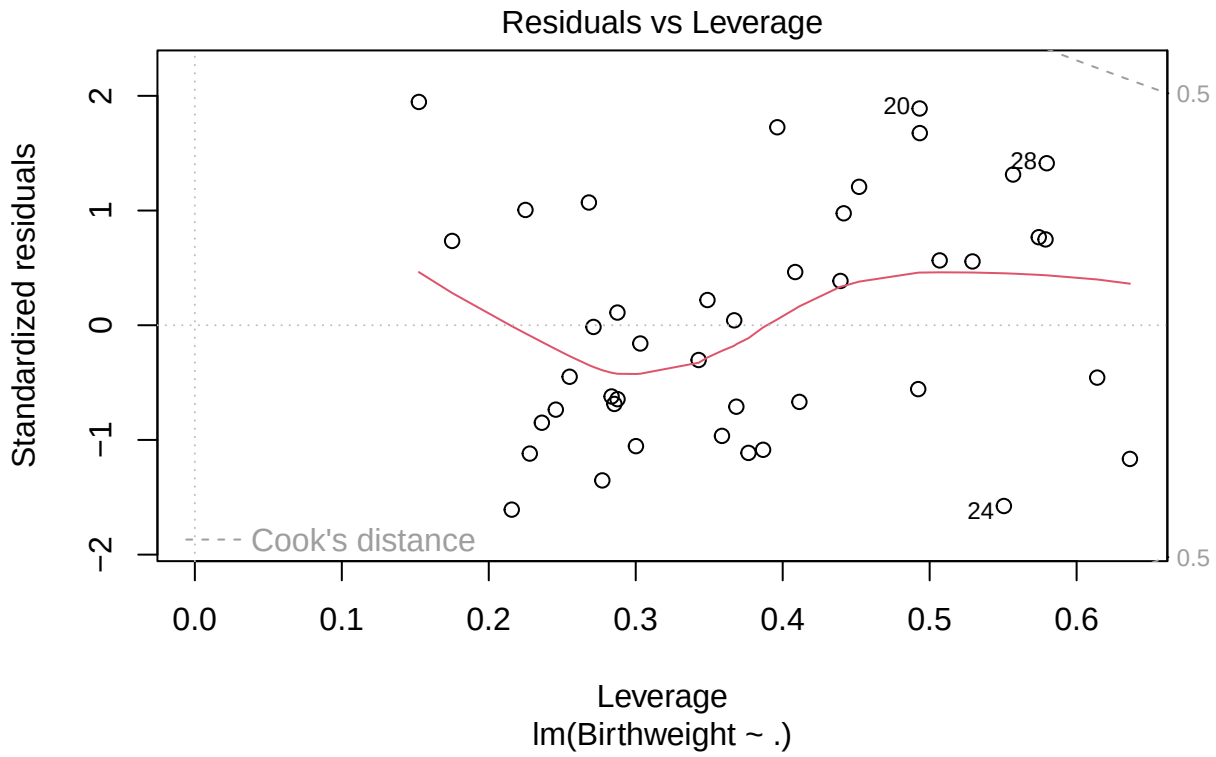
These plots reveal

```
# par(mfrow=c(2,2))
plot(model_full)[1]
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
−0.2
−0.6
20
2
15
2.5
3.0
3.5
4.0
Fitted values
lm(Birthweight ~ .)


Q−Q Residuals
Standardized residuals
−1
0
1
2
−2
−1
0
1
2
Theoretical Quantiles
lm(Birthweight ~ .)
1
20
2


Scale−Location
√|Standardized residuals|
0.0
0.4
0.8
1.2
20
2
1
2.5
3.0
3.5
4.0
Fitted values
lm(Birthweight ~ .)

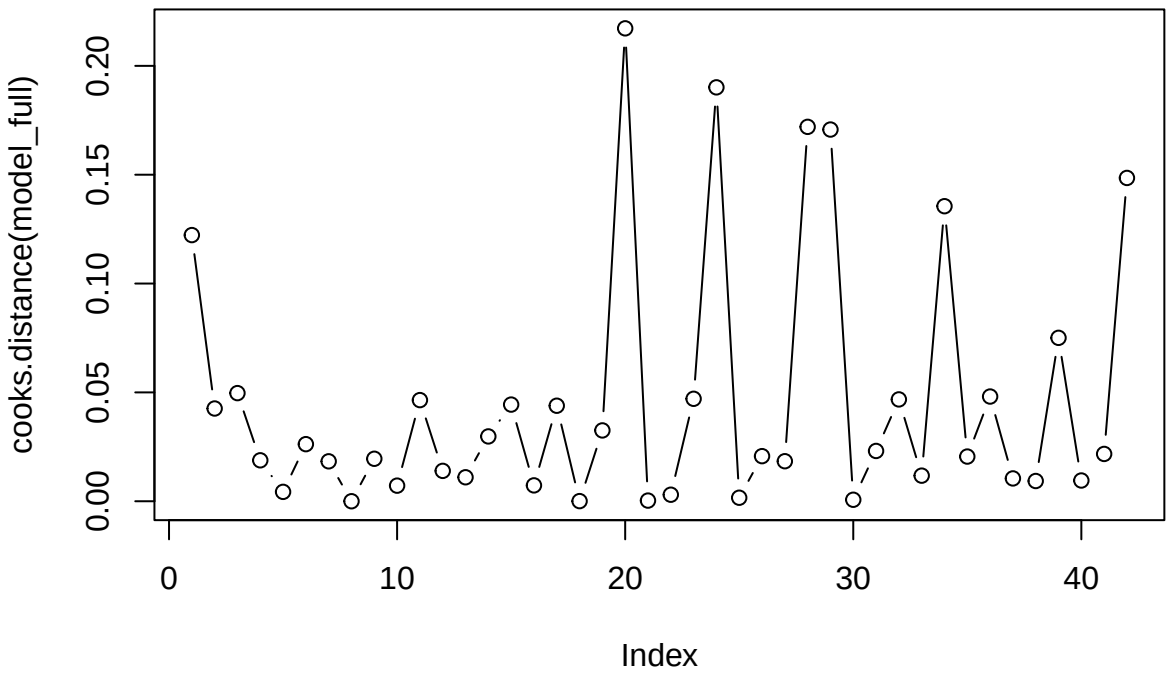



```
## NULL
```

```
#
resid_ordered <- order(abs(residuals(model_full)))
u = rep(0,length(resid_ordered))
u[length(resid_ordered)] = 1
#
cooks.distance(model_full)
```

```
##        1        2        3        4        5        6        7        8        9       10
## 1.2e-01 4.3e-02 5.0e-02 1.9e-02 4.3e-03 2.6e-02 1.8e-02 5.2e-06 2.0e-02 7.2e-03
##       11       12       13       14       15       16       17       18       19       20
## 4.6e-02 1.4e-02 1.1e-02 3.0e-02 4.4e-02 7.3e-03 4.4e-02 6.6e-05 3.3e-02 2.2e-01
##       21       22       23       24       25       26       27       28       29       30
## 3.1e-04 3.0e-03 4.7e-02 1.9e-01 1.6e-03 2.1e-02 1.8e-02 1.7e-01 1.7e-01 6.9e-04
##       31       32       33       34       35       36       37       38       39       40
## 2.3e-02 4.7e-02 1.2e-02 1.4e-01 2.1e-02 4.8e-02 1.0e-02 9.3e-03 7.5e-02 9.5e-03
##       41       42
## 2.2e-02 1.5e-01
```

```
plot(cooks.distance(model_full), type="b")
```




```
#forbeslm_42 = lm(y~x+u11); summary(forbeslm11)
```

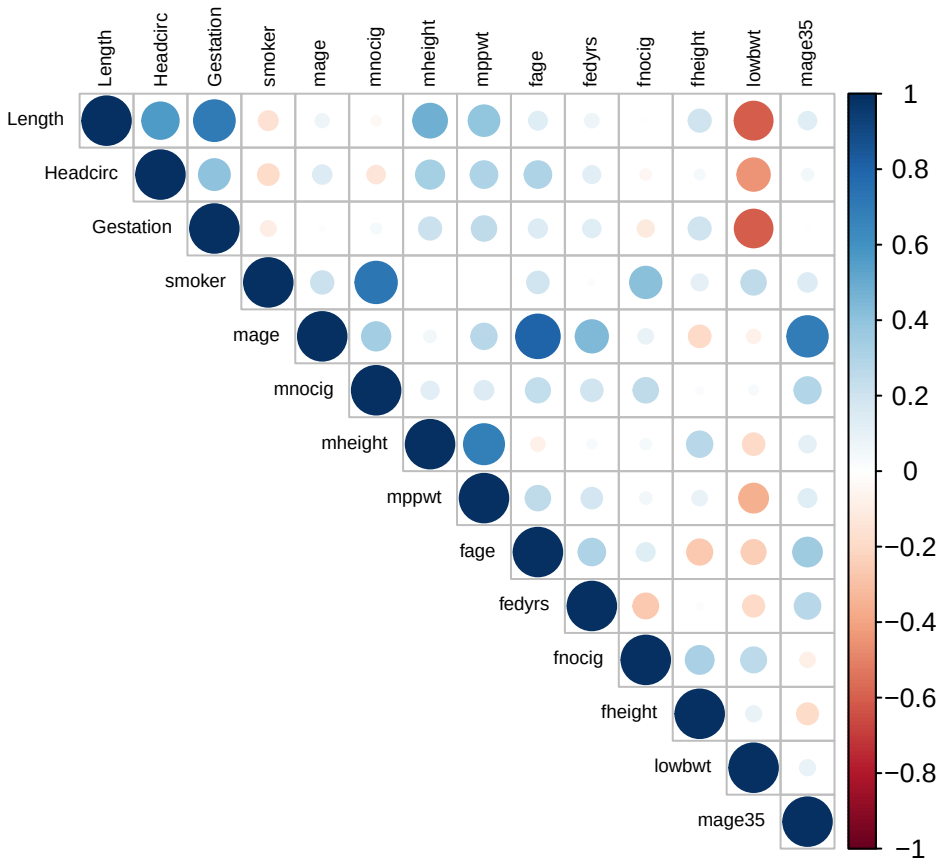
Investigating multi-collinearity:

```
car::vif(model_full)
```

```
##        ID    Length  Headcirc Gestation    smoker      mage    mnocig   mheight
##       1.6       4.8       1.9       2.6       2.8       9.9       2.9       3.4
##     mppwt      fage    fedyrs    fnocig   fheight    lowbwt    mage35
##       2.5       6.7       1.7       2.0       1.8       3.0       4.3
```

When we look into `car::vif()`, if we follow the rule of thumb, we see that no variable has a VIF value that is higher than 5. Hence, we do not seem to have any issues regarding multi-collinearity.

```
# Calculate correlation matrix
cor_matrix <- cor(data[, -c(1,3)]) # Exclude response variable Birthweight

# Plot the correlation matrix
corrplot(cor_matrix,
         method = "circle", type = "upper", tl.col = "black", tl.cex = 0.6)
```



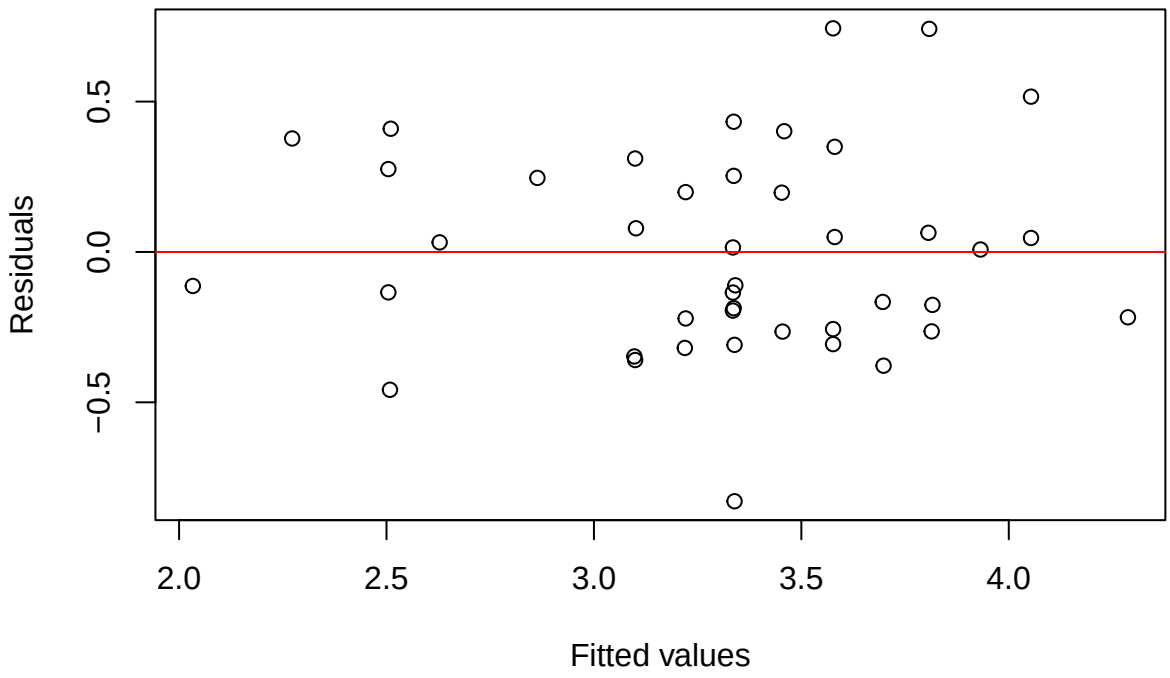
Question 2b)

```
# step_down_model <- step(model, direction = 'backward')
model <- lm(
    Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mheight + mppwt + fage + fedyrs + fno
    data = data
)
# remove fage
model_1 <- lm(
    Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mheight + mppwt + fedyrs + fnocig + fl
    data = data
)
# remove mheight
model_2 <- lm(
    Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mppwt + fedyrs + fnocig + fheight,
    data = data
)
# remove fedyrs
model_3 <- lm(
    Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mppwt + fnocig + fheight,
    data = data
)
# remove fnocig as it's not significant and Adjusted R-squared is comparable
model_4 <- lm(
```

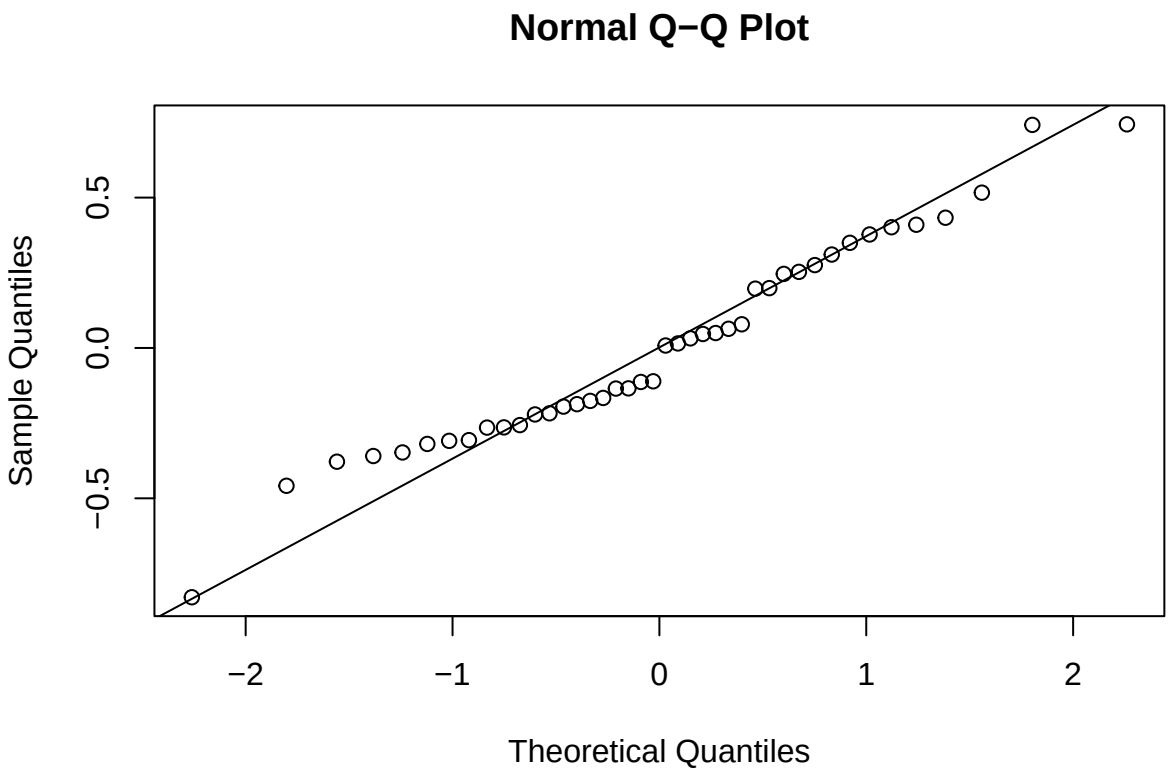
```
    Birthweight ~ Length + Headcirc + Gestation + mage + mnocig + mppwt + fheight,
    data = data
)

# remove mnocig as it's not significant and Adjusted R-squared is comparable
model_5 <- lm(
    Birthweight ~ Length + Headcirc + Gestation + mage + mppwt + fheight,
    data = data
)

# remove length as it's not significant and Adjusted R-squared is comparable
model_6 <- lm(
    Birthweight ~ Headcirc + Gestation + mage + mppwt + fheight,
    data = data
)

# remove fheight as it's not significant and Adjusted R-squared is comparable
model_7 <- lm(
    Birthweight ~ Headcirc + Gestation + mage + mppwt,
    data = data
)

# remove mage as it's not significant and Adjusted R-squared is comparable
model_8 <- lm(
    Birthweight ~ Headcirc + Gestation + mppwt,
    data = data
)


# remove mppwt as it's not significant and Adjusted R-squared is comparable
step_down_model <- lm(
    Birthweight ~ Headcirc + Gestation,
    data = data
)
```

- Tests below are done for the model assumption..

```
# linearity
plot(predict(step_down_model), residuals(step_down_model), xlab = "Fitted values", ylab = "Residuals")
abline(h = 0, col = "red")
```




```
# normality of residuals
qqnorm(residuals(step_down_model))
qqline(residuals(step_down_model))
```



Question 2c)

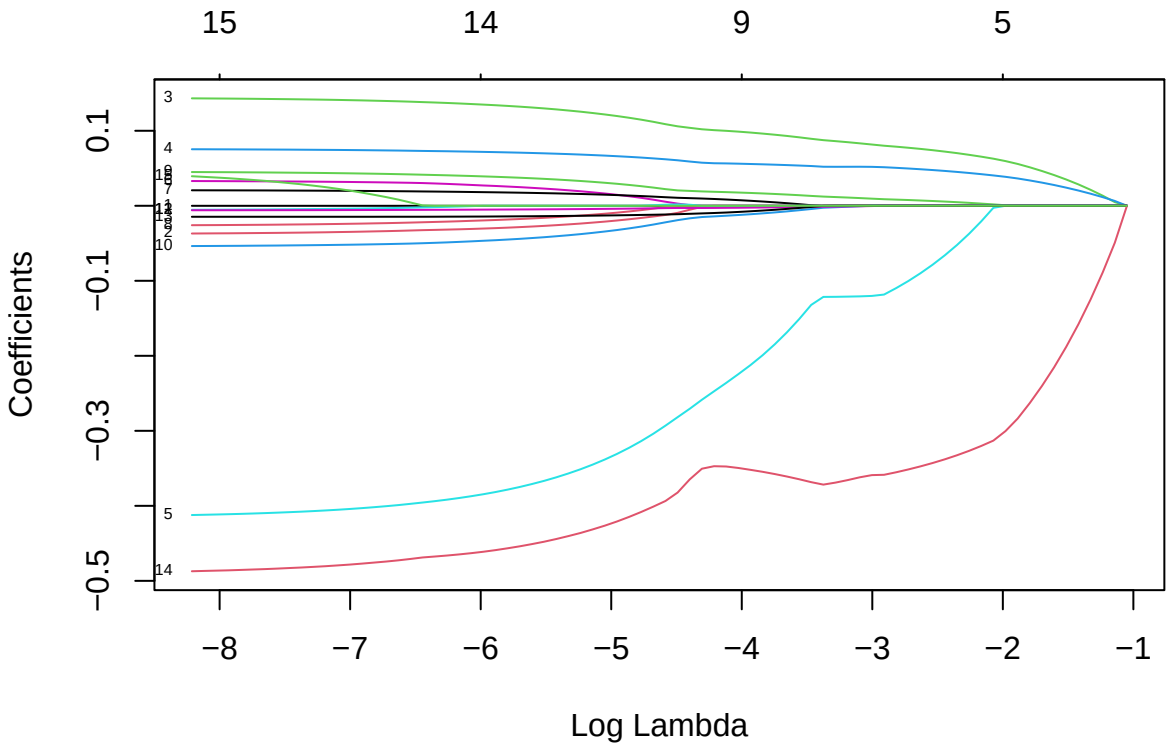
```
head_mean <- mean(data$Gestation)
gest_mean <- mean(data$Headcirc)

df <- data.frame(Gestation = gest_mean, Headcirc = head_mean)

ci <- predict(step_down_model, newdata = df, interval = 'confidence')
pi <- predict(step_down_model, newdata = df, interval = 'prediction')

ci; pi
```

```
##   fit lwr upr
## 1 3.3 2.9 3.7
```

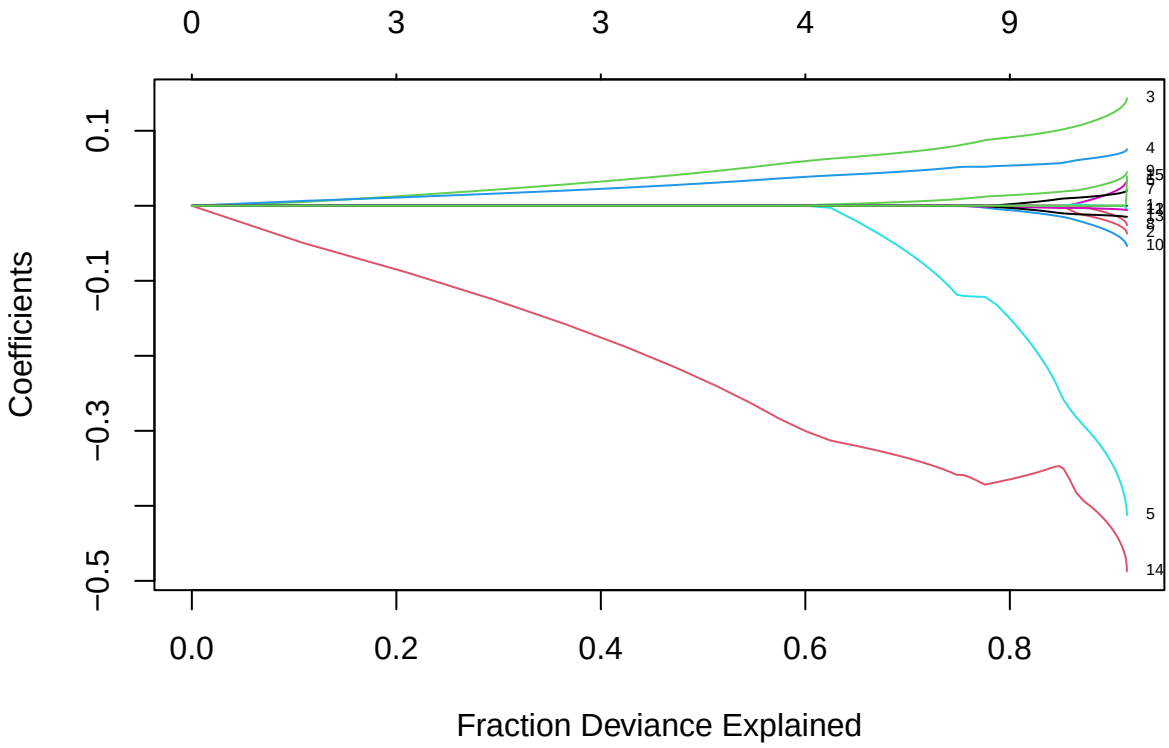
```
##   fit lwr upr
## 1 3.3 2.5 4.1
```

Question 2d)

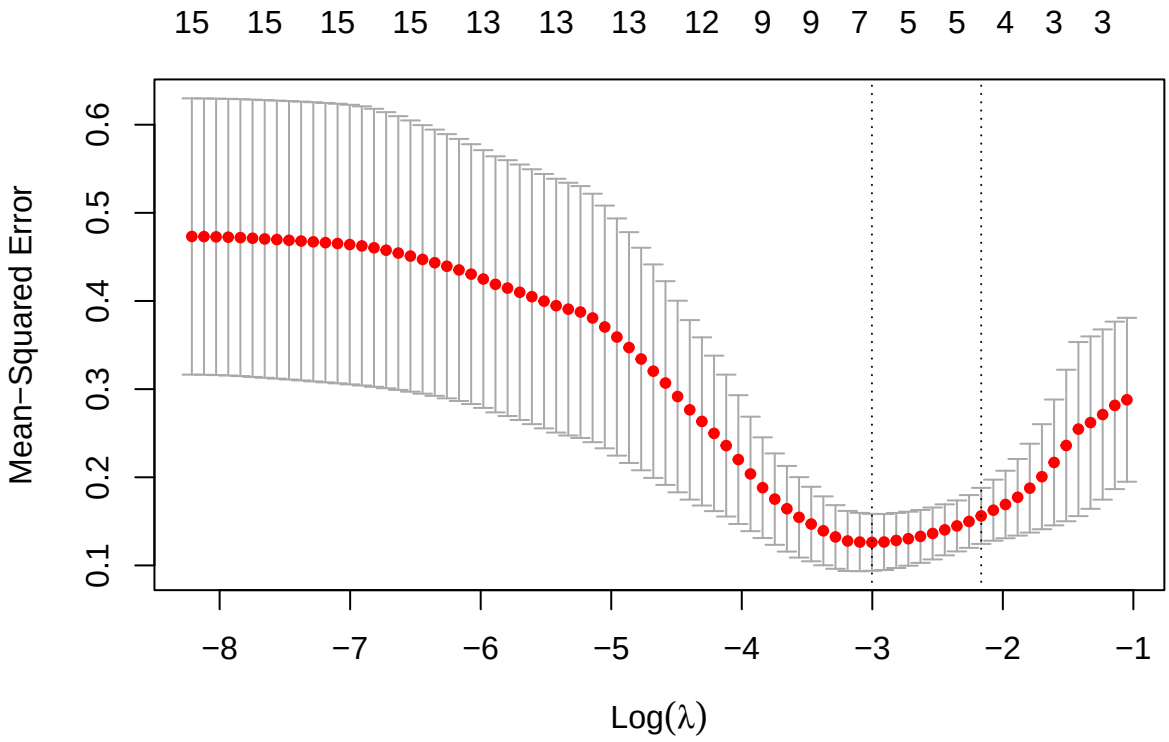
```r
x <- subset(data, select = -Birthweight)
y <- data$Birthweight

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]

y_predict_lm=predict(step_down_model,newdata=data[-train,]) # predict for the test rows
mse_lm=mean((y_test-y_predict_lm)^2)# prediction quality by the linear model

lasso_model <- glmnet(x_train, y_train, alpha = 1)

lasso_cv <- cv.glmnet(
    as.matrix(x_train),
    y_train,
    alpha = 1,
    type.measure="mse",
    nfolds=5
)

lasso_pred1=predict(lasso_model,s=lasso_cv$lambda.min,newx=as.matrix(x_test))
lasso_pred2=predict(lasso_model,s=lasso_cv$lambda.1se,newx=as.matrix(x_test))
mse1_lasso=mean((y_test-lasso_pred1)^2)
mse2_lasso=mean((y_test-lasso_pred2)^2)


plot(lasso_model, xvar="lambda", label = TRUE)
```




```
plot(lasso_model, xvar="dev", label = TRUE)
```




```
plot(lasso_cv)
```




```
coef(lasso_cv, s= lasso_cv$lambda.1se)
```

```
## 16 x 1 sparse Matrix of class "dgCMatrix"
##                    s1
## (Intercept) -0.7277
## ID           .
## Length       .
## Headcirc     0.0655
## Gestation    0.0422
## smoker      -0.0208
## mage         .
## mnocig       .
## mheight      .
## mppwt        0.0029
## fage         .
## fedyrs       .
## fnocig       .
## fheight      .
## lowbwt      -0.3202
## mage35       .
```

```
# coef(lasso_cv_model, s= lasso_cv_model$lambda.min)

mse_lm; mse1_lasso; mse2_lasso
```

```
## [1] 0.13
```

```
## [1] 0.17
```

```
## [1] 0.21
```

Question 2e)

```
data_new <- data[,c("lowbwt","Gestation","smoker","mage35")]
data_new$Gestation <- as.numeric(data_new$Gestation)
data_new$smoker <- factor(data_new$smoker, levels = 0:1, labels = c("No","Yes"))
data_new$mage35 <- factor(data_new$mage35, levels = 0:1, labels = c("No","Yes"))
```

To investigate "Do *smoking mothers* seem to have lighter babies?" and "Do *older mothers* seem to have lighter babies?", we can check the **Crosstabs**:

```
xtabs(lowbwt~smoker+mage35,data=data_new) / xtabs(~smoker+mage35,data=data_new)
```

```
##       mage35
## smoker    No   Yes
##    No  0.053 0.000
##    Yes 0.211 0.333
```

```
# Aggregate over mage35 and smoker
xtabs(lowbwt~smoker,data=data_new) / xtabs(~smoker,data=data_new)
```

```
## smoker
##   No  Yes
## 0.05 0.23
```

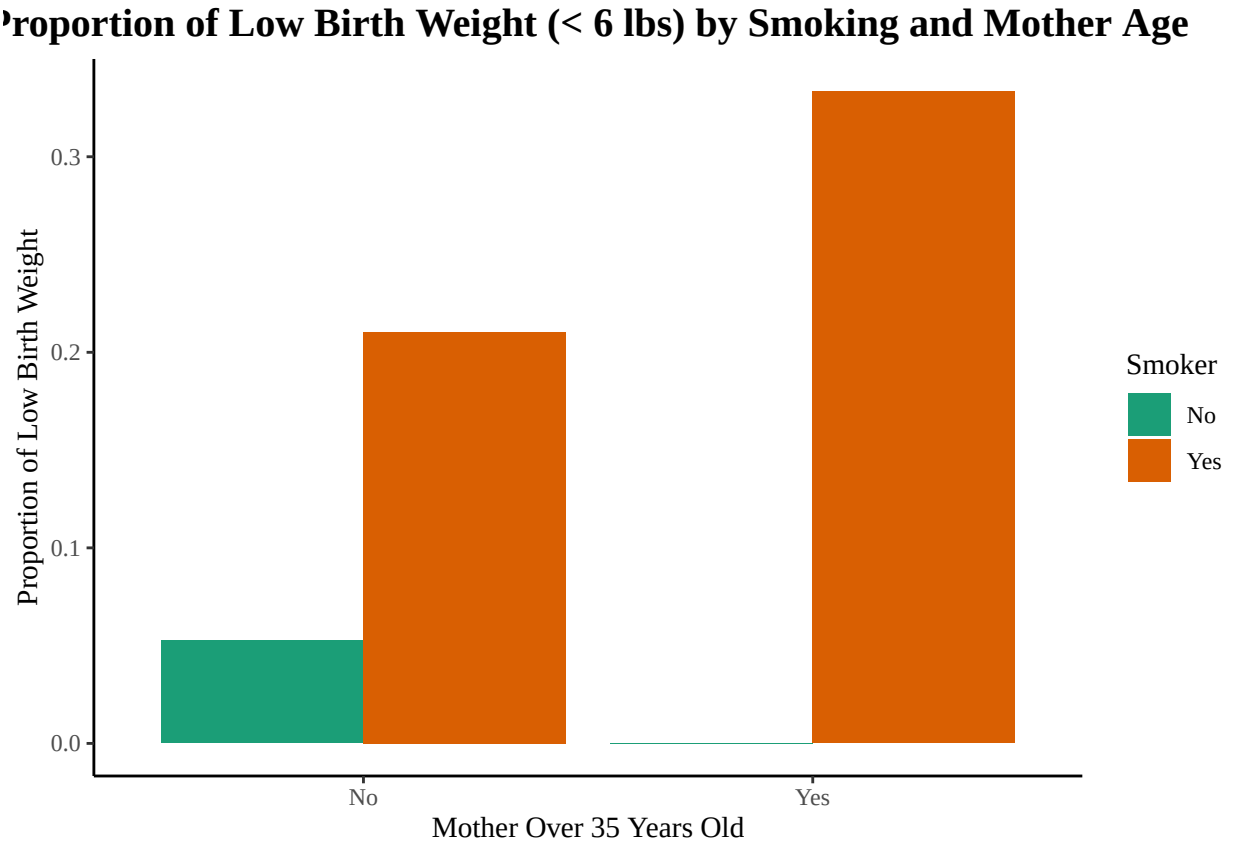
```
xtabs(lowbwt~mage35,data=data_new) / xtabs(~mage35,data=data_new)
```

```
## mage35
##   No  Yes
## 0.13 0.25
```

```
# Factorizing the DV after the crosstabs:
data_new$lowbwt <- factor(data_new$lowbwt, levels = 0:1, labels = c("No","Yes"))
```

To get an even more visual look, we can further plot these proportion of low birth weight by *Smoking* and *Mother Age*.

```
aggregated_data <- data %>%
  group_by(smoker, mage35) %>%
  summarise(proportion_dv = mean(lowbwt), .groups = 'drop')
aggregated_data$smoker <- factor(aggregated_data$smoker, levels = 0:1, labels = c("No","Yes"))
aggregated_data$mage35 <- factor(aggregated_data$mage35, levels = 0:1, labels = c("No","Yes"))

ggplot(aggregated_data, aes(x = mage35, y = proportion_dv, fill = as.factor(smoker))) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(x = "Mother Over 35 Years Old", y = "Proportion of Low Birth Weight", fill = "Smoker",
       title = "Proportion of Low Birth Weight (< 6 lbs) by Smoking and Mother Age") +
  theme_classic(base_family = "Times") + scale_fill_brewer(palette = color_choice) +
  theme(plot.title = element_text(face = "bold", hjust = 0.5, size = 15))
```



Based on the crosstabs and the bar plot:
1) It is immediately clear that there is *a higher proportion of low birth weight* in babies given birth by smoking mothers, compared to babies given birth by non-smoker mothers. If we aggregate over `mage35` variable in crosstabs, we can see that the difference between the proportions of 0.05 (non-smoker) and 0.23 (smoker) appears to be quite considerable. Hence, smoking mothers indeed seem to have lighter babies.
2) It appears "older mothers seem to have lighter babies", however the *main effect* is not as apparent as being a smoker.
- **Within the smoker mothers**, age of the mother seems to influence low birth weight: compared to smoker mothers that are younger than 35, smoker mothers that are 35 or older seem to have a higher proportion of Low Birth Weight in babies.
- **Within the non-smoker mothers**, age of the mother does not seem to influence low birth weight: there appears to be no difference between non-smoker mothers that are younger than 35 and non-smoker mothers that are 35 or older in Low Birth Weight in babies (since the difference in proportions of 0.05 and 0 seem negligible). - If we aggregate over `smoker` variable in crosstabs, we can see that there seems to be a difference between the proportions of 0.13 (under 35 y.o.) and 0.25 (over 35 y.o.). Yet, we need further testing to see whether this difference is actually significant. Although, considering the above points, it is possible that there is a `smoker*mage35` interaction.

Question 2f)

```
model_glm_0 <- stats::glm(lowbwt ~ 1, family=binomial("logit"), data=data_new)
model_glm_smoker <- stats::glm(lowbwt ~ smoker, family=binomial("logit"), data=data_new)
model_glm_mage35 <- stats::glm(lowbwt ~ mage35, family=binomial("logit"), data=data_new)
model_glm_both <- stats::glm(lowbwt ~ smoker + mage35, family=binomial("logit"), data=data_new)
```

```r
# Testing the Predictors through model comparison:
results_1 <- anova(model_glm_0, model_glm_smoker, test="Chisq")
results_2 <- anova(model_glm_0, model_glm_mage35, test="Chisq")
# Odds
exp(model_glm_smoker$coefficients[2])
```

```
## smokerYes
##       5.6
```

```r
exp(model_glm_mage35$coefficients[2])
```

```
## mage35Yes
##       2.2
```

The results of the binomial logistic regression model, with `glm`, shows that neither of the predictors have a significant main effect. We can see this through testing each of the predictors with making model comparison with `anova(model_1, model_2, test="Chisq")`.

We compared a base model, which includes only the intercept, to models including whether the mother is a smoker (`smoker`), age of the mother (`mage35`), and both of these predictors predictors (`smoker + mage35`) to assess the significance of each predictor. The analysis of deviance using Chi-square tests revealed that neither of the predictors are significant in their effect on low birth weight for newborns (`lowbwt`).

For the comparison between the base model and the model including `smoker`, there was no significant improvement in model prediction, $Deviance = 2.93$, $p = .087$. Similarly, for the comparison between the base model and the model including `mage35`, there was no significant improvement in model prediction, $Deviance = 0.35$, $p = .55$.

Although, it needs to be noted that the predictor of `smoker` is very close to significance with $p = .087$, which is in line with the graphs we have produced for 2e, although we would have expect it to be significant. Regarding `mage35`, based on the graphs, we had doubts whether this was a significant predictor, and now we can see that it is not.

When we investigate the **odds**, we can take the exponential of the coefficients in the model. The results of the models show that the odds of having a low birth weight baby for smoker mothers are about 5.6 times the odds for non-smoker mothers. And, the odds of having a low birth weight baby for mothers older than 35 are about 2.2 times the odds for mothers younger than 35. Although, as stated above, neither of these predictors seem to be significant with the default alpha level of 0.05.

Question 2g)

Investigate the interaction of predictor *Gestation* with *smoker*, and the interaction of *Gestation* with *mage35* (one interaction at a time). From this and f), choose a resulting model.

```r
model_glm_Gestation <- stats::glm(lowbwt ~ Gestation, family=binomial("logit"), data=data_new)
model_glm_int_smoker <- stats::glm(lowbwt ~ smoker*Gestation, family=binomial("logit"), data=data_new)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
model_glm_int_mage35 <- stats::glm(lowbwt ~ mage35*Gestation, family=binomial("logit"), data=data_new)

# Testing with ANOVA
```

```
results_1 <- anova(model_glm_0, model_glm_Gestation, test="Chisq")
results_2 <- anova(model_glm_Gestation, model_glm_int_smoker, test = "Chisq")
results_3 <- anova(model_glm_Gestation, model_glm_int_mage35, test = "Chisq")
```

We followed a similar model comparison procedure as 2f. As a first step, we can see that the model with `Gestation` as a predictor, compared to only the model with the intercept is significantly better in terms of model prediction, $Deviance = 16.4$, $p < .001$.

Given that the models with `smoker` and `mage35` predictors are not significantly better than the model with only the intercept, when we add the interaction terms (seperately to two models), we can deduce whether the interaction terms `smoker*Gestation` and `mage35*Gestation` provide significant improvement to model prediction.

We can see that the `smoker*Gestation` model is significantly better than just having `Gestation` as a predictor, $Deviance = 7.87$, $p = .02$. However, the model with `mage35*Gestation` model is not a significant improvement than just having `Gestation` as a predictor.

To note, running this `glm` model gives a `Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred`. Upon some research, we have discovered that this warning means that the model almost perfectly predict the outcome variable. We interpret this as a good sign.

However, we need to consider that the model improvement for `lowbwt ~ smoker * Gestation` might simply depend on the near-significant effect (see our answer for 2f) of `smoker` as a predictor. So, we can further test this by running a Chi-square test for the interaction model.

```
results <- anova(model_glm_int_smoker, test = "Chisq")
```

We can see that the interaction term `smoker:Gestation` is not significant, $p = 0.198$. We also further confirm this result through a model comparison with `smoker + Gestation` and `smoker*Gestation`:

```
model_glm_noint_smoker <- stats::glm(lowbwt ~ smoker + Gestation, family=binomial("logit"), data=data_n
anova(model_glm_noint_smoker, model_glm_int_smoker, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: lowbwt ~ smoker + Gestation
## Model 2: lowbwt ~ smoker * Gestation
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        39       11.8
## 2        38       10.2  1     1.66      0.2
```

Hence, even though `smoker` variable is nearly non-significant, due to the model comparison results, we choose our final model as `lowbwt ~ smoker + Gestation`.

Question 2h)

h) For the resulting model from g), estimate the probability of low baby weight for each combination of levels of the involved factors and the gestation of 40 weeks.

```
new_data <- data.frame(Gestation = 40, smoker = factor(c("No", "Yes"), levels = c("No", "Yes")))
probabilities <- predict(model_glm_noint_smoker, newdata = new_data, type = "response")
new_data$probability_lowbwt <- probabilities
print(new_data)
```

```
##   Gestation smoker probability_lowbwt
## 1        40     No             0.00007
## 2        40    Yes             0.01651
```

Question 2i)

i) Another approach to address the questions in e) would be to apply a contingency table test. Implement the relevant test(s). Is this approach wrong? Name both an advantage and a disadvantage of this approach as compared to the one from f).

```
## Chi-squared test for smoking mothers
smoker_table <- table(data_new$smoker, data_new$lowbwt)
chisq_test_smoker <- chisq.test(smoker_table)
```

```
## Warning in chisq.test(smoker_table): Chi-squared approximation may be incorrect
```

```
chisq_test_smoker
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  smoker_table
## X-squared = 1, df = 1, p-value = 0.2
```

```
## Chi-squared test for mothers over 35
mage35_table <- table(data_new$mage35, data_new$lowbwt)
chisq_test_mage35 <- chisq.test(mage35_table)
```

```
## Warning in chisq.test(mage35_table): Chi-squared approximation may be incorrect
```

```
chisq_test_mage35
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  mage35_table
## X-squared = 3e-31, df = 1, p-value = 1
```