### Exercise 1

The file `catheter.rda` can be downloaded from the website and can be read with `load()`.

```
dat = load("data/catheter.rda")
```

The variable `height` describes the height of a patient in cm, the variable `weight` describes his weight in kg. The target variable `catlength` is the optimal length of a catheter that is used for an examination of the heart. The goal is to estimate this quantity from the available data set.

(a) Do a simple linear regression for both `catlength` $\sim$ `height` and `catlength` $\sim$ `weight`. Are the predictors significant?

```
# linear regressions
mod1 <- lm(catlength ~ height, data=catheter)
mod2 <- lm(catlength ~ weight, data=catheter)
summary(mod1)
```

```
##
## Call:
## lm(formula = catlength ~ height, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0929 -0.7298 -0.2608  1.1652  6.6879
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.12706    4.24700   2.855 0.017090 *
## height       0.23774    0.04034   5.893 0.000152 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.009 on 10 degrees of freedom
## Multiple R-squared:  0.7764,Adjusted R-squared:  0.7541
## F-statistic: 34.73 on 1 and 10 DF,  p-value: 0.0001525
```

```
summary(mod2)
```

```
##
## Call:
## lm(formula = catlength ~ weight, data = catheter)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.9676 -1.4963 -0.1386  2.0980  7.0205
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 25.62631    2.00264  12.796 1.59e-07 ***
## weight       0.61613    0.09759   6.313 8.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.797 on 10 degrees of freedom
## Multiple R-squared:  0.7994,Adjusted R-squared:  0.7794
## F-statistic: 39.86 on 1 and 10 DF,  p-value: 8.755e-05

# The predictor is highly significant in both cases
```

(b) Fit a multiple linear regression `catlength` $\sim$ `height + weight`. Is there an influence of the predictors on the target overall? Is it significant?

```
# multiple regression
mod   <- lm(catlength ~ height + weight, data=catheter)
summary(mod)

##
## Call:
## lm(formula = catlength ~ height + weight, data = catheter)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0497 -1.2753 -0.2595  1.9095  6.9933
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21.08527    8.77037   2.404   0.0396 *
## height       0.07681    0.14412   0.533   0.6070
## weight       0.42752    0.36810   1.161   0.2753
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 3.94 on 9 degrees of freedom
## Multiple R-squared:  0.8056,Adjusted R-squared:  0.7624
## F-statistic: 18.65 on 2 and 9 DF,  p-value: 0.0006301

# Yes, there is an influence of the predictors on the target variable
# overall. This is assessed by the global F-test. Its p-value is smaller
# than 0.01 so that the null hypothesis is rejected at the 1\% level.
# At least one of the predictors is necessary.
```

(c) Test the null hypotheses $H_0 : \beta_1 = 0$ and $H_0 : \beta_2 = 0$. Compare the results with those from the two simple linear regressions. Comment and explain the differences if there are any.

```
# As we can see from the summary output (see above), both null hypotheses
# are retained, i.e. the predictors are not significant. Is this a
# contradiction to the results from the two simple linear regressions?
# No, in multiple regression the hypotheses tests assess whether we need
# (e.g.) the predictor height when we already know the predictor weight.
# The answer is no and the same holds vice versa.
# On the other hand, the global F-test indicates that we need at least
# one of the two predictors. So we do not need to include both predictors
# simultaneously but we need one of them. This situation occurs when the
# predictors are strongly correlated. Due to the smaller p-value we would
# prefer the predictor weight in this case.
```

(d) For a child that is 120cm tall and has a weight of 25kg, compute the 95% prediction interval with the multiple regression model as well as with the simple regression models. In practice, a prediction error of $\pm$ 2cm would be acceptable. Do the data and the models allow for a prediction of `catlength` that is sufficiently precise? Does it make sense to use both predictors?

```
## prediction intervals
newdat <- data.frame(height=120, weight=25)
predict(mod1, newdata=newdat, interval="prediction")

##        fit      lwr      upr
## 1 40.65609 31.20891 50.10327

predict(mod2, newdata=newdat, interval="prediction")

##        fit      lwr      upr
## 1 41.02954 32.06162 49.99747
```

```
predict(mod,   newdata=newdat, interval="prediction")

##        fit      lwr      upr
## 1 40.99072 31.53989 50.44154

# The predictions differ slightly. We note that the prediction interval
# is not shortest for the multiple regression model which one might
# expect since it uses the largest amount of information. However, the
# multiple model requires estimating one additional parameter based on
# the available 12 data points. This is associated with a larger
# estimation error of each single parameter. In most practical cases
# the prediction accucacy increases by including an additional parameter
# but in our case the increased estimation error has a stronger, negative
# influence. This is due to the fact that the two predictors are strongly
# correlated -- adding the second predictor when the first one is already
# present does hardly yield additional information.
#
# In practice, a prediction error of $\pm$ 2cm would be acceptable.
# Thus, the data and the models do not allow for a prediction of
# catlength that is sufficiently precise.
```

**Exercise 2**

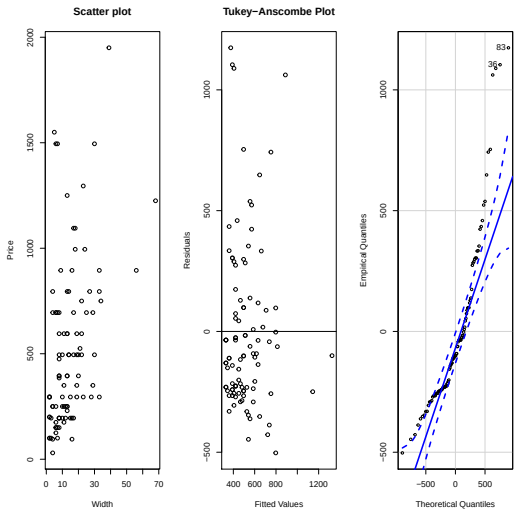
The left figure shows the price of 100 books ($y$; in Pence) as a function of their width ($x$; in mm). The data were taken for the estimation of a potential damage loss of a household insurance. The following linear regression model has been fitted to the data:

$$y_i = \alpha + \beta x_i + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2) iid$$

Here, you see a part of the R output:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  300.485     57.468   5.229      ???
## Dicke         15.071      3.171   4.752      ???
##
## Residual standard error: ??? on 98 degrees of freedom
## Multiple R-squared: 0.1873,Adjusted R-squared: 0.179
```

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




```
## [1] 83 36
```

(a) There is a significant correlation between width and price of books ($\beta$ is significantly different from 0).

(a) true

(b) false

```
# true, the p-value corresponding to the t-value is
(p.val <- (1 - pt(4.752, 98))*2)

## [1] 6.904541e-06
```

(b) Which of the following intervals is an exact 95% confidence interval for $\beta$ under the the assumption of normally distributed errors?

(a) $15.071 \pm 1.984 \cdot 3.171$

(b) $15.071 \pm 1.984 \cdot 4.752$

(c) $15.071 \pm \frac{1.984}{\sqrt{100}} \cdot 3.171$

(d) $15.071 \pm \frac{1.984}{\sqrt{100}} \cdot 4.752$

(e) None of the indicated intervals

```
# (a) the quantile is
qt(0.975,df=98)

## [1] 1.984467
```

(c) What's the estimate for $\hat{\sigma}$ approximately ("?" in the output)?

(a) $0 \leq \hat{\sigma} < 10$

(b) $10 \leq \hat{\sigma} < 100$

(c) $100 \leq \hat{\sigma} < 1000$

(d) $1000 \leq \hat{\sigma}$

```
# (c) the standard deviation of the errors can be estimated by
# looking at the distribution of the residuals in the Tukey-
# Anscombe plot.
```

(d) How much does a book of a width of 30 mm approximately cost (in Pence), based on the regression fit?

(a) 500

(b) 750

(c) 1000

(d) 1250

(e) 1500

```
# (b) Based on the estimated coefficients, the predicted cost is
300.485 + 30*15.071

## [1] 752.615
```

(e) Do the model assumptions hold for the fitted data set?

(a) The connection between width and price is non-linear.

(b) The errors are not normally distributed.

(c) No deviations from the model assumptions are visible.

```
# (b) can be seen from the QQ Plot
```

**Exercise 3**

The following data give the income, number of cows and area for a number of American farms.

| Income (Dollar) | 960 | 830 | 1260 | 610 | 590 | 900 | 820 | 880 | 860 | 760 |
|---|---|---|---|---|---|---|---|---|---|---|
| Number of cows (cows) | 18 | 0 | 14 | 6 | 1 | 9 | 6 | 12 | 7 | 2 |
| Size of farm (acres) | 60 | 220 | 180 | 80 | 120 | 100 | 170 | 110 | 160 | 230 |
| | | | | | | | | | | |
| Income (Dollar) | 1020 | 1080 | 960 | 700 | 800 | 1130 | 760 | 740 | 980 | 800 |
| Number of cows (cows) | 17 | 15 | 7 | 0 | 12 | 16 | 2 | 6 | 12 | 15 |
| Size of farm (acres) | 70 | 120 | 240 | 160 | 90 | 110 | 220 | 110 | 160 | 80 |

To these data, the linear regression model

$$\text{Dollar}_i = \beta_0 + \beta_1 \text{cows}_i + \beta_2 \text{acres}_i + E_i$$

with iid $E_i \sim \text{N}(0, \sigma^2)$ was fitted.
This is part of the output from R:

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  285.457     81.379   3.508   0.0027 **
## cows          32.569      3.728     ??? 1.08e-07 ***
## acres          2.138      0.394   5.434 4.47e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 76.45 on ??? degrees of freedom
## Multiple R-squared: 0.8179,Adjusted R-squared: 0.7965
## F-statistic: 38.17 on ??? and ??? DF,  p-value: 5.165e-07
```

(a) The size of a farm has a statistically significant influence on its income.
   (a) True
   (b) False

```
# (a) the p-value of the variable ``acres'' is clearly below $0.05$.
```

(b) The number of cows on a farm has a statistically significant influence on its income.
   (a) True
   (b) False

```
# (a) the p-value of the variable ``cows'' is clearly below $0.05$.
```

(c) What is the outcome of the test of the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$?

(a) Keep $H_0$

(b) Reject $H_0$

```
# (b)
```

(d) How many degrees of freedom are there in this model fit?

(a) $\infty$

(b) 20

(c) 18

(d) 17

(e) 3

```
# (d)
```

(e) Which of the following is an exact 95% confidence interval for $\beta_1$?

(a) $32.569 \pm 2.11 \cdot 3.7276$

(b) $32.569 \pm 1.96 \cdot 3.7276$

(c) $32.569 \pm \frac{2.11}{\sqrt{17}} \cdot 5.45$

(d) None of these

```
# (a)
qt(0.975,df=17)

## [1] 2.109816
```

(f) How high an income would you predict for a 100-acre farm without cows?

(a) 285

(b) 213

(c) 499

(d) 325

```
# (c) based on the estimate coefficients, the income is
285.457 + 100*2.138
```

```
## [1] 499.257
```

(g) In a simple linear regression model using the area of a farm as the only explanatory variable, would it (the area) have a significant influence on the income?

(a) Definitely

(b) Definitely not

(c) It isn't clear

```
# (c)
```