Computer exercises 2

R packages

If there is some functionality that is not implemented in base R there is most probably a package for it. You can install packages with the function `install.packages`. Note that the package name has to be given as a character string for the function `install.packages`. For example, the package `car` that is required for this exercise session can be installed with the following line of code.

```
install.packages("car")
```

Once the package is installed you can use functionality inside the package by specifying the correct namespace and using double colon `::` between the namespace and the function. Below we use the function `vif` from the package `car`.

```
data <- read.table("data/hald.txt", header = TRUE, sep = "\t")
car::vif(lm(HEAT ~ . - SUM, data = data))
```

```
##     CHEM1     CHEM2     CHEM3     CHEM4
##  38.49621 254.42317  46.86839 282.51286
```

Namespaces are a useful concept since there can be functions with the same name in different packages. For example function `lag` can be found at least in two different packages.

```
?stats::lag
?dplyr::lag
```

Instead of specifying the namespace one can attach the package with the function `library`. For an example, see following lines of code.

```
library(car)
data <- read.table("data/hald.txt", header = TRUE, sep = "\t")
# No reference to namespace
vif(lm(HEAT ~ . - SUM, data = data))
```

```
##     CHEM1     CHEM2     CHEM3     CHEM4
##  38.49621 254.42317  46.86839 282.51286
```

If you decide to attach packages it is a good practice to put `library` commands at the top of your script, instead of scattering them all around.

Some useful packages

A bundle of packages called tidyverse provides many useful packages for data science. For example, tidyverse includes below packages among others.

- `ggplot2` – Produce quality figures.
- `purrr` – Replaces R base solutions for functional programming.
- `tibble` – Enhances the common data type `data.frame`.

Homepage of tidyverse: https://www.tidyverse.org/

Demo exercises

2.1

This exercise is 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 `7 = USA`. Compare the results with the homework assignment of the previous week.

Solution

First we read the data. We can remove redundant variables.

```
smoking <- read.table("data/tobacco.txt", header = TRUE, sep = "\t",
                      row.names = "COUNTRY")
smoking <- smoking[, c("CONSUMPTION", "ILL")]
str(smoking)
```

```
## 'data.frame':    11 obs. of  2 variables:
##  $ CONSUMPTION: int  220 250 310 510 380 455 1280 460 530 1115 ...
##  $ ILL        : int  58 90 115 150 165 170 190 245 250 350 ...
```

Then estimate the linear regression model.

```
model <- lm(ILL ~ CONSUMPTION, data = smoking)
```

a)

Scatter plot with regression line is showed in Figure 1.

```
# Country labels
countries <- c("Iceland", "Norway", "Sweden", "Canada", "Denmark", "Austria",
               "USA", "Netherlands", "Switzerland", "Finland", "England")

# Plotting
plot(smoking$CONSUMPTION, smoking$ILL,
     ylab = "Cases in 1950",
     xlab = "CONSUMPTION in 1930",
     main = "CONSUMPTION/ILL per 100 000 individuals",
     pch = 16, cex = 1.5, col = "midnightblue",
     xlim = c(0, max(smoking$CONSUMPTION)),
     ylim = c(min(smoking$ILL), max(smoking$ILL) + 50))
abline(model, lty = 2, lwd = 2)

# Add labels for points
text(smoking$CONSUMPTION, smoking$ILL, labels = countries, cex = 0.5,
     pos = 3)
```

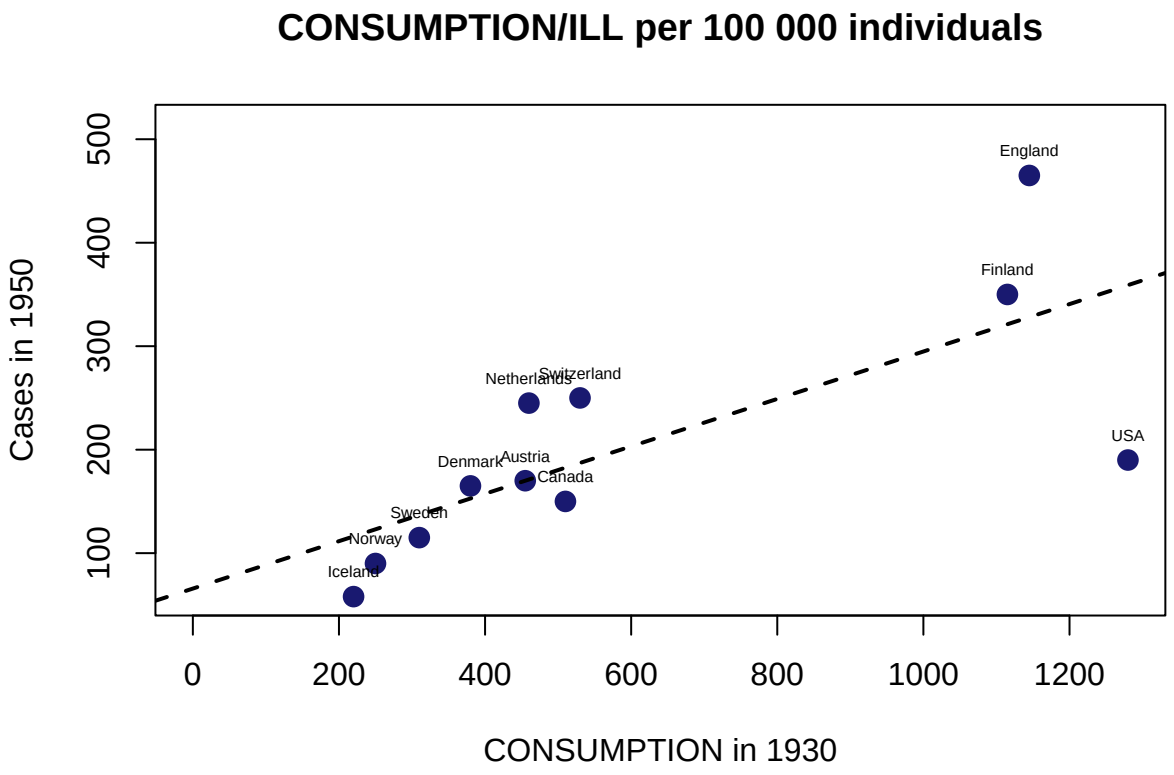

Figure 1: Estimated regression line and scatter plot of the variables.

b)

Remember that `model` is a S3 object of class `lm` (essentially, a named list). There are multiple object oriented systems in R, S3 system being one of them.

```
names(model)
```

```
##  [1] "coefficients"  "residuals"     "effects"       "rank"
##  [5] "fitted.values" "assign"        "qr"            "df.residual"
##  [9] "xlevels"       "call"          "terms"         "model"
```

```
fit <- model$fitted.values
res <- model$residuals
```

c)

Figure 2 shows scatter plot where fitted values $\widehat{y}_i$ are plotted against observed values $\texttt{ILL}_i$.

```
plot(smoking$ILL, fit, pch = 21, bg = "skyblue", cex = 1.5, ylab = "Fits",
     xlab = "Sick")
abline(a = 0, b = 1, col = "grey")

# label USA with ifelse function
text(smoking$ILL, fit, labels = ifelse(countries == "USA", "USA", NA), pos = 1)
```

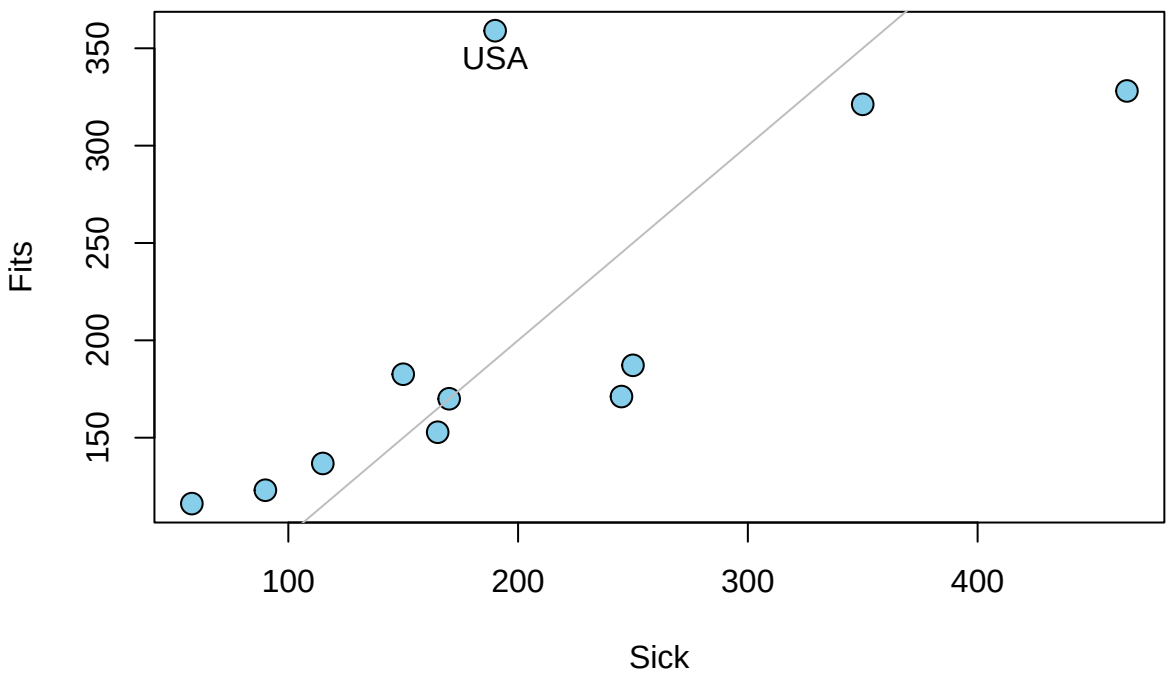

Figure 2: Scatter plot (`ILL`, `fit`).

Coefficient of determination R^2 is equal to the squared Pearson correlation coefficient between observed values $\texttt{ILL}_i$ and fitted values $\hat{y}_i$.

```
cor(smoking$ILL, fit)^2
```

```
## [1] 0.54904
```

```
summary(model)$r.squared
```

```
## [1] 0.54904
```

The scatter plot in Figure 2 illustrates the goodness of the fit:

- The closer the points $(\texttt{ILL}_i, \hat{y}_i), i = 1, 2, \ldots, n$ are to the line $f(x) = x$, the better the model is.
- Outliers are usually visible (see `USA`).
- Nonlinear shapes indicate that the functional form of the model part is not well selected.

Figure 3 shows scatter plot where fitted values $\hat{y}_i$ are plotted against residuals e_i.

```
plot(fit, res, pch = 21, bg = "skyblue", cex = 1.5,
     xlab = "Fits", ylab = "Residuals")
abline(h = 0, col = "grey")

# Another way to label USA
text(fit[7], res[7], labels = "USA", pos = 3)
```

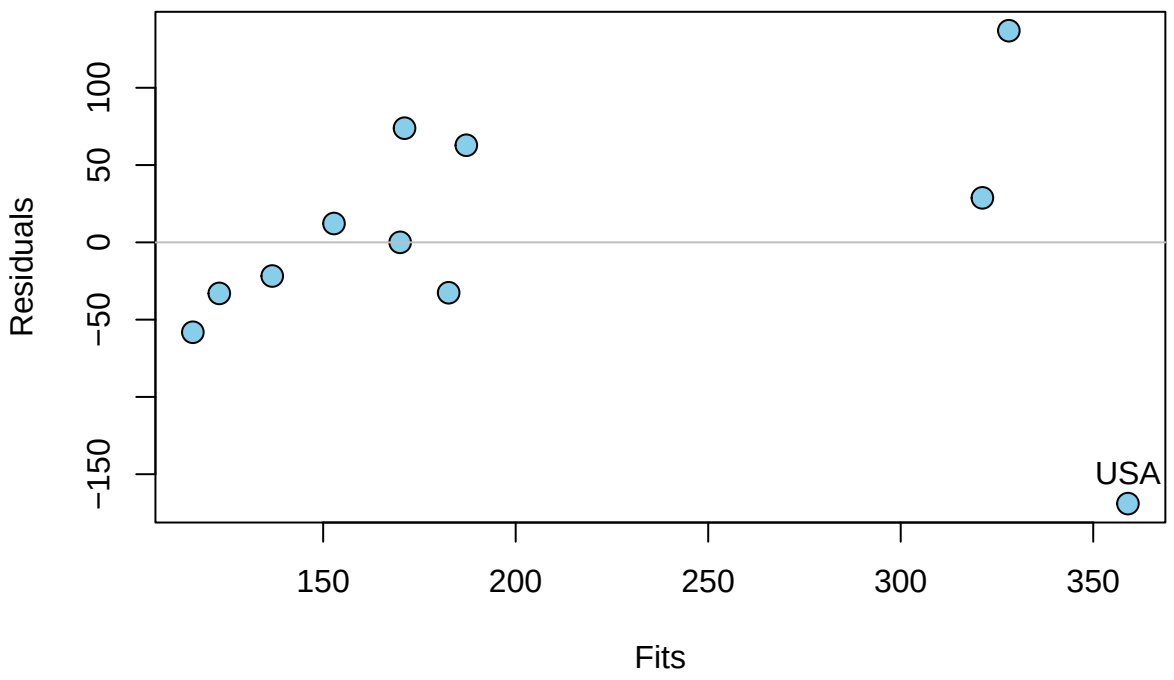

Figure 3: Scatter plot (`fit`, `res`).

The scatter plot in Figure 3 illustrates the goodness of the model:

- The closer the points $(\hat{y}_i, e_i), i = 1, 2, \ldots, n$ are to the line $f(x) = 0$, the better the model is.
- Outliers are usually visible (see `USA`).
- Nonlinear shapes indicate that the functional form of the model part is not well selected.
- If the height of the scatter plot is not approximately the same everywhere, the residuals might be heteroscedastic.

d)

Figures 2 and 3 suggest that the observation `USA` is an outlier.

e)

Figure 4 shows Cook's distance for each observation. Also, according to Cook's distances observation `USA` is an outlier.

```
cooksd <- cooks.distance(model)

# xaxt = "n" => No ticks at x-axis
# type = "h" => Plot vertical lines instead of points
plot(cooksd, xaxt = "n", type = "h", lwd = 3, xlab = NA,
     ylab = "Cook's distances")
```

```
# side = 1 => Modify x-axis (bottom axis)
# at = 1:11 => Add tick marks to places 1:11
# las = 2 => Rotate text 90 degrees
# cex.axis = 0.9 => Smaller x-axis labels
axis(side = 1, at = 1:11, labels = countries, las = 2, cex.axis = 0.9)
```

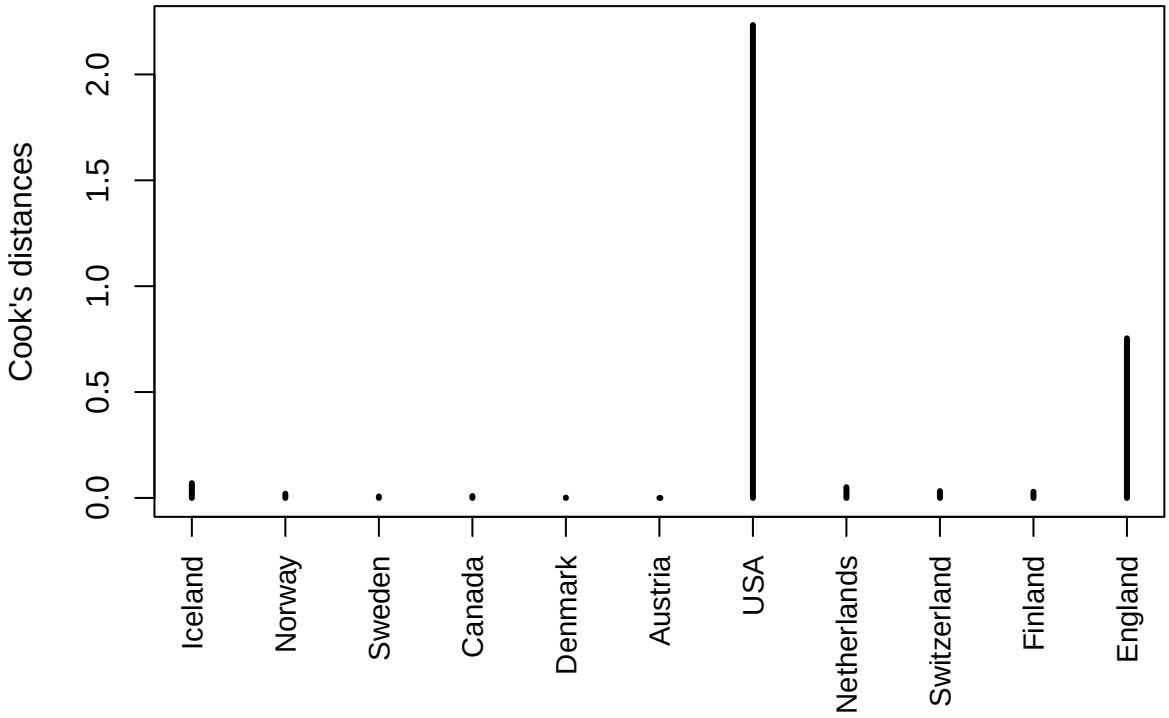

Figure 4: Cook's distance for each country.

f)

Contextual explanation for the outlyingness of the observation `USA` is that during the corresponding time period, tobacco was milder in the USA, when compared to the other countries of the study. Furthermore, cigarettes sold in the USA had filters, whereas the cigarettes sold in the other countries did not have filters.

For this reason it is appropriate to analyze observation `USA` separately. In our case, this means that we estimate linear regression models with and without the observation `USA`. However, there exists more sophisticated ways to handle outliers, for example, models that are separated in two parts and robust estimation methods. **Remember that one may not just remove data points that are unpleasant.**

Next, let us estimate linear regression model without the observation `USA`.

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

summary(model)
```

```
##
```

```
## Call:
## lm(formula = ILL ~ CONSUMPTION, data = smoking)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -169.016  -32.813    0.004   45.804  136.914
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 65.74886   48.95871   1.343  0.21217
## CONSUMPTION  0.22912    0.06921   3.310  0.00908 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 84.13 on 9 degrees of freedom
## Multiple R-squared:  0.549,  Adjusted R-squared:  0.4989
## F-statistic: 10.96 on 1 and 9 DF,  p-value: 0.009081
```

```
summary(model2)
```

```
##
## Call:
## lm(formula = ILL ~ CONSUMPTION, data = smoking2)
##
## 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  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
```

We can compare above summaries (`summary(model)` and `summary(model2)`), however, it is easier to see the differences of estimated models by plotting them. Figure 5 shows estimated models with and without the observation `USA`.

```
plot(smoking$CONSUMPTION, smoking$ILL,
     ylab = "Cases in 1950",
     xlab = "CONSUMPTION in 1930",
     main = "CONSUMPTION/ILL per 100 000 individuals",
     pch = 16,
     cex = 1.5,
     col = ifelse(countries != "USA", "blue", "red"),
     xlim = c(0, max(smoking$CONSUMPTION)))
abline(model2, lty = 1, lwd = 2)
abline(model, lty = 2, lwd = 2)
legend("topleft", legend = c("No USA", "With USA"), lty = 1:2)
```

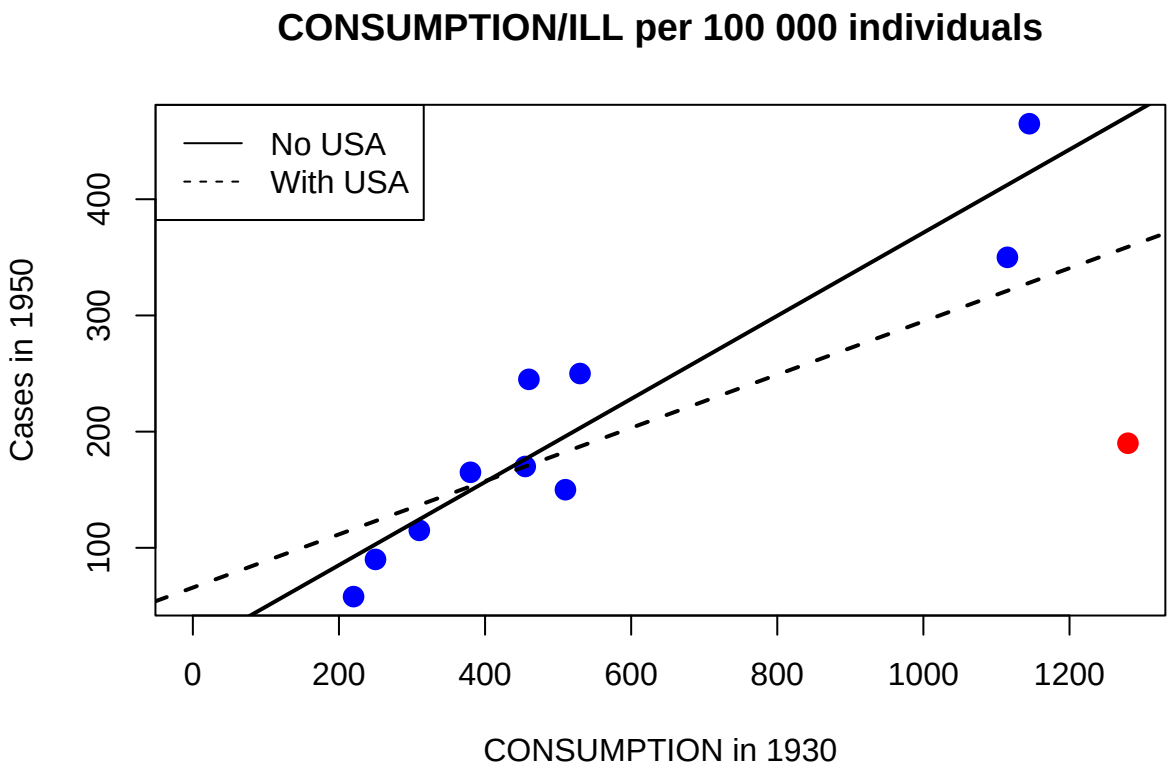

Figure 5: Estimated linear regression models with and without the observation `USA`. Observation `USA` is colored as red.

Figure 5 shows that 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).

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),
`CHEMi` = ingredients of cement (% of the dry substance), $\texttt{i} \in \{1, 2, 3, 4\}$.

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

First we read the data.

```
hald <- read.table("data/hald.txt", header = TRUE, sep = "\t")
str(hald)
```

```
## 'data.frame':    13 obs. of  6 variables:
##  $ CHEM1: int  7 1 11 11 7 11 3 1 2 21 ...
##  $ CHEM2: int  26 29 56 31 52 55 71 31 54 47 ...
##  $ CHEM3: int  6 15 8 8 6 9 17 22 18 4 ...
##  $ CHEM4: int  60 52 20 47 33 22 6 44 22 26 ...
##  $ HEAT : num  78.5 74.3 104.3 87.6 95.9 ...
##  $ SUM  : int  99 97 95 97 98 97 97 98 96 98 ...
```

a)

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 examine the correlations between the different variables.

```
cor(hald)
```

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

$$\texttt{HEAT}_j = \beta_0 + \beta_1 \texttt{CHEM1}_j + \beta_2 \texttt{CHEM2}_j + \beta_3 \texttt{CHEM3}_j + \beta_4 \texttt{CHEM4}_j + \varepsilon_j, \quad j \in \{1, 2, \dots, 13\}. \tag{1}$$

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

```
##
## Call:
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM3 + CHEM4, data = hald)
##
## 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 (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 β_0, β_1, β_2, β_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.

Note that F-test and t-test are reliable only when residuals ε_i are normally distributed. By Figure 6 it seems plausible that residual might not be normally distributed. Thus one should not make too definitive conclusions based on t-test and F-test.

```
b <- seq(-4, 4, length.out = 9)
hist(fullmodel$residuals, breaks = b, border = FALSE, col = "skyblue")
```

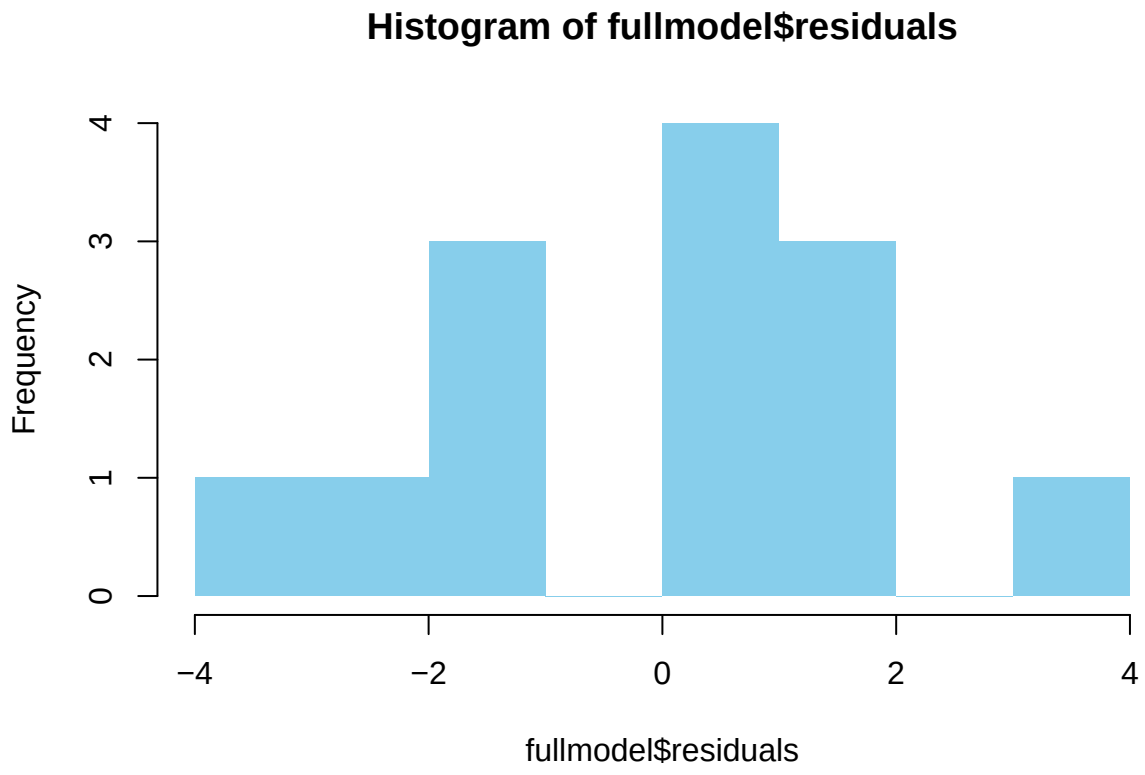

Figure 6: Histogram of estimated residuals of the `fullmodel`.

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.

Consider VIF-coefficient for the explatonary variable `CHEM1`,

$$\text{VIF}_1 = \frac{1}{1 - \text{R}_1^2},$$

where R_1^2 is the coefficient of determination of the model

$$\texttt{CHEM1}_j = \alpha_0 + \alpha_2 \texttt{CHEM2}_j + \alpha_3 \texttt{CHEM3}_j + \alpha_4 \texttt{CHEM4}_j + \delta_j, \quad j \in \{1, 2, \ldots, 13\}.$$

Thus VIF_1 can be calculated manually in the following way.

```
model1 <- lm(CHEM1 ~ CHEM2 + CHEM3 + CHEM4, data = hald)
r1 <- summary(model1)$r.squared
vif1 <- 1 / (1 - r1)
vif1
```

```
## [1] 38.49621
```

We could calculate VIF-coefficiets for other explatonary variables `CHEM2`, `CHEM3` and `CHEM4` in a similar fashion. However, all the VIF-coefficients can be computed with the function `vif` from the package `car`. Install the package `car` at this point if you have not yet, and attach it if you wish to.

```
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.

Multicollinearity of the model (1) is explained by noting that cement consists almost entirely of the substances `CHEM1`, `CHEM2`, `CHEM3` and `CHEM4`, as can be seen by inspecting variable `SUM`.

```
hald$SUM
```

```
##  [1] 99 97 95 97 98 97 97 98 96 98 98 98 98
```

The sum of variables `CHEMi` is somewhere between 95-99% for each batch of cement. 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)

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).

Model selection criteria are often based on minimizing/maximizing a function that is of the form,

$$f(p, \hat{\sigma}_p^2),$$

where p is the number of estimated parameters and $\hat{\sigma}_p^2$ is the estimated residual variance. In general, we expect the following from a criterion function:

- Maximal coefficient of determination using as few explanatory variables as possible.

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`.

In the next step, by removing any explanatory variable from the estimated model

$$\texttt{HEAT}_j = b_0 + b_1\texttt{CHEM1}_j + b_2\texttt{CHEM2}_j + b_4\texttt{CHEM4}_j, \tag{2}$$

we just increase AIC. Thus we are left with the estimated model (2).

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

```
##
## Call:
## lm(formula = HEAT ~ CHEM1 + CHEM2 + CHEM4, data = hald)
##
## 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 6 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 assumption of residuals and thus, it is the safer alternative here.

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

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

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?