# Biostatistics: Exercise 09

Beate Sick, Lisa Herzog

10.11.2020

## Exercise 1

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

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

The variables `height (in cm)` and `weight (in kg)` describe the height in centimeter and the weight in kg for a respective patient. The target variable `catlength` is the optimal length of a catheter that is used for an examination of the patient's heart. The goal is to estimate this quantity from the available dataset.

- Do a simple linear regression for both *catlength* $\sim$ *height* and *catlength* $\sim$ *weight*. Is there a significant influence of the predictors on the target?

```
# simple linear regression models
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
```

```
# Both predictors are highly significant.
```

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

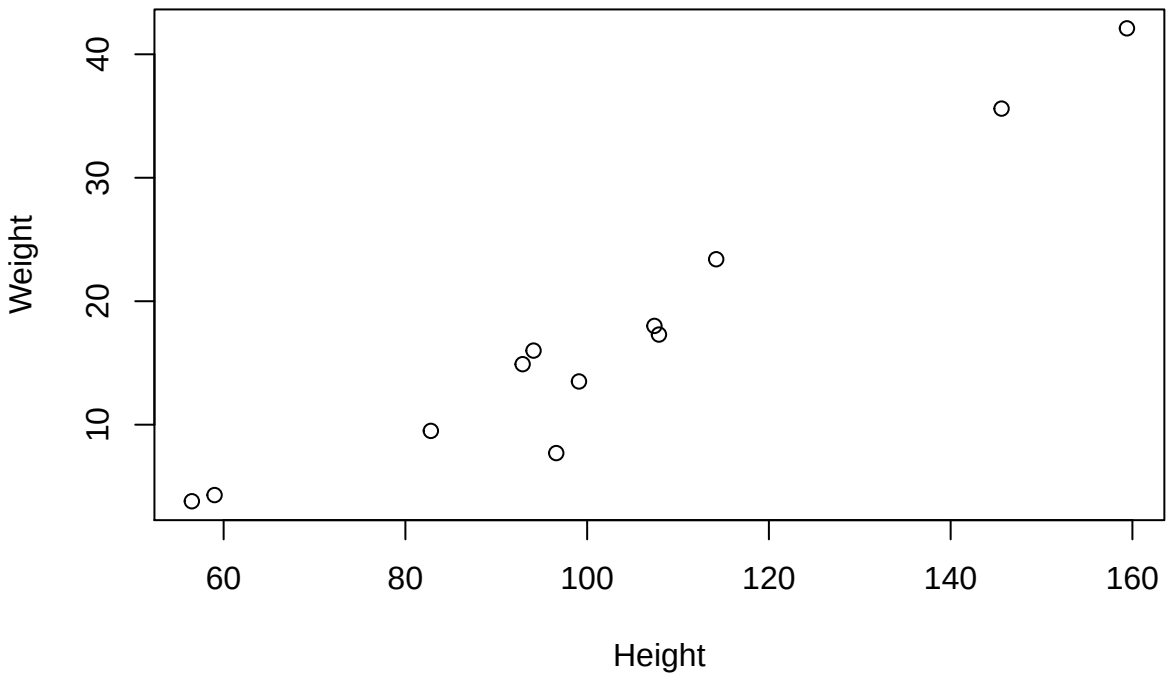
```
# multiple regression
mod   <- lm(catlength ~ height + weight, data = catheter)
summary(mod)
```

```
##
## 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 such that the null hypothesis is rejected at the 1% level.
# We know that at least one of the predictors is necessary.
```

- 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 regression models. Comment and explain the differences if there are any.

```
plot(catheter$height, catheter$weight,
     xlab = "Height", ylab = "Weight")
```




```
# We tested the null hypotheses already in the previous exercise with the
# multiple regression model and both null hypotheses are retained,
# i.e. the predictors are not significantly different from 0.
# Is this a contradiction to the results from the two simple linear regression models?
# 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 (s. plot).
# Due to the smaller p-value we would prefer the predictor weight in this case.
```

- For a child with height 120cm and weight 25kg, compute the 95% prediction interval once with the simple regression models and once with the multiple regression model. In practice, a prediction error of ± 2cm was 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? Why do we use a prediction and not a confidence interval?

```
# 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 accuracy 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 was acceptable.
# Thus, the data and the models do not allow for a prediction of
# catlength that is sufficiently precise.

# The prediction interval gives us the range in which the observation
# will fall. The confidence interval is a measure for the mean.
```

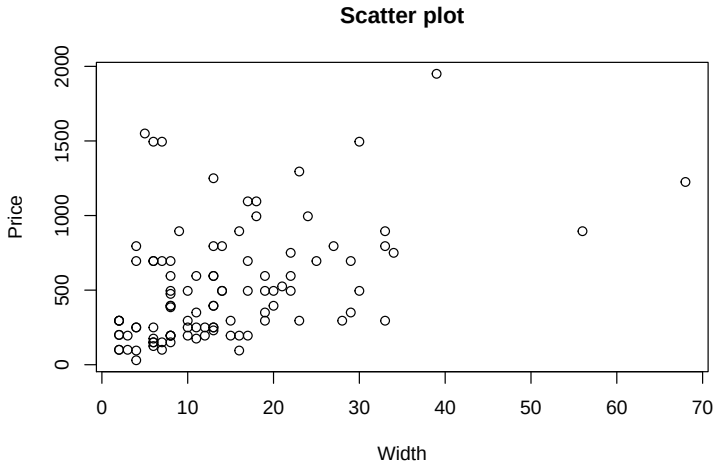
## Exercise 2

The figure below 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 was fitted to the data and we assume that the model assumptions are not violated:

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

Consider the R output and the plot to answer the following questions.

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




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

```
# True, the p-value corresponding to the t-value is
(1 - pt(4.752, 98))*2
```

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

- Which of the following intervals is an exact 95% confidence interval for $\beta$ under the assumption of normally distributed errors?
  - $15.071 \pm 1.984 \cdot 3.171$
  - $15.071 \pm 1.984 \cdot 4.752$
  - $15.071 \pm \frac{1}{\sqrt{100}}\ 1.984 \cdot 3.171$
  - $15.071 \pm \frac{1}{\sqrt{100}}\ 1.984 \cdot 4.752$
  - None of the indicated intervals

```
# The correct interval is the second one with
# 15.071 +/- qt(0.975, df = 98) * 3.171
# and
qt(0.975, df = 98)
```

```
## [1] 1.984467
```

- What does a book of width 30mm on average approximately costs (in pence), based on the regression fit?
  - 500
  - 750
  - 1000
  - 1250
  - 1500

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

```
## [1] 752.615
```

```
# which is on average 750.
```

## Exercise 3

The following dataset summarizes the income (in dollar), the number of cows and the size of the farm (in acres) for 20 American farms.

```
str(farm)
```

```
## 'data.frame':    20 obs. of  3 variables:
##  $ Dollar: int  960 830 1260 610 590 900 820 880 860 760 ...
##  $ cows  : int  18 0 14 6 1 9 6 12 7 2 ...
##  $ acres : int  60 220 180 80 120 100 170 110 160 230 ...
```

We fit the following linear regression model to the dataset:

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

with $E_i \sim N(0, \sigma^2)$ iid..

Answer the following questions with the information above and this 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
```

- The size of a farm has a statistically significant influence on its income given that the number of cows is kept fixed.
  - True
  - False

```
# True: The p-value of the variable acres is clearly below 0.05.
```

- The number of cows on a farm has a statistically significant influence on its income given that the acres are kept fixed.
  - True
  - False

```
# True: The p-value of the variable ``cows'' is clearly below $0.05$.
```

- What is the outcome of the test of the null hypothesis $H_0 : \beta_2 = 0$ against the alternative $H_A : \beta_2 \neq 0$?
  - Keep $H_0$
  - Reject $H_0$

```
# Reject H0: The very small p-value for acres indicates that we have strong
# evidence againt the null.
```

- How many degrees of freedom are there in this model fit?
  - $\infty$
  - 20
  - 18
  - 17
  - 3

```
# 17: from the dataset we know that the model was trained on n = 20
# observations and p = 3 parameters are estimated (intercept and two coefficients).
# Therefore, df = n - p = 17.
```

- Which of the following is an exact 95% confidence interval for $\beta_1$?
  - $32.569 \pm 2.11 \cdot 3.7276$
  - $32.569 \pm 1.96 \cdot 3.7276$
  - $32.569 \pm \frac{1}{\sqrt{17}} 2.11 \cdot 5.45$
  - None of the above

```
# 32.569 +/- qt(0.975, df = 17) * 3.7276:
# the exact CIs of the coefficients of a linear regression model are using
# quantiles from the t-distribution
qt(0.975,df = 17)
```

```
## [1] 2.109816
```

- What is the predicted income for a 100 acre farm without cows?
  - 285
  - 213
  - 499
  - 548

```
# 499: based on the estimate coefficients, the income is
285.457 + 100 * 2.138
```

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