2. Computer exercises

Demo exercises

2.1 Continuation of the homework.

a) Generate a scatter plot (CONSUMPTION, ILL). Add the estimated regression line to the figure.

b) Determine the fitted values $\hat{y}$ and estimated residuals e from the corresponding model and assign them to variables FIT and RES, respectively.

c) Generate scatter plots (ILL, FIT) and (FIT, RES).

d) Study whether the observation 7=USA is an outlier by using the plots of part (c).

e) Study whether the observation 7=USA is an outlier by using Cook's distances.

f) Estimate the model without the observation USA. Compare the results with the homework assignment of the previous week.

Solution.

```
smoking <- read.table("tobacco.txt",header=T,sep="\t")
model <- lm(ILL~CONSUMPTION,data=smoking)
countries <- c("Iceland","Norway","Sweden","Canada","Denmark",
          "Austria","USA","Netherlands","Switzerland","Finland",
          "England")
```

a) Scatter plot (Figure 1):

```
plot(smoking$CONSUMPTION,smoking$ILL, ylab="Cases in 1950",
     xlab="CONSUMPTION in 1930", pch=16,
     main="CONSUMPTION/ILL per 100 000 individuals")
abline(model,col="red")
text(smoking$CONSUMPTION, smoking$ILL, labels=countries, cex= 0.8,
     pos=3)
```

Alternatively, you can use the function `identify` to label the observations.

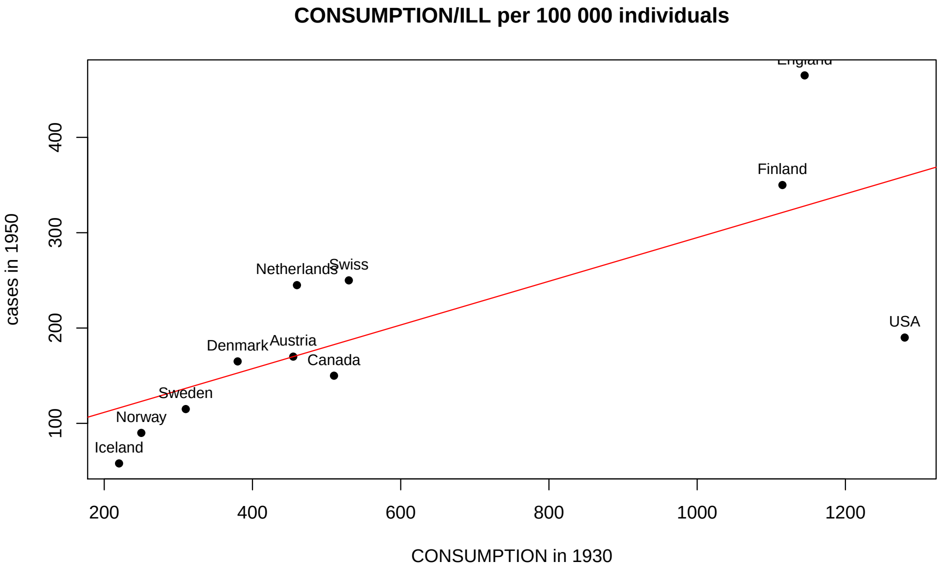

Figure 1: Scatter plot of CONSUMPTION and ILL.

b) The fitted values and the estimated residuals correspond to `fitted.values` and `residuals` from the estimated model, and they can be accessed by

```
FIT <- model$fit
RES <- model$res
```

c) Scatter plot (observed values, fitted values) (Figure 2).
Plot the fitted values $\hat{y}_i$ against the observed values of the variable ILL.

```
plot(smoking$ILL,FIT, ylab="Fits",xlab="Sick",pch=16)
text(smoking$ILL,FIT,  labels = ifelse(rownames(smoking)=="7",
     countries, NA),pos=2)
```

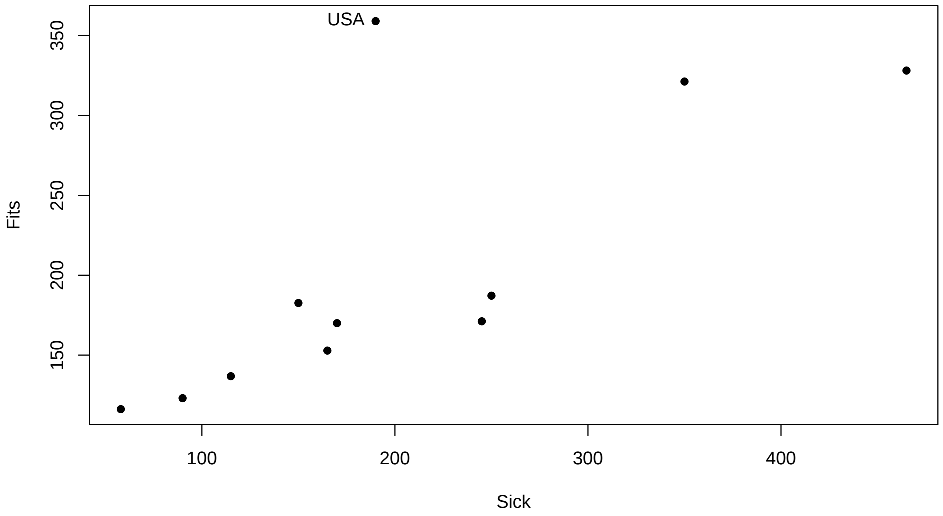

Figure 2: Scatter plot of the observed values and the fitted values.

The scatter plot illustrates the goodness of the model:

- The closer the points $(y_i, \hat{y}_i), i = 1, 2, ..., n$ are to the line with slope of 1, the better the model is.
- Outliers are usually visible.

Note that, the squared Pearson correlation coefficient given by the points $(y_i, \hat{y}_i), i = 1, 2, ..., n$ is equal to the coefficient of determination:

$$[\mathrm{Cor}(y, \hat{y})]^2 = R^2.$$

Scatter plot (fitted values, residuals) (Figure 3). Plot the residuals e_i against the fitted values $\hat{y}_i$.

```
plot(FIT,RES, xlab="Fits",ylab="Residuals",pch=16)
text(FIT,RES,  labels = ifelse(rownames(smoking)=="7",
     countries, NA),pos=3)
```

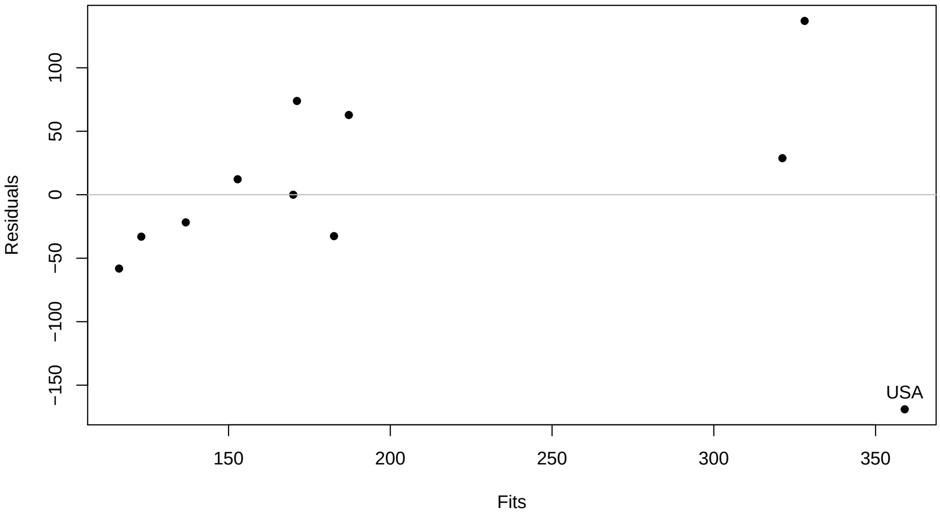

Figure 3: Scatter plot of the fitted values and the residuals.

The scatter plot illustrates the goodness of the model:

- The closer the points $(\hat{y}_i, e_i), i = 1, 2, ..., n$ are to the line $e = 0$, the better the model is.
- Outliers are usually visible.

d) Especially, by the scatter plot (FIT,RES), the observation 7=USA looks like an outlier.

e) Assign the Cook's distances to `cooksd` and plot the distances. See Figure 4.

```
cooksd <- cooks.distance(model)
x <-plot(cooksd,xaxt="n",xlab=" ",ylab="Cook’s distances")
axis(side=1,at=1:11, labels=countries,las=2 )
```

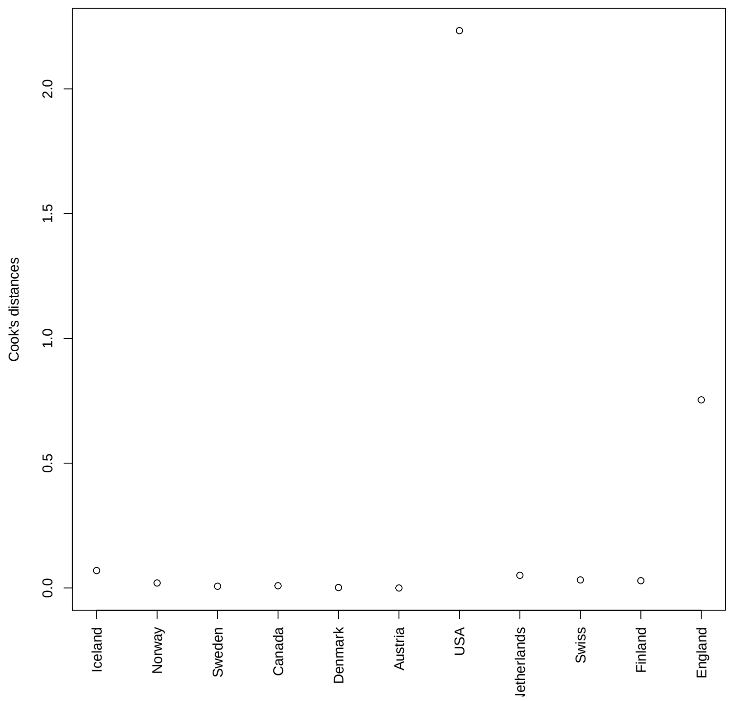

Figure 4: Cook's distances of the model.

f) Estimate the model without the observation 7=USA.

```
smoking2 <- smoking[-7,]
model2 <- lm(ILL~CONSUMPTION,data=smoking2)
summary(model2)

Residuals:
    Min      1Q  Median      3Q     Max
-62.353 -28.923  -7.861  35.321  66.919

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.55343   28.26713   0.479    0.644
CONSUMPTION[-7]  0.35767    0.04547   7.867 4.93e-05 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 44.92 on 8 degrees of freedom
Multiple R-squared:  0.8855,Adjusted R-squared:  0.8712
F-statistic: 61.88 on 1 and 8 DF,  p-value: 4.928e-05
```

Compared to the first homework assignment, the estimate for the slope has increased from 0.23 to 0.36. This implies a stronger linear dependence between lung cancer cases and consumption of cigarettes among the remaining observations (countries).

Question: Can we remove the observation 7=USA?

Answer: During the corresponding time period, tobacco was milder in the USA, when compared to the other countries of the study. Furthermore, the cigarettes sold in the USA had filters, whereas the cigarettes sold in the other countries did not have filters.

As we have found a contextual explanation, the observation USA can be regarded as an outlier and its removal from the data is justified. Remember that disregarding data without valid explanations is not allowed!

2.2 When cement hardens, heat is produced. The amount of heat depends on the composition of the cement. From file `hald.txt`, you can find the following information regarding 13 different batches of cement:

HEAT	=heat energy (cal/g)
CHEM1, CHEM2, CHEM3, CHEM4	=ingredients of cement (% of the dry substance)

a) Estimate a linear regression model with all explanatory variables. Compare statistical significances of the regression coefficients and examine the variance inflation factors of the corresponding explanatory variables.

b) Find the best combination of explanatory variables by using Akaike information criterion (AIC).

Solution. The goal of the exercise is to find out which of the explanatory variables CHEM1, CHEM2, CHEM3, CHEM4 are significant in explaining the behavior of the response variable HEAT.

First, we import the data and install the package `car` for later use.

```
install.packages("car")
library(car)
hald=read.table("hald.txt",header=T)
```

a) **Estimation of the full model**

In situations, where it is not known which of the explanatory variables affect the response variable, it is first usually reasonable to estimate the full model, i.e. the model with all candidates for explanatory variables.

First, we should examine the correlations between the different variables.

```
cor(hald)
            CHEM1      CHEM2      CHEM3      CHEM4       HEAT
CHEM1  1.00000000  0.2285795 -0.8241338 -0.2454451  0.7307175
CHEM2  0.22857947  1.0000000 -0.1392424 -0.9729550  0.8162526
CHEM3 -0.82413376 -0.1392424  1.0000000  0.0295370 -0.5346707
CHEM4 -0.24544511 -0.9729550  0.0295370  1.0000000 -0.8213050
HEAT   0.73071747  0.8162526 -0.5346707 -0.8213050  1.0000000
SUM    0.05010722 -0.2604492 -0.1102512  0.3290769 -0.1645805
SUM
0.05010722
-0.26044918
-0.11025122
0.32907694
-0.16458053
1.00000000
```

The variable HEAT correlates strongly with all explanatory candidates. Correlation is positive with the variables CHEM1 and CHEM2, and negative with CHEM3 and CHEM4. There is a strong negative correlation between variables CHEM1 and CHEM3, as well as between variables CHEM2 and CHEM4.

We begin by estimating the full model:

$$\text{HEAT} = \beta_0 + \beta_1\text{CHEM1} + \beta_2\text{CHEM2} + \beta_3\text{CHEM3} + \beta_4\text{CHEM4} + \epsilon \qquad (1)$$

```
fullmodel=lm(HEAT~CHEM1+CHEM2+CHEM3+CHEM4,data=hald)
summary(fullmodel)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1750 -1.6709  0.2508  1.3783  3.9254

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  62.4054    70.0710   0.891   0.3991
CHEM1         1.5511     0.7448   2.083   0.0708 .
CHEM2         0.5102     0.7238   0.705   0.5009
CHEM3         0.1019     0.7547   0.135   0.8959
CHEM4        -0.1441     0.7091  -0.203   0.8441
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 2.446 on 8 degrees of freedom
Multiple R-squared:  0.9824,Adjusted R-squared:  0.9736
F-statistic: 111.5 on 4 and 8 DF,  p-value: 4.756e-07
```

The model (1) has a high coefficient of determination (98.2%). The value of the F-test statistics for the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

is 111.5 and the p-value is close to zero, i.e. the model is statistically significant and at least one of the regression coefficients $\beta_0, \beta_1, \beta_2, \beta_3$ deviates from zero.

However, none of the explanatory variables of the model (1) is statistically significant with a 5%:n level of significance. This is due to the multicollinearity of the explanatory variables.

Multicollinearity of the explanatory variables can be measured with VIF-coefficients. The VIF-coefficient is 1 for an explanatory variable whose sample correlation is 0 with other explanatory variables. The stronger a variable is linearly dependent on the other variables, the larger the VIF-coefficient of the variable is. If

$$\text{VIF} > 10,$$

then multicollinearity might be a problem.

VIF-coefficients can be computed with the function `vif` of the package `car`.

```
vif(fullmodel)
    CHEM1     CHEM2     CHEM3     CHEM4
 38.49621 254.42317  46.86839 282.51286
```

In model (1), the VIF-coefficients of the variables CHEM2 and CHEM4 are larger than 200, which indicates that strong multicollinearity is present in the model.

Next, we further study the existing multicollinearity by estimating two regression models, where CHEM2 and CHEM4 are explained with all the other explanatory variables of the original model (1).

Consider the model:

$$\text{CHEM2} = \alpha_0 + \alpha_1\text{CHEM1} + \alpha_3\text{CHEM3} + \alpha_4\text{CHEM4} + \delta, \tag{2}$$

which can be estimated using,

```
model2 <- lm(CHEM2 ~ CHEM1+CHEM3+CHEM4,data=hald)
summary(model2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.2494 -0.7280  0.3881  0.7033  0.9512

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 96.59382    2.16253   44.67 7.06e-12 ***
CHEM1       -0.97860    0.10602   -9.23 6.94e-06 ***
```

```
CHEM3       -1.00350    0.09443  -10.63 2.15e-06 ***
CHEM4       -0.97759    0.02111  -46.30 5.12e-12 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.126 on 9 degrees of freedom
Multiple R-squared:  0.9961,Adjusted R-squared:  0.9948
F-statistic: 760.3 on 3 and 9 DF,  p-value: 3.864e-11
```

The coefficient of determination of the model is 99.6% implying that CHEM2 is strongly linearly dependent on the other explanatory variables. Note that the VIF-coefficient of CHEM2 in the model (1) is

$$\mathrm{VIF}_2 = \frac{1}{1 - R_2^2},$$

where R_2^2 is the coefficient of determination for model (2).
Consider the model,

$$\mathrm{CHEM4} = \alpha_0 + \alpha_1\mathrm{CHEM1} + \alpha_2\mathrm{CHEM2} + \alpha_3\mathrm{CHEM3} + \delta, \tag{3}$$

which can be estimated using,

```
Call:
lm(formula = CHEM4 ~ CHEM1 + CHEM2 + CHEM3)

Residuals:
    Min      1Q  Median      3Q     Max
-2.3264 -0.6836  0.4439  0.7463  1.0379

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 98.65079    1.94627  50.687 2.27e-12 ***
CHEM1       -1.00504    0.10175  -9.878 3.96e-06 ***
CHEM2       -1.01865    0.02200 -46.303 5.12e-12 ***
CHEM3       -1.02809    0.09187 -11.191 1.39e-06 ***
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Residual standard error: 1.15 on 9 degrees of freedom
Multiple R-squared:  0.9965,Adjusted R-squared:  0.9953
F-statistic: 844.5 on 3 and 9 DF,  p-value: 2.413e-11
```

The coefficient of determination of the model is 99.7% implying that CHEM4 is strongly linearly dependent on the other explanatory variables.

Note that the VIF-coefficient of CHEM4 in the model (1) is

$$\mathrm{VIF}_4 = \frac{1}{1 - R_3^2},$$

where R_3^2 is the coefficient of determination of the model (3).

Multicollinearity of the model (1) is explained by noting that cement consists almost entirely of the substances CHEM1, CHEM2, CHEM3 and CHEM4. The sum of these variables is somewhere between 95-99%. Therefore, by increasing the amount of a substance, we have to reduce the amount of some other substances in the mixture. This explains the strong negative correlations between the variable pairs (CHEM1, CHEM3) and (CHEM2, CHEM4).

b) **The best combination of explanatory variables**

There exists different strategies for choosing the explanatory variables of a regression model. When searching for the best combination of explanatory variables, different models are compared to each other by using some criterion for model selection.

Some well-known criteria for model selection are, e.g., Akaike information criterion (AIC), Schwarz bayesian information criterion (SBIC) and Hannan-Quinn criterion (HQ).

The criterion functions of model selection methods are of the form,

$$\min_{M \subseteq (1,\ldots,q)} C(|M|, \hat{\sigma}_M^2),$$

where M is a combination of explanatory variables and $\hat{\sigma}_{|M|}^2$ is the maximum likelihood estimator for the variance of the residuals of the corresponding model. Furthermore, C is an increasing function with respect to the two arguments. In general, we expect the following from a criterion function:

- Maximal coefficient of determination,
- Using as few explanatory variables as possible.

In R, the function `step()` gives the combination of explanatory variables that minimizes the value of AIC. Note that `step()` computes AIC by assuming normally distributed residuals.

```
step(fullmodel)

Start:  AIC=26.94
HEAT ~ CHEM1 + CHEM2 + CHEM3 + CHEM4

        Df Sum of Sq    RSS    AIC
- CHEM3  1    0.1091 47.973 24.974
- CHEM4  1    0.2470 48.111 25.011
- CHEM2  1    2.9725 50.836 25.728
<none>               47.864 26.944
```

```
- CHEM1  1   25.9509 73.815 30.576

Step:  AIC=24.97
HEAT ~ CHEM1 + CHEM2 + CHEM4

        Df Sum of Sq    RSS    AIC
<none>                47.97 24.974
- CHEM4  1      9.93  57.90 25.420
- CHEM2  1     26.79  74.76 28.742
- CHEM1  1    820.91 868.88 60.629

Call:
lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM4, data = hald)

Coefficients:
(Intercept)        CHEM1        CHEM2        CHEM4
    71.6483       1.4519       0.4161      -0.2365
```

The output can be interpreted as follows. The AIC of the full model is 26.944. When CHEM3 is omitted from the model, the AIC is 24.974. When CHEM4 is omitted, the AIC is 25.011. When CHEM2 is omitted, the AIC is 25.728 and when CHEM1 is omitted, the AIC is 30.576. We wish to minimize the model selection criterion and hence, we estimate the model without CHEM3.

Consider the model,

$$\text{HEAT} = \beta_0 + \beta_1\text{CHEM1} + \beta_2\text{CHEM2} + \beta_4\text{CHEM4}. \tag{4}$$

Now the AIC of model (4) is 24.974. From the output of R, we see that omitting any of the remaining explanatory variables (CHEM1, CHEM2, CHEM4) would increase the AIC value. Next, we estimate the model (4).

```
model4 <- lm(HEAT ~ CHEM1  + CHEM2 + CHEM4 , data=hald)
summary(model4)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0919 -1.8016  0.2562  1.2818  3.8982

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  71.6483    14.1424   5.066 0.000675 ***
CHEM1         1.4519     0.1170  12.410 5.78e-07 ***
CHEM2         0.4161     0.1856   2.242 0.051687 .
CHEM4        -0.2365     0.1733  -1.365 0.205395
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.309 on 9 degrees of freedom
Multiple R-squared:  0.9823,Adjusted R-squared:  0.9764
F-statistic: 166.8 on 3 and 9 DF,  p-value: 3.323e-08
```

Note that the variables CHEM2 and CHEM4 are not statistically significant with 5% significance level. Figure 5 illustrates the estimated residuals of the full model. The shape of the histogram indicates that the normality assumption does not hold, which on the other hand means that AIC is not a reliable method for model selection. In homework assignment 2.3, the model selection is done using the permutation test. The permutation test does not require normality and thus, it is the safer alternative here.

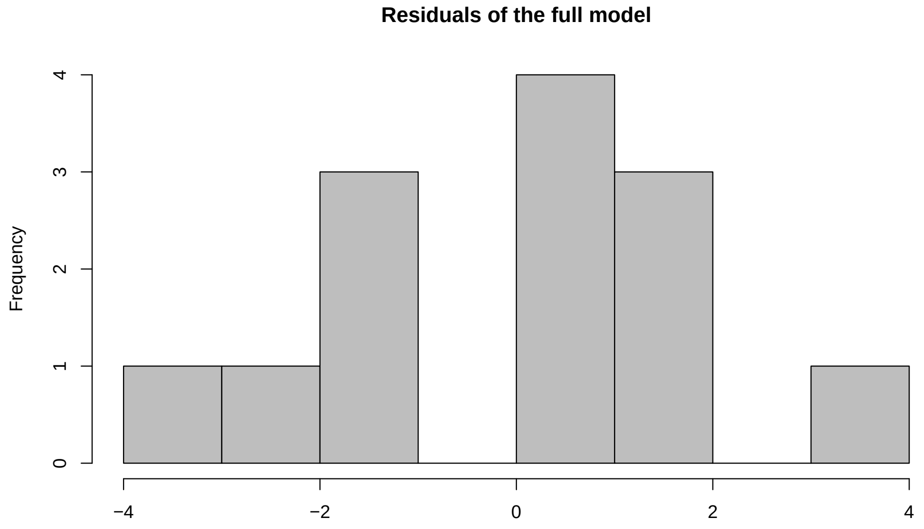

Figure 5: The residuals of the full model.

Remark: It is not possible to use the error sum of squares or the coefficient of determination as a criterion for model selection, since minimizing the error sum of squares as well as maximizing the coefficient of determination always leads to the full model (the model with all possible explanatory variables).

Homework

2.3 Continuation to Exercise 2.2. Use backward elimination to choose the model. Perform the backward elimination using the permutation test. You may utilize lecture slides and demo exercises of the previous week. Compare results with part (b) of Problem 2.2. Use level of significance $\alpha = 5\%$.

In backward elimination, the first step is to estimate the full model and examine statistical significance of the explanatory variables. The least significant variable is removed from the model and after that, a new model is estimated. Variables are removed from the model one at a time, until all remaining variables are statistically significant.

2.4 The quantity of a fertilizer affects the yield of wheat. The effect was studied by altering the quantity of the fertilizer (11 levels) in 33 different cultivations (the same amount of fertilizer in 3 cultivations) and by measuring the yield of each cultivation. Results of the study are given in the file `crop.txt`. The variables are,

Yield	= Yield (kg/unit of area)
Fertilizer	= the amount of the fertilizer (kg/unit of area)

a) Estimate a linear regression model, where Yield is a response variable and Fertilizer is an explanatory variable. Using regression graphics, study whether the model is sufficient.

b) Estimate a linear regression model, where you have added the explanatory variable

$$\text{LSqrd} = \text{Fertilizer} \cdot \text{Fertilizer}$$

to the model of the part a). That is, LSqrd consists of the squared elements of the variable Fertilizer. Using regression graphics, study whether the model is sufficient.

c) Compare the results obtained in parts a) and b). Which of the models is more suitable here?