Series 4

1. The file `catheter.rda` is a data set from a medical application. 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) Examine the marginal distributions of the three variables and comment on what you notice. Additionally, examine and comment on the two-dimensional scatter plots.

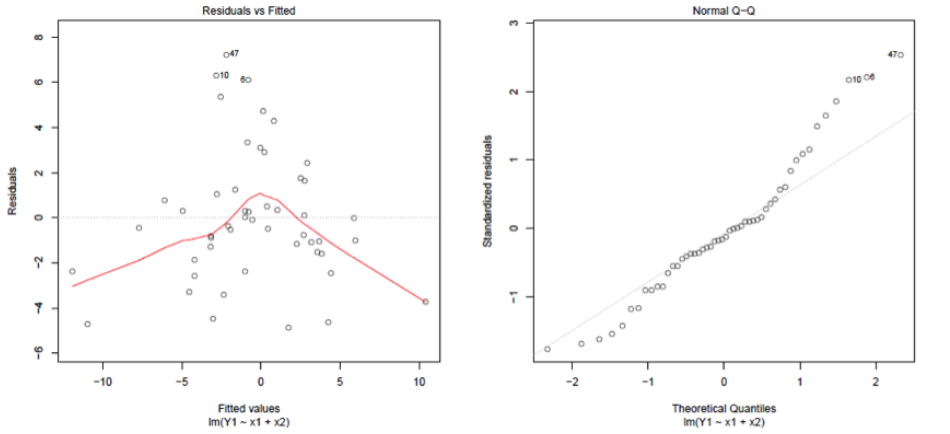
b) Do a simple linear regression for both `catlength` $\sim$ `height` and `catlength` $\sim$ `weight`. Are the predictors significant?

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

d) 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.

e) 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?

2. At a degustation the taste of cake was evaluated. The response y is the score and the predictors are x_1 : *salt content* and x_2 : *sugar content*. The model $\log(y) \sim x_1 + x_2$ was fitted and yielded the following residual plots:



a) The model does not fit well. What are the problems?

b) Now, a larger and better model is fitted by including a quadratic term for the salt content: $\log(y) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \beta_3 \cdot x_2 + E$. Explain why it makes sense to add a quadratic term for the salt content. Why do we expect that $\beta_2 < 0$?

c) The residual plots of this new model look much better indeed (no figure). The summary output looks as follows:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.4150     0.5000  -0.830   0.4109
x1            4.0609     0.4995   8.130 1.91e-10 ***
I(x1^2)      -1.0725     0.4143  -2.589   0.0128 *
x2            2.0109     0.4110   4.893 1.26e-05 ***

Residual standard error: 2.784 on 46 degrees of freedom
Multiple R-squared: 0.7117,Adjusted R-squared: 0.6929
F-statistic: 37.85 on 3 and 46 DF,  p-value: 1.768e-12
```

How many observations are available?

d) Consider a cake with $x_1^* = 3.5$ and $x_2^* = 1$. Compute the conditional median $median(y^*|x_1^*, x_2^*)$ and the conditional expectation $\mathcal{E}(y^*|x_1^*, x_2^*)$ for the cake score expectation on the original scale.

e) Lastly, another, even larger model is fitted by including another predictor: $\log(y) = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \beta_3 \cdot x_2 + \beta_4 \cdot x_3 + E$. We would like to compare the new model with the one from subproblem c) by using a partial F-test. In addition to the summary output from above, we have the following line from the summary output of the new model:
`Residual standard error:  2.7 on 45 degrees of freedom`
Write down the null hypothesis and the alternative, the test statistic and its distribution, the observed value of the test statistic as well as the p-value and the test decision. Use a significance level of 5%.

3. This exercise covers further aspects of the *Conconi* test. As previously mentioned, not only one but several runners were tested. We compared the confidence intervals of the estimated coefficients to test whether the increase in pulse was significantly different between Dani and Marcel. This time we want to assess this question more systematically. The file `conconi2.rda` contains the export of an Excel spread sheet.

a) We first need to preprocess the data. Create a data frame that contains all observations (hint: there are 39) of the variables `puls`, `speed`, and `runner`. `runner` should be a categorical variable with the levels "Dani" and "Marcel", indicating what person the corresponding observation belongs to.

b) Now fit an OLS regression model for the main effects: `puls ~ speed + runner`. What does this model assume with respect to the initial pulse and the increase in pulse of the two runners?

c) Perform a residual analysis by plotting the "residuals vs. fitted" plot and the Normal plot. Which model violations can we detect and what might be their causes?

d) To find the cause of the bad fit we modify the plot "residuals vs. fitted". Use different colors for the points beloning to Dani and Marcel, respectively. What do you observe?

e) Now, fit a model with an interaction term between `speed` and `runner`. What does this model assume with respect to the initial pulse and the increase in pulse of the two runners? Also perform a residual analysis.

f) Compute the estimates for the intial pulse (i.e. when `speed=0`) of Dani and Marcel as well as the estimates for the amount the pulse increases with every additional km/h in speed. Is the difference significant?

4. The Australian Bureau of Agricultural and Resource Economics conducts an annual survey of the agroindustry. In 1991, 451 farms in New South Wales took part. The raw data is contained in the file `farm.rda`. The variables have the following meanings:

`ertrag` : target variable, total revenue of the farm
`aufwand` : predictor, total costs of the farm
`region` : predictor, code for different regions within New South Wales
`industry` : predictor, code for the cultivation (1=Weizen (wheat), 2=Weizen_Schaf_Rind (wheat, sheep, cattle), 3=Schaf (sheep), 4=Rind (cattle), 5=Schaf_Rind (sheep, cattle)).

The aim is to fit a suitable regression model that explains the revenue of a farm. You will need to perform the following steps:

a) Preprocess the data as needed. I.e. define the necessary factor variables, assess whether transformations are necessary, etc.

b) Fit the main effect model and check the residuals. Does the model fit?

c) For a cattle farm in region 111 with costs of 100'000, what is the expected revenue?

d) Test whether `region` has a significant influence on `revenue` when the other predictors are given. Hint: use `drop1(fit, test="F")`.

e) Add an interaction term between `region` and `industry`.

- How many parameters are estimated in total?

- Do we have sufficiently many observations to estimate this model?
- Is the interaction term significant? Do a suitable test.
- What is the intuitive meaning of the interaction term, how do we have to interpret it in the context of the model? What is the conclusion?

f) We now have several models of varying complexity. First, there is the model with all predictors and the interaction term between `region` and `industry`. Then, there is the model with the main effects followed by the one without the variable `region`. Lastly, the least complex model is the one that does not differentiate between `region` and `industry`. Which model is best suited for predicting the revenue of a farm?

Preliminary discussion: Monday, October 26.

Deadline: Monday, November 2.