Biostatistics: Exercise 08

Beate Sick, Lisa Herzog

3.11.2020

Exercise 01: Linear Regression I

We consider the `agefat` dataset from library `HSAUR2`.

a) Investigate the relationship between age (`age`) and body fat percentage (`fat`) as well as between gender (`gender`) and body fat percentage graphically.

```
library(HSAUR2)
```

```
## Loading required package: tools
```

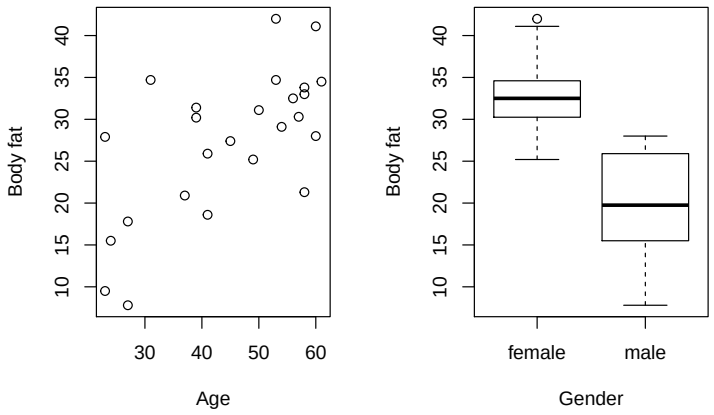
```
data("agefat")
dat = agefat
head(dat)
```

```
##   age  fat gender
## 1  24 15.5   male
## 2  37 20.9   male
## 3  41 18.6   male
## 4  60 28.0   male
## 5  31 34.7 female
## 6  39 30.2 female
```

```
par(mfrow=c(1,2))

# Since age and body fat are continuous, we use a scatterplot
plot(dat$age, dat$fat,
     ylab="Body fat",
     xlab="Age")

# Since gender is dichotomous, we use a boxplot
boxplot(dat$fat~dat$gender,
        names = c('female','male'),
        ylab='Body fat',
        xlab='Gender')
```



b) Fit a linear regression model using the `lm()` function. Use age and gender as covariates. Interpret the estimates for the intercept, age and gender. (**R-Hint:** to fit the model use `mod` ← `lm()`. To consider the results use `summary(mod)`)

```
mod = lm(fat ~ age + gender, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = fat ~ age + gender, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.4691 -2.0505  0.0276  2.1442  8.2773
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.64786    4.10781   4.783 8.92e-05 ***
## age           0.26556    0.07953   3.339  0.00297 **
## gendermale  -10.54892    2.09140  -5.044 4.74e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.644 on 22 degrees of freedom
## Multiple R-squared:  0.7375, Adjusted R-squared:  0.7137
## F-statistic: 30.91 on 2 and 22 DF,  p-value: 4.069e-07
```

```
# Intercept: mean body fat percentage if both explanatory
# variables are 0, i.e. if age=0 and gender=female (response).

# Age: If age increases by one unit, the mean body fat precentage increases
# by 0.2 units given gender stays contant. That is, with increasing age,
# body fat percentage increases if gender is the same.

# Gender: The mean body fat percentage is about 10 units smaller for males
# than for females, given that age doesn't change.
```

c) Check the model assumptions using a Tukey-Ascombe and a normal QQ-plot (**R-Hint**: To get the

fitted values and the residuals for the Tukey-Ascombe plot, you can use `fitted(mod)` and `resid(mod)`. For the QQ-plot use the function `qqPlot()` from library `car`).

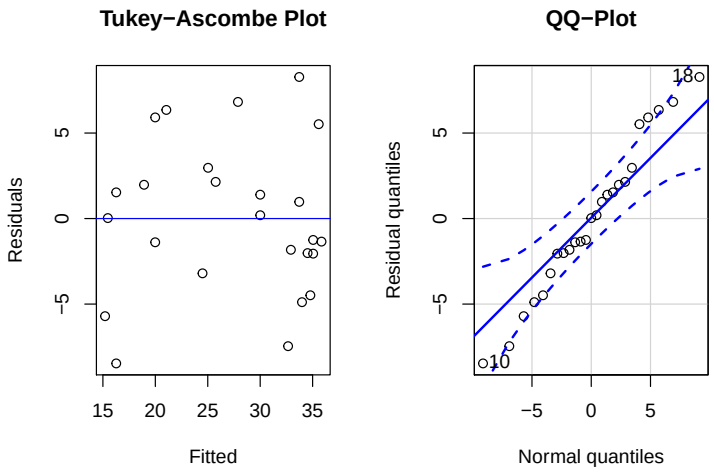
```
par(mfrow=c(1,2))

# TA plot: fitted vs. residuals
plot(fitted(mod), resid(mod),
     main='Tukey-Ascombe Plot',
     ylab='Residuals',
     xlab='Fitted')
abline(a=0, b=0, col='blue')

# Normal QQ-plot with the residuals
library(car)
```

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

```
qqPlot(resid(mod), dist = "norm",
  mean = mean(resid(mod)),
  sd = sd(resid(mod)),
  ylab='Residual quantiles',
  xlab='Normal quantiles',
  main="QQ-Plot")
```




```
## [1] 10 18
```

```
# The model assumptions hold.
```

d) Write down the equation of the model (Y = ...). Then predict the mean body fat percentage for a 40 year old woman. You can calculate it by hand or by using the R-function `predict()`.

```
# The formular is:
# Y = 19.65 + 0.27*age - 10.55*gender

# The mean body fat percentage for a 40 year old woman is:
# Y = 19.65 + 0.27*40 - 10.55*0 = 30.45
predict(mod, newdata=data.frame(age=40, gender="female"))
```

```
##        1
## 30.27037
```

e) Given the model is correct - how much will the body fat change on average if a person gets 2 years older.

```
# The formular is:
# Y = 19.65 + 0.27*age - 10.55*gender

# The change of the mean body fat percentage when increasing the age by 2 years
# is given by 2-times the age-coefficient:
# 2*0.27=0.54 is the mean fat percentage increase
# According to this linear model, this is valid for every specific value of age' and sex':
# Y(age'+2, sex')-Y(age', sex')
# = 19.65 + 0.27*(age'+2) - 10.55*(sex') -(19.65 + 0.27*(age') - 10.55*(sex*))
# = 2*0.27  # 0.54
```

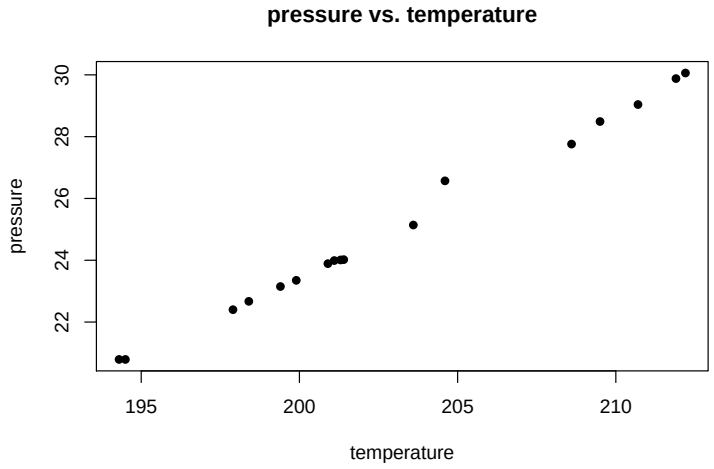
Exercise 2: Linear regression II]

The data set of Forbes lists the boiling point of water (in ^{o}F) and the atmospheric pressure (in inches of mercury) at different places in the alps. We want to investigate the association between the temperature (`Temp`) and the pressure (`Press`). You can read the data into R using:

```
url = "https://polybox.ethz.ch/index.php/s/uZJZavllfYbxldy/download"
dat = read.table(url, header = TRUE)
```

a) Investigate the relationship between pressure and temperature graphically. Is it reasonable to fit a linear regression model?

```
plot(dat$Temp, dat$Press, pch=16,
     main='pressure vs. temperature',
     ylab='pressure',
     xlab='temperature')
```




```
# If we consider the data it seems to be reasonable to fit a linear regession
# model. We see, that the atmospheric pressure increases with the increase
# of the boiling point of water.
```

b) Perform a linear regression. Investigate the influence of the temperature (covariate) on the pressure (outcome).

```r
mod = lm(Press~Temp, data = dat)
summary(mod)
```

```
##
## Call:
## lm(formula = Press ~ Temp, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25717 -0.11246 -0.05102  0.14283  0.64994
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -81.06373    2.05182  -39.51   <2e-16 ***
## Temp          0.52289    0.01011   51.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2328 on 15 degrees of freedom
## Multiple R-squared:  0.9944, Adjusted R-squared:  0.9941
## F-statistic:  2677 on 1 and 15 DF,  p-value: < 2.2e-16
```

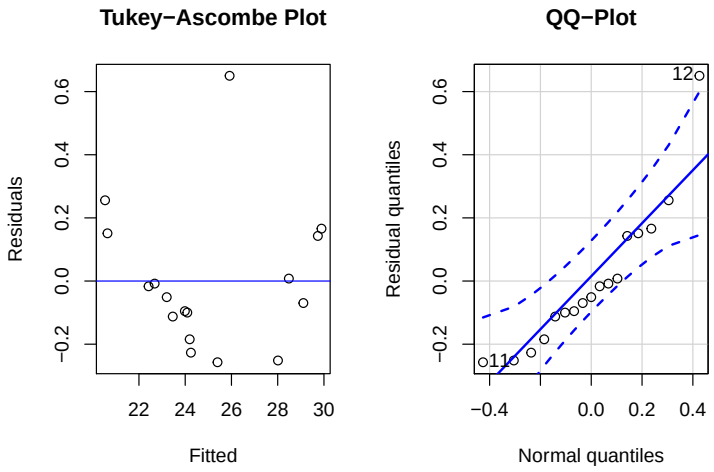
```r
# If the temperature increases by one point, the atmospheric pressure
# increases by 0.5 inches. That is, the higehr the temperature for the
# boiling point of the water, the higher the atmospheric pressure.
```

c) Generate a Tukey-Anscombe plot and a normal Q-Q plot of the residuals. Are there any hints that the model assumptions are violated?

```r
par(mfrow=c(1,2))

# TA plot: fitted vs. residuals
plot(fitted(mod), resid(mod),
     main='Tukey-Ascombe Plot',
     ylab='Residuals',
     xlab='Fitted')
abline(a=0, b=0, col='blue')

# Normal QQ-plot with the residuals
library(car)
qqPlot(resid(mod), dist = "norm",
  mean = mean(resid(mod)),
  sd = sd(resid(mod)),
  ylab='Residual quantiles',
  xlab='Normal quantiles',
  main="QQ-Plot")
```




```
## [1] 12 11
```

```
# The Tukey Ascombe plot shows that there are some violations of the
# assumptions. The residuals are not in a band around the horizontal
# line.
# The QQ-plot looks good but it shows an outlier (observation 12)
```

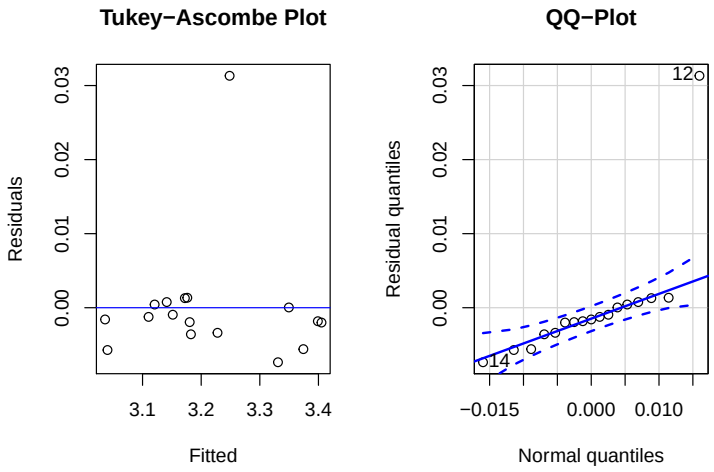
d) To get a better fit, we transform the outcome variable `Press`. Calculate the log of the variable and fit a linear regression model with the transformed outcome variable. Calculate a Tukey-Ascombe and a QQ-plot. What about the model assumptions?

```
# calculate the log
dat$log_Press = log(dat$Press)
log_mod = lm(log_Press ~ Temp, data = dat)

par(mfrow=c(1,2))

# TA plot: fitted vs. residuals
plot(fitted(log_mod), resid(log_mod),
     main='Tukey-Ascombe Plot',
     ylab='Residuals',
     xlab='Fitted')
abline(a=0, b=0, col='blue')

# Normal QQ-plot with the residuals
library(car)
qqPlot(resid(log_mod), dist = "norm",
  mean = mean(resid(log_mod)),
  sd = sd(resid(log_mod)),
  ylab='Residual quantiles',
  xlab='Normal quantiles',
  main="QQ-Plot")
```




```
## [1] 12 14
```

```
# The Tukey Ascombe plot shows no violation anymore.
# However, the outlier is still a problem which  should be investigated further.
# Assume that it turned out, that the value is not valid since the temperature meter was uncalibrated.
```

e) Identify and remove the outlier. Calculate a Tukey-Ascombe and a QQ-plot. What about the model assumptions now?

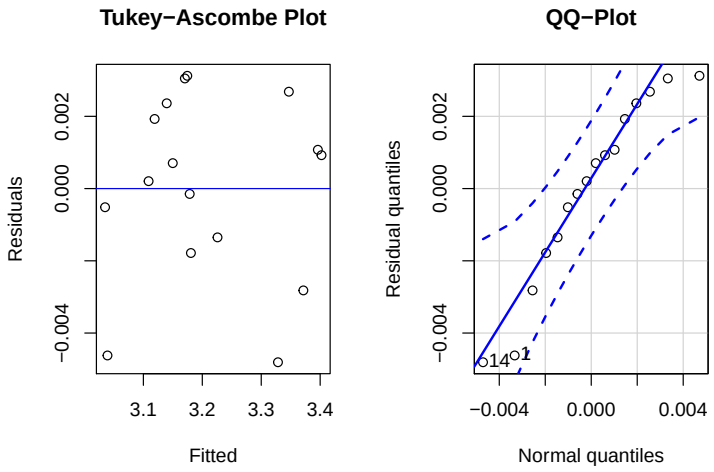
```
# Since we discovered taht the extreme observation 12 is unreliable, we remove it from the data
dat_new = dat[-12,]

mod_new = lm(log_Press ~ Temp, data = dat_new)

par(mfrow=c(1,2))

# TA plot: fitted vs. residuals
plot(fitted(mod_new), resid(mod_new),
     main='Tukey-Ascombe Plot',
     ylab='Residuals',
     xlab='Fitted')
abline(a=0, b=0, col='blue')

# Normal QQ-plot with the residuals
library(car)
qqPlot(resid(mod_new), dist = "norm",
  mean = mean(resid(mod_new)),
  sd = sd(resid(mod_new)),
  ylab='Residual quantiles',
  xlab='Normal quantiles',
  main="QQ-Plot")
```




```
## 14  1
## 13  1
```

```
# All assumptions hold now.
```