Additional Homework 9

Tauqeer Kasam Rumaney

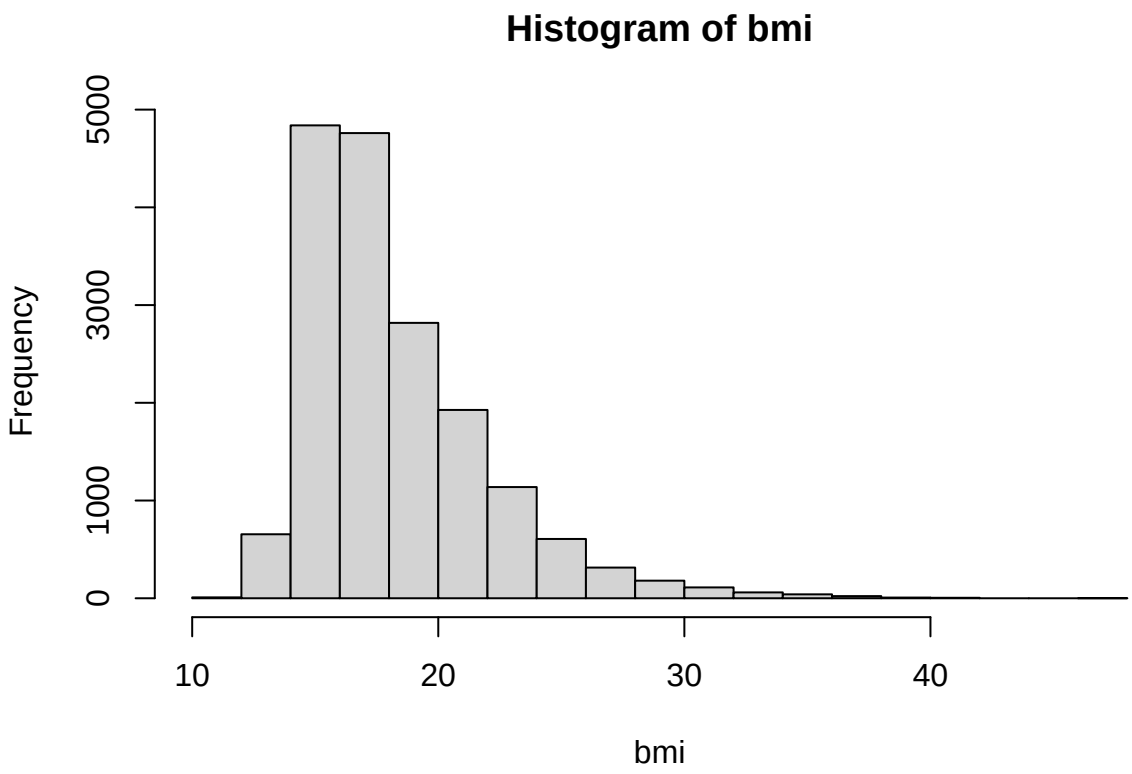
December 17, 2022

Download this R Markdown file, save it on your computer, and perform all the below tasks by inserting your answer in text or by inserting R chunks below. After you are done, upload this file with your solutions on Moodle.

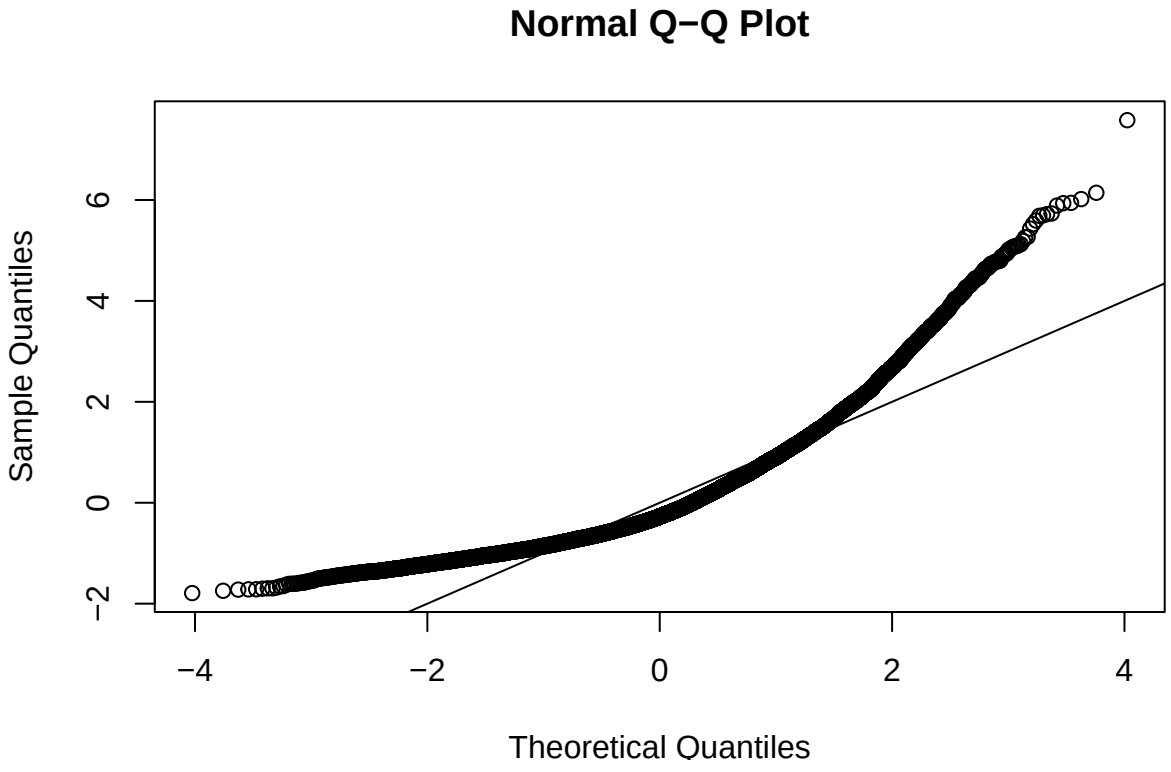
Exercise 1: Assumptions of linear regression

Load the KiGGS dataset and compute a regression predicting BMI by sex and age groups (age2):

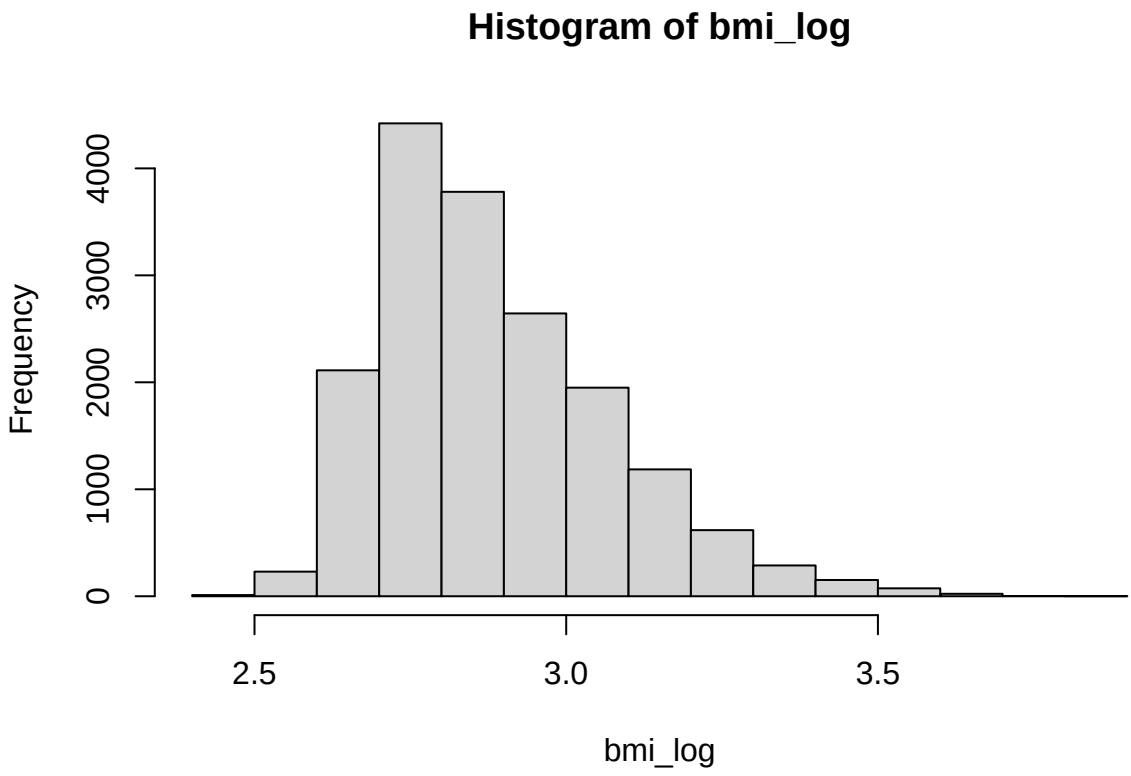
```
# load data
dat_link <- url("https://www.dropbox.com/s/pd0z829pv2otzqt/KiGGS03_06.RData?dl=1")
load(dat_link)
dat <- KiGGS03_06
bmi <- dat$bmiB
sex <- dat$sex
age <- dat$age2
bmi_log <- log(bmi)

# Regression:
fit1 <- lm(bmi ~ sex + age)

# Histogram of Y (BMI):
hist(bmi)
```



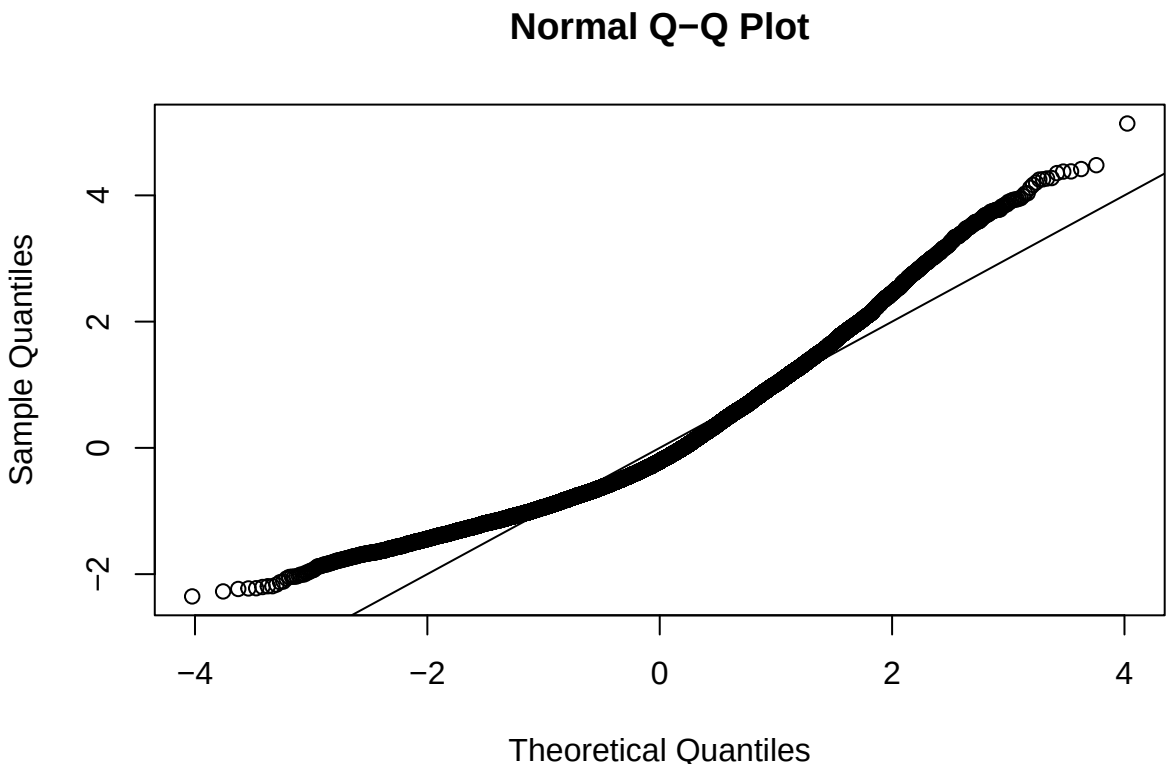

```
qqnorm(scale(bmi)); abline(0,1)
```



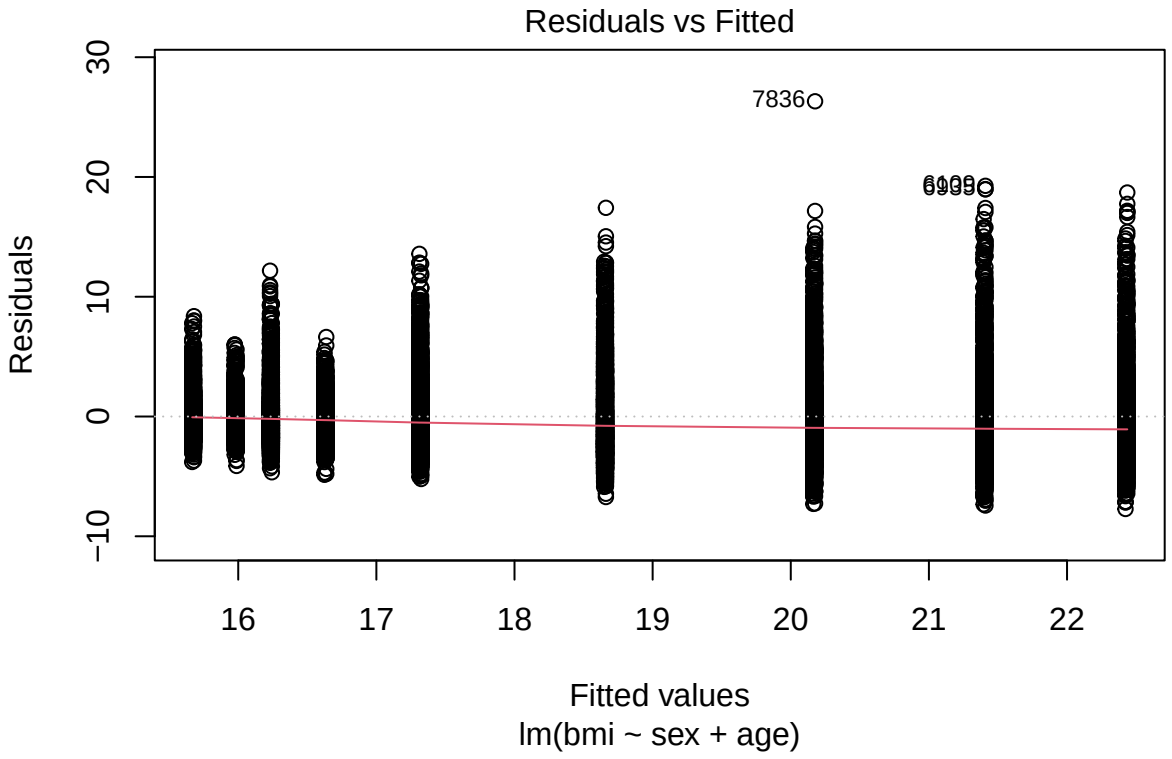

```
hist(bmi_log)
```



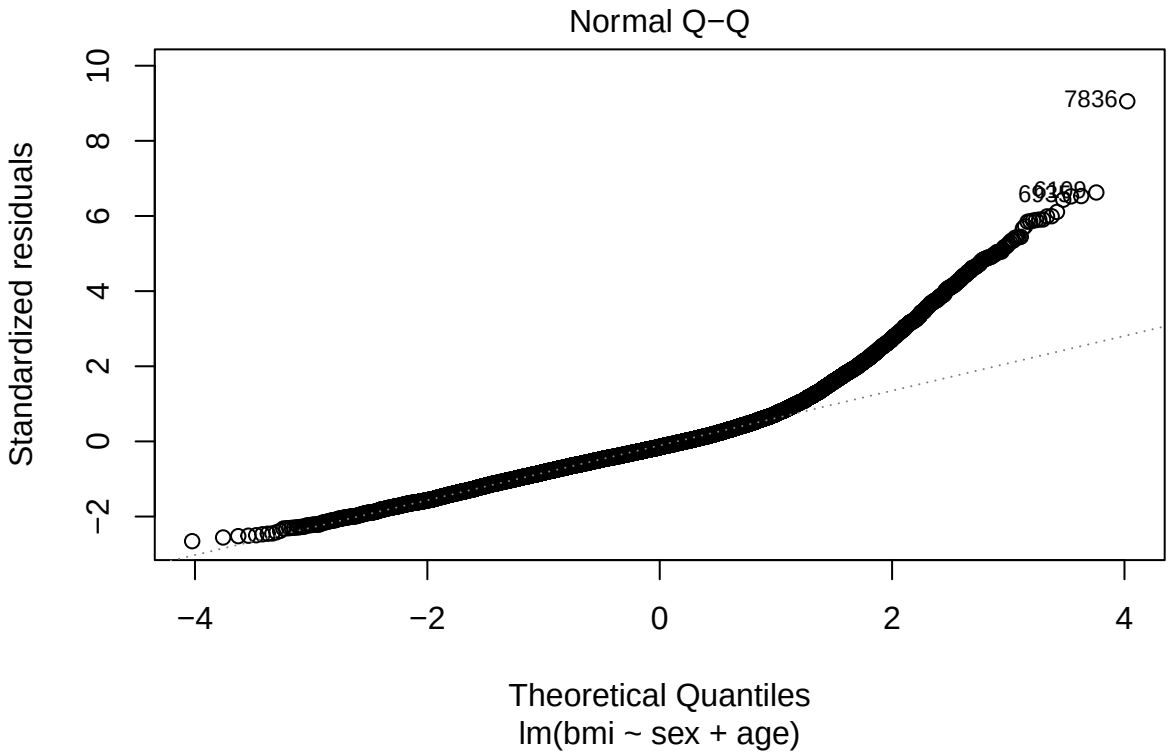

```
qqnorm(scale(bmi_log)); abline(0,1)
```



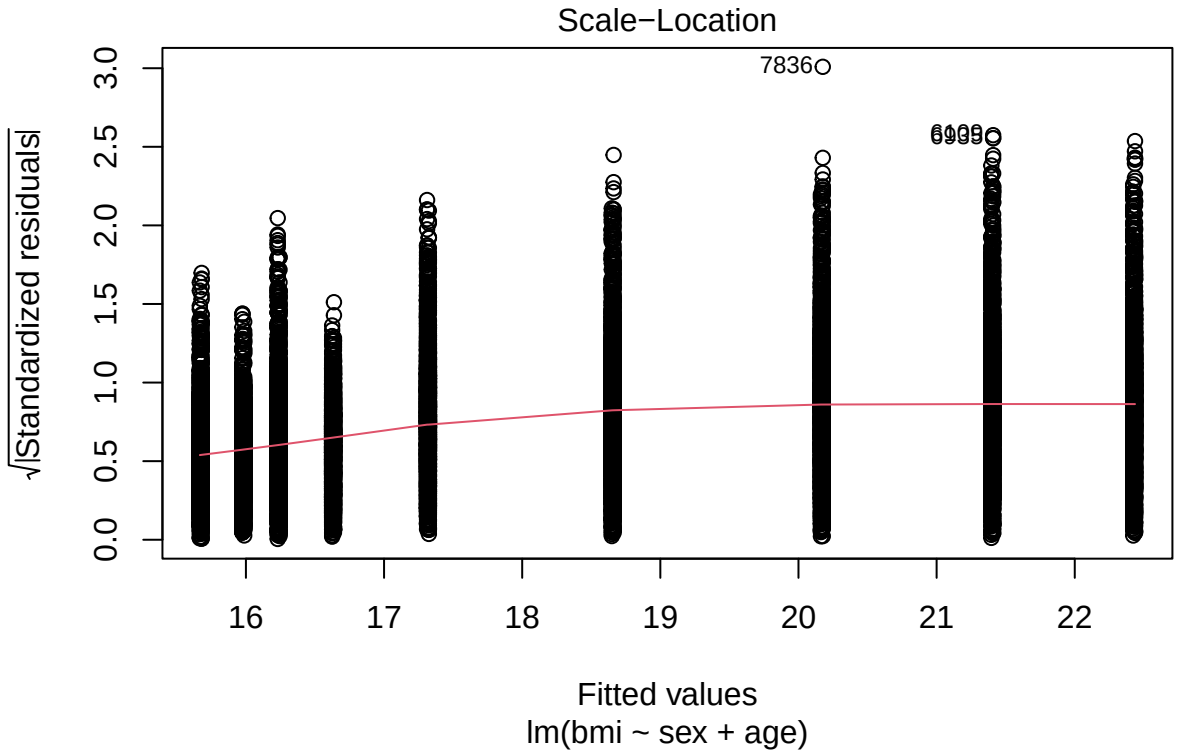

```
# Linearity of Data (first plot):
plot(fit1,1)
```




```
#In the graph, we can see that there is no particular pattern as well as the horizontal line is close t

# Distribution of residuals (second plot):
plot(fit1,2)
```

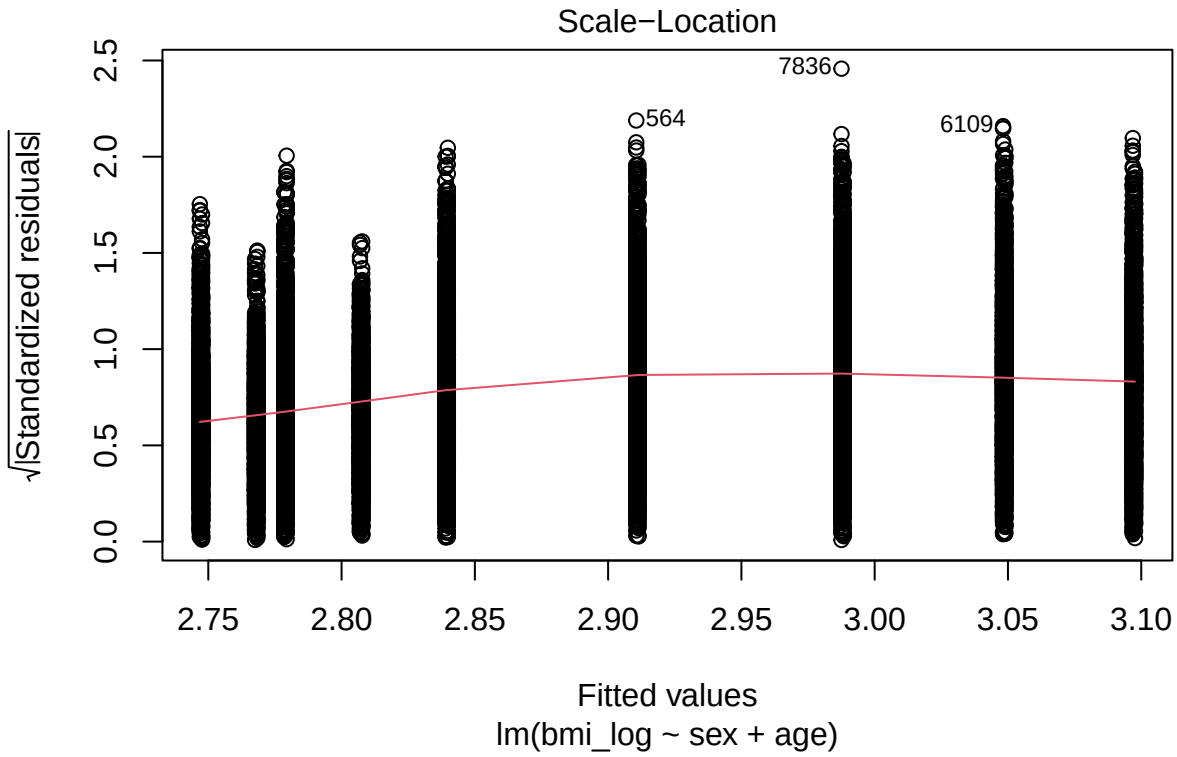


```
#Q-Q plot shows normality of Residues, and as we can see the distribution follows the straight line in

# Homoscedasticity of the residuals (third plot)
plot(fit1,3)
```




```
fit2 <- lm(bmi_log ~ sex + age)
plot(fit2, 3)
#install.packages('lmtest')
library(lmtest)
```

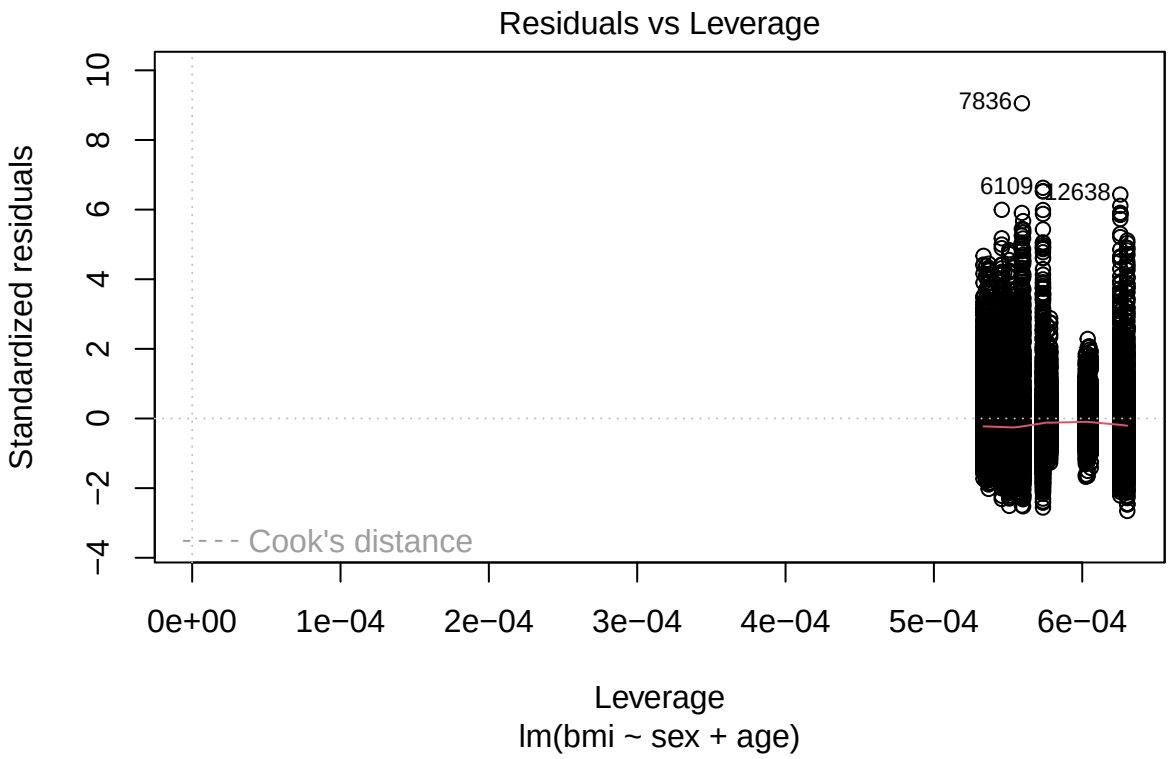
```
## Loading required package: zoo
##
## Attaching package: 'zoo'
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```




```
bptest(fit2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fit2
## BP = 1059.7, df = 9, p-value < 2.2e-16
```

```
#p < 0.05, suggesting that our data is not homoscedastic.

# Outliers and high levarage points (fifth plot)
plot(fit1,5)
```



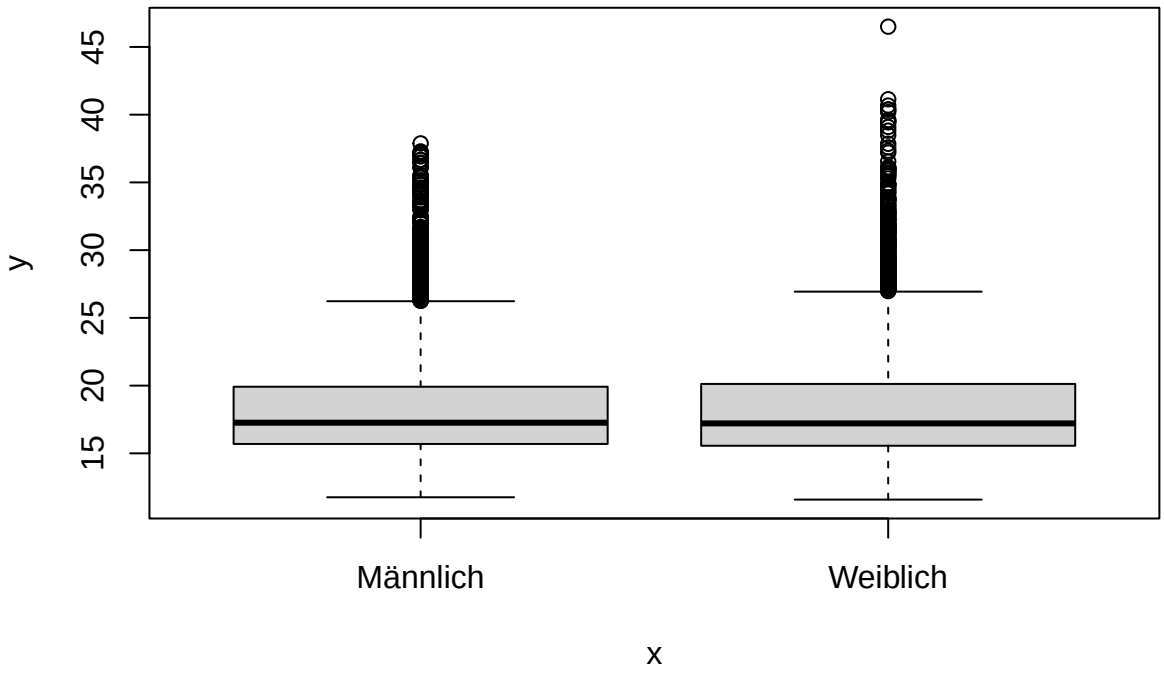

```
#The plot above highlights the top 3 most extreme points (#6109, #7836 and #12638), with a standardized
#Additionally, there is no high leverage point in the data. That is, all data points, have a leverage s

# Independence of Predictors
library(car)
```

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

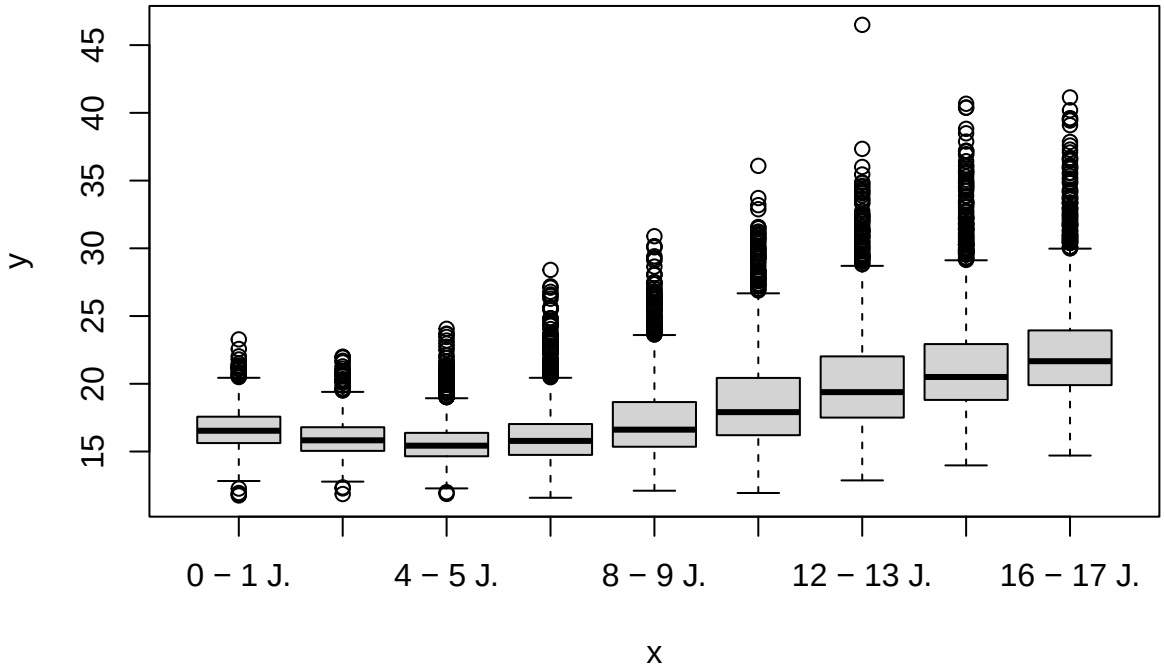
```
durbinWatsonTest(fit1)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      0.02214128      1.955575   0.006
##  Alternative hypothesis: rho != 0
```

```
#p < 0.05, so the errors are autocorrelated. We have violated the independence assumption.

# Linear relationship of Y with predictors:
plot(sex, bmi)
```




```
plot(age, bmi)
```




```
# Multicollinear
cor(as.numeric(sex), as.numeric(age), use = "complete.obs")
```

```
## [1] -0.006083543
```

```
#Since the value is very low, we can conclude that both the variables "sex" and "age2" are not related.

# results:
summary(fit1)
```

```
##
## Call:
## lm(formula = bmi ~ sex + age)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.7198 -1.7437 -0.4326  1.1148 26.3190
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.62433    0.07137 232.918  < 2e-16 ***
## sexWeiblich    0.01294    0.04399   0.294    0.769
## age2 - 3 J.   -0.64945    0.09617  -6.753 1.49e-11 ***
## age4 - 5 J.   -0.95770    0.09494 -10.088  < 2e-16 ***
## age6 - 7 J.   -0.39391    0.09374  -4.202 2.66e-05 ***
## age8 - 9 J.    0.68712    0.09301   7.388 1.56e-13 ***
## age10 - 11 J.  2.02457    0.09332  21.694  < 2e-16 ***
## age12 - 13 J.  3.53873    0.09394  37.669  < 2e-16 ***
## age14 - 15 J.  4.77202    0.09447  50.513  < 2e-16 ***
## age16 - 17 J.  5.79891    0.09717  59.676  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.908 on 17483 degrees of freedom
##   (147 observations deleted due to missingness)
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.3905
## F-statistic:  1246 on 9 and 17483 DF,  p-value: < 2.2e-16
```

In this model, investigate and judge whether the assumptions listed on slide 13 in lecture 9 are satisfied.

Exercise 2: Model selection in linear regression (optional)

In the KiGGS dataset, aim to select relevant predictors for sys12 (systolic blood pressure). Use 2 of the model selection approaches described on slide 26, apply them to the KiGGS dataset and compare the results.

Exercise 3: Linear regression with multiple imputation (optional)

Run the code in the Rmd file R_9b_linear_regression_MI.Rmd, inspect the R code what it is doing, and look at the results. Apply the same to the linear regression model of another variable of your choice.