Solution to Series 1

1. a) Let the first observation go to infinity. Then:

$$\lim_{x_1 \to \infty} \frac{1}{n} \sum_{i=1}^{n} x_i = \infty$$

This means the arithmetic mean breaks. Thus the breaking point is $\varepsilon^* = 0$.

b) We order the observations $x_1, \ldots, x_n$ (n =odd):

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(k)} \leq x_{(k+1)} \leq x_{(k+2)} \leq \ldots \leq x_{(n)}$$

The median is the observation in the middle:

$$\text{med}(x_1, \ldots, x_n) = x_{(k+1)} = x_{\left(\frac{n+1}{2}\right)}$$

In order for the median to break, the absolute value of $x_{(k+1)}$ has to become large. At the same time, the absolute values of either $x_{(1)}, \ldots, x_{(k)}$ or $x_{(k+2)}, \ldots, x_{(n)}$ must become large as well. Because a total of $k+1$ observations need to be changed, the breaking point is $\varepsilon^*_n(\text{median}; x_1, \ldots, x_n) = k/(2k+1)$. For $k \to \infty$ this number converges to $1/2$.

2. a)

```
> d.ertrag <- scan(url("http://stat.ethz.ch/Teaching/Datasets/WBL/ertrag.dat"))
```

Robust estimation of the expected value:

```
> library(robustbase)
> (muh <- huberM(d.ertrag, k = 1.345, se = TRUE))
$mu
[1] 35.8

$s
[1] 0.297

$it
[1] 12

$SE
[1] 0.141
```

The robust confidence interval is then given by

```
> muh$mu + c(-1, 1) * qt(0.975, length(d.ertrag) - 1) * muh$SE
[1] 35.4 36.1
```

b) The classical confidence interval is given by

```
> t.test(d.ertrag)
        One Sample t-test

data:  d.ertrag
t = 56, df = 8, p-value = 1e-11
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 33.7 36.6
sample estimates:
mean of x
     35.1
```

Note that the classical confidence interval is much larger than the robust one.

3. **a)**

```
> library(MASS)
> D.oats <-  read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/oatsM16.dat",
                        header = TRUE)
> D.oats$FBlock <- as.factor(D.oats$Block)
> D.oats$FVariety <- as.factor(D.oats$Variety)
> str(D.oats)
'data.frame':        40 obs. of  6 variables:
 $ Variety  : int  1 2 3 4 5 6 7 8 1 2 ...
 $ Block    : int  1 1 1 1 1 1 1 1 2 2 ...
 $ ValuesOrg: int  296 402 437 303 469 345 324 488 357 390 ...
 $ Values   : num  287 402 480 303 469 ...
 $ FBlock   : Factor w/ 5 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 2 2 ...
 $ FVariety : Factor w/ 8 levels "1","2","3","4",..: 1 2 3 4 5 6 7 8 1 2 ...
> OatsOrg.lm <- lm(ValuesOrg ~ FVariety + FBlock, data = D.oats)
> ## robust
> OatsOrg.rlm <- rlm(ValuesOrg ~ FVariety + FBlock, data = D.oats, psi = psi.huber,
                     method = "M", maxit = 50)
```

- Residual standard error

```
> summary(OatsOrg.lm)
Call:
lm(formula = ValuesOrg ~ FVariety + FBlock, data = D.oats)

Residuals:
   Min     1Q Median     3Q    Max
-67.02 -20.39  -4.39  16.48  75.17

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  363.025     19.775   18.36  < 2e-16 ***
FVariety2     42.200     22.834    1.85  0.07517 .
FVariety3     28.200     22.834    1.23  0.22710
FVariety4    -47.600     22.834   -2.08  0.04635 *
FVariety5    105.000     22.834    4.60  8.3e-05 ***
FVariety6     -3.800     22.834   -0.17  0.86902
FVariety7    -14.000     22.834   -0.61  0.54475
FVariety8     49.800     22.834    2.18  0.03775 *
FBlock2      -25.500     18.052   -1.41  0.16880
FBlock3        0.125     18.052    0.01  0.99452
FBlock4      -42.000     18.052   -2.33  0.02745 *
FBlock5      -75.750     18.052   -4.20  0.00025 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36.1 on 28 degrees of freedom
Multiple R-squared:  0.75,       Adjusted R-squared:  0.651
F-statistic: 7.62 on 11 and 28 DF,  p-value: 7e-06
> summary(OatsOrg.rlm)
Call: rlm(formula = ValuesOrg ~ FVariety + FBlock, data = D.oats, psi = psi.huber,
    maxit = 50, method = "M")
Residuals:
   Min     1Q Median     3Q    Max
-65.31 -12.52  -1.94  12.83  92.02

Coefficients:
            Value   Std. Error t value
(Intercept) 361.306  17.615     20.511
FVariety2    40.268  20.340      1.980
FVariety3    16.568  20.340      0.815
```

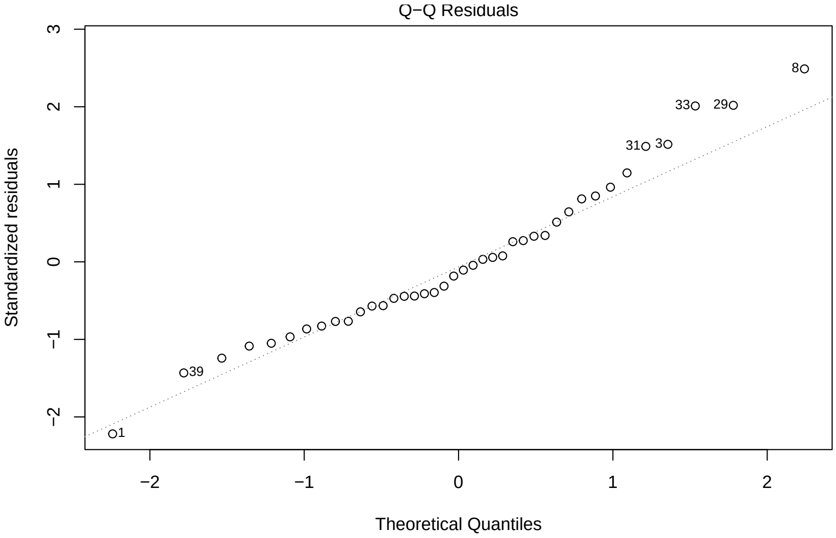
```
FVariety4   -49.532  20.340     -2.435
FVariety5    95.095  20.340      4.675
FVariety6    -5.732  20.340     -0.282
FVariety7   -17.580  20.340     -0.864
FVariety8    34.674  20.340      1.705
FBlock2     -16.838  16.081     -1.047
FBlock3       5.661  16.081      0.352
FBlock4     -43.831  16.081     -2.726
FBlock5     -69.863  16.081     -4.345

Residual standard error: 19.4 on 28 degrees of freedom
```

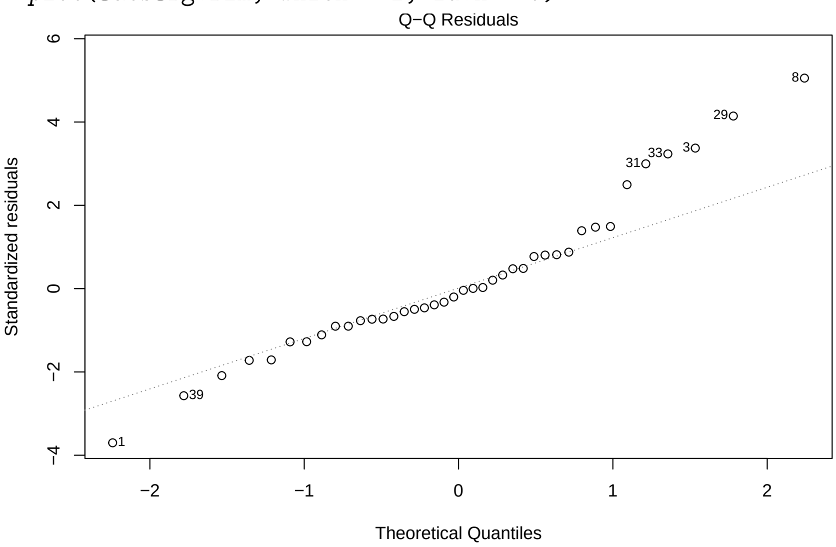
The residual standard error of the classical fit ($\hat{\sigma} = 36.1$) is almost two times larger than that of the robust fit ($\hat{\sigma} = 19.4$).

- Normal plot

```
> plot(OatsOrg.lm, which = 2, id.n = 7)
```




```
> plot(OatsOrg.rlm, which = 2, id.n = 7)
```



In the classical fit, there is only weak evidence that the distribution is slightly right skewed (Looking only at the qq-plot of the classical fit, we would assume the normality assumption is not violated). The robust fit shows this evidence much more clearly. In contrast to the classical fit we can identify a few outliers (this result depends on the definition of outlier):

```
> which(abs(resid(OatsOrg.lm)) > 2.2 * 36.1)   ## none
named integer(0)
> which(abs(resid(OatsOrg.rlm)) > 2.2 * 19.4) ## 1  3  8 29 31 33 39
```

```
 1  3  8 29 31 33 39
 1  3  8 29 31 33 39
```

- L_1 distance between the estimated coefficients

```
> sum(abs(coef(OatsOrg.rlm) - coef(OatsOrg.lm)))
[1] 69.7
> round(cbind(coef(OatsOrg.rlm), coef(OatsOrg.lm)), 1)
             [,1]  [,2]
(Intercept) 361.3 363.0
FVariety2    40.3  42.2
FVariety3    16.6  28.2
FVariety4   -49.5 -47.6
FVariety5    95.1 105.0
FVariety6    -5.7  -3.8
FVariety7   -17.6 -14.0
FVariety8    34.7  49.8
FBlock2     -16.8 -25.5
FBlock3       5.7   0.1
FBlock4     -43.8 -42.0
FBlock5     -69.9 -75.8
```

Clearly different values have been obtained as estimates.

- Are the two factor variables significant on the 5% level?

```
> drop1(OatsOrg.lm, test = "F")
Single term deletions

Model:
ValuesOrg ~ FVariety + FBlock
         Df Sum of Sq    RSS AIC F value  Pr(>F)
<none>                 36498 297
FVariety  7     76827 113324 328    8.42 1.6e-05 ***
FBlock    4     32443  68941 314    6.22   0.001 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Yes, both factor variables are significant on the 5% level. According to the normal plot, there is no real evidence that the result may not be valid. However, the robust fit indicates thatthe residuals are not normally distributed, but have a longer tail on the right-hand side.

b)

```
> OatsM.lm <- lm(Values ~ FVariety + FBlock, data = D.oats)
> ## robust
> OatsM.rlm <- rlm(Values ~ FVariety + FBlock, data = D.oats, psi = psi.huber,
                   method = "M", maxit = 50)
```

- Residual standard error

```
> summary(OatsM.lm)  ## sigma = 53.89
Call:
lm(formula = Values ~ FVariety + FBlock, data = D.oats)

Residuals:
   Min     1Q Median     3Q    Max
 -89.7  -27.7  -10.5   15.6  112.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   377.11      29.52   12.78  3.3e-13 ***
FVariety2      31.02      34.08    0.91   0.3705
FVariety3      25.62      34.08    0.75   0.4585
FVariety4     -58.78      34.08   -1.72   0.0956 .
FVariety5     102.42      34.08    3.01   0.0055 **
FVariety6     -14.98      34.08   -0.44   0.6637
```

```
FVariety7      -7.98      34.08   -0.23   0.8166
FVariety8      47.22      34.08    1.39   0.1769
FBlock2       -35.18      26.94   -1.31   0.2024
FBlock3        -9.55      26.94   -0.35   0.7257
FBlock4       -35.55      26.94   -1.32   0.1977
FBlock5       -77.36      26.94   -2.87   0.0077 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 53.9 on 28 degrees of freedom
Multiple R-squared:  0.574,        Adjusted R-squared:  0.407
F-statistic: 3.43 on 11 and 28 DF,  p-value: 0.00409
```

```
> summary(OatsM.rlm)  ## sigma = 19.37
```

```
Call: rlm(formula = Values ~ FVariety + FBlock, data = D.oats, psi = psi.huber,
    maxit = 50, method = "M")
Residuals:
   Min     1Q Median     3Q    Max
-73.90 -12.53  -1.94  12.83 138.10

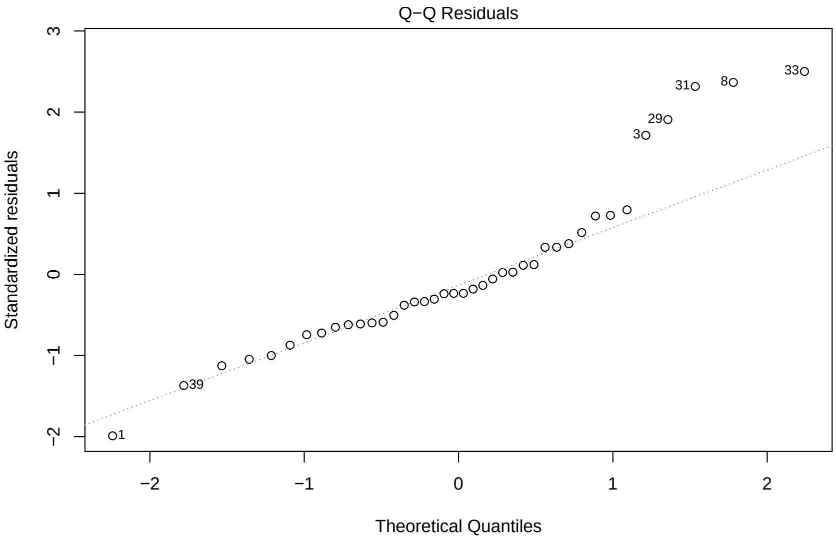
Coefficients:
            Value   Std. Error t value
(Intercept) 361.299  17.621     20.504
FVariety2    40.277  20.347      1.980
FVariety3    16.586  20.347      0.815
FVariety4   -49.523  20.347     -2.434
FVariety5    95.103  20.347      4.674
FVariety6    -5.723  20.347     -0.281
FVariety7   -17.572  20.347     -0.864
FVariety8    34.686  20.347      1.705
FBlock2     -16.842  16.085     -1.047
FBlock3       5.658  16.085      0.352
FBlock4     -43.830  16.085     -2.725
FBlock5     -69.868  16.085     -4.344

Residual standard error: 19.4 on 28 degrees of freedom
```

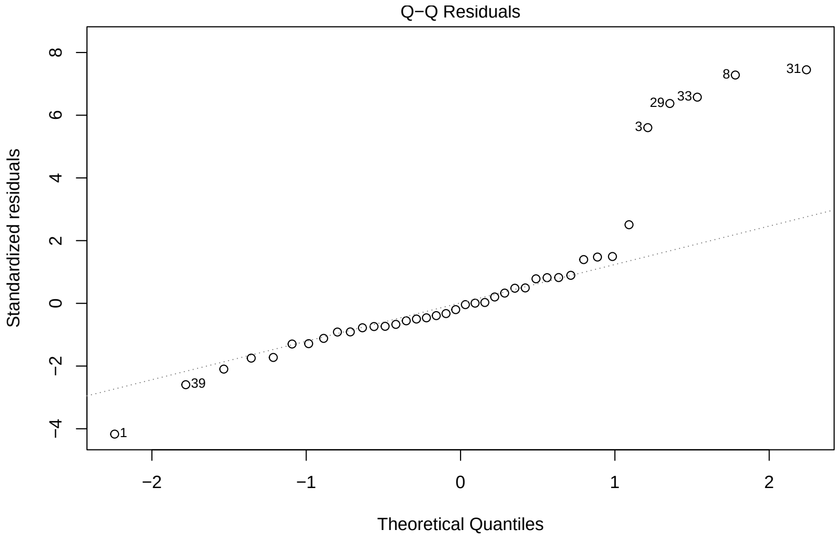
The residual standard error of the classical fit is almost three times larger than that of the robust fit. Compared to **a)**, the estimates of the robust fit have not changed.

- Normal plot

```
> plot(OatsM.lm, which = 2, id.n = 7)
```




```
> plot(OatsM.rlm, which = 2, id.n = 7)
```



In the classical fit, there is evidence that the distribution is right skewed. The robust fit shows this evidence much more clearly. In contrast to the classical fit we can again identify a few outliers (this result depends on the definition of outlier)

```
> which(abs(resid(OatsM.lm)) > 2.2 * 53.89)  ## none
named integer(0)
> which(abs(resid(OatsM.rlm)) > 2.2 * 19.37) ## 1  3  8 29 31 33 39
 1   3   8 29 31 33 39
 1   3   8 29 31 33 39
```

- L_1 distance between the estimated coefficients

  ```
  > sum(abs(coef(OatsM.rlm) - coef(OatsM.lm)))
  [1] 131
  ```

  The difference between these two estimates is even larger than in task **a)**.
- Are the two factor variables significant on the 5% level?

  ```
  > drop1(OatsM.lm, test = "F")
  Single term deletions

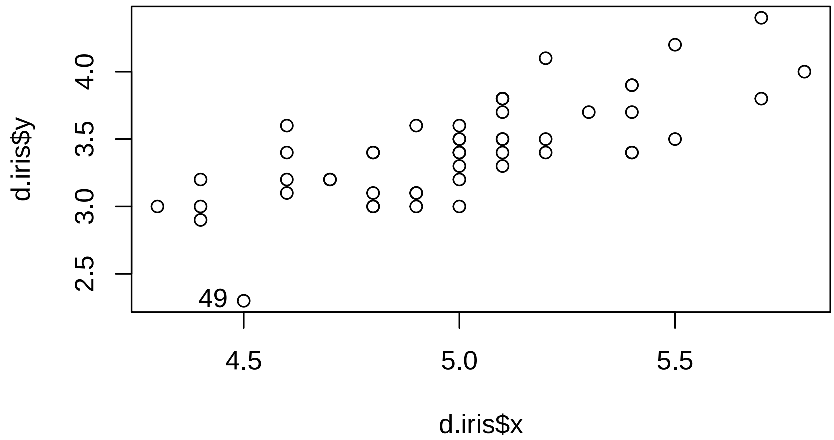
  Model:
  Values ~ FVariety + FBlock
           Df Sum of Sq    RSS AIC F value Pr(>F)
  <none>                 81315 329
  FVariety  7     80713 162028 342    3.97 0.0039 **
  FBlock    4     28859 110173 333    2.48 0.0664 .
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
  ```

  No, just factor variable `Variety` is significant on the 5% level, but not the factor variable `Block`. According to the normal plot of the classical fit, there is a weak evidence that the result may not be valid.
- ```
  > sum(abs(coef(OatsM.lm) - coef(OatsOrg.lm)))
  [1] 88.8
  > sum(abs(coef(OatsM.rlm) - coef(OatsOrg.rlm)))
  [1] 0.0918
  ```

  The estimates of the classical fits have changed in contrast to the robust fits which are (almost) identical.

4. a)

```
> d.iris <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/irisset.dat",
                       header = TRUE)
> plot(d.iris$x, d.iris$y)
> identify(d.iris$x, d.iris$y)
```
```




```
> fit0 <- lm(y ~ x, data = d.iris)
> beta0 <- coef(fit0)[2]
> d.iris[42, "y"] <- 2.5
> fit1 <- lm(y ~ x, data = d.iris)
> beta1 <- coef(fit1)[2]
> d.iris[42, "y"] <- 2.9
> fit2 <- lm(y ~ x, data = d.iris)
> beta2 <- coef(fit2)[2]
> d.iris[42, "y"] <- 3.3
> fit3 <- lm(y ~ x, data = d.iris)
> beta3 <- coef(fit3)[2]
> d.iris[42, "y"] <- 4.1
> fit4 <- lm(y ~ x, data = d.iris)
> beta4 <- coef(fit4)[2]
> c(beta0, beta1, beta2, beta3, beta4)
```

```
        x         x         x         x         x
0.7985283 0.7819060 0.7486613 0.7154167 0.6489274
```

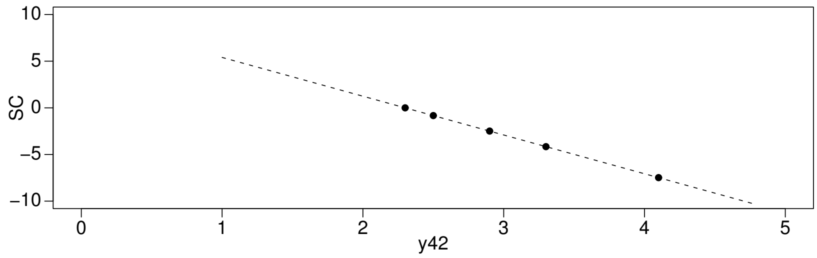
```
> ## SC
> SC1 <- 50 * (beta1 - beta0)
> SC2 <- 50 * (beta2 - beta0)
> SC3 <- 50 * (beta3 - beta0)
> SC4 <- 50 * (beta4 - beta0)
> c(SC1, SC2, SC3, SC4)
```

```
         x          x          x          x
-0.8311159 -2.4933478 -4.1555796 -7.4800434
```

If we vary the observation y_{42}, the estimation $\widehat{\beta}$ of the slope changes as follows:

	original value	new values			
y_{42}	2.3	2.5	2.9	3.3	4.1
$\widehat{\beta}$	0.7985	0.7819	0.7487	0.7154	0.6489
SC	0	-0.831	-2.493	-4.156	-7.480

The empirical influence function $SC = \frac{\widehat{\beta}(y_{42})-\widehat{\beta}_{(Orig)}}{1/50}$ is a straight line (compare with the results from **b)**).



b) For the slope coefficient $\widehat{\beta}$ it holds (see script):

$$\begin{aligned}
\widehat{\beta} &= \frac{\sum_{i=1}^{50}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{50}(x_i - \overline{x})^2} = \frac{1}{\mathrm{SS}_X} \cdot \sum_{i=1}^{50}(y_i - \overline{y})(x_i - \overline{x}) \\
&= \frac{1}{\mathrm{SS}_X} \cdot \left(\sum_{i=1}^{50} y_i(x_i - \overline{x}) - \sum_{i=1}^{50} \overline{y}(x_i - \overline{x}) \right) \\
&= \frac{1}{\mathrm{SS}_X} \cdot \left(\sum_{i=1}^{50} y_i(x_i - \overline{x}) - \overline{y} \cdot \sum_{i=1}^{50} (x_i - \overline{x}) \right) \\
&= \frac{1}{\mathrm{SS}_X} \cdot \sum_{i=1}^{50} y_i(x_i - \overline{x})
\end{aligned}$$

If we want to see how $\widehat{\beta}$ depends on y_{42}, we can write:

$$\begin{aligned}
\widehat{\beta}(y_{42}) &= \frac{1}{\mathrm{SS}_X} \cdot \sum_{i \neq 42} y_i(x_i - \overline{x}) + \frac{1}{\mathrm{SS}_X} \cdot y_{42}(x_{42} - \overline{x}) = \ldots \\
&= 0.989 - 0.083 \cdot y_{42}
\end{aligned}$$

and get a **linear** relation.
($x_{42} = 4.5$; $\overline{x} = 5.006$; $\mathrm{SS}_X = 6.088$; $\sum_{i \neq 42} y_i(x_i - \overline{x}) = 6.025$ – see R-Output)

For the empirical influence function it follows:

$$SC = \frac{\widehat{\beta}(y_{42}) - \widehat{\beta}}{1/50} = \frac{0.989 - 0.083 \cdot y_{42} - 0.7985}{1/50} = 9.558 - 4.156 \cdot y_{42}$$

R-Output:

```
> d.iris <- read.table("http://stat.ethz.ch/Teaching/Datasets/WBL/irisset.dat",
                       header = TRUE)
> ##   x_42
> d.iris[42, "x"]
[1] 4.5
> ##   mean(x)
> t.x <- d.iris[, "x"]
> mean(t.x)
[1] 5.006
> ##   SS_x
> sum((t.x - mean(t.x))^2) # or var(t.x) * (50-1)
[1] 6.0882
> ##   Sum of y_i * (x_i - mean(x)) without i=42
> t.ind <- rep(TRUE, nrow(d.iris))
> t.ind[42] <- FALSE
> t.y <- d.iris[, "y"]
> sum(t.y[t.ind] * (t.x[t.ind] - mean(t.x)))
```

```
[1] 6.0254
```

c) Regression-lines for the different y_{42}:

```
> plot(d.iris)
> abline(fit0)
> abline(fit1, lty = 2); points(d.iris[42, 1], 2.5, lty = 2, pch = 2)
> abline(fit2, lty = 3); points(d.iris[42, 1], 2.9, lty = 3, pch = 3)
> abline(fit3, lty = 4); points(d.iris[42, 1], 3.3, lty = 4, pch = 4)
> abline(fit4, lty = 5); points(d.iris[42, 1], 4.1, lty = 5, pch = 5)
```