Lösungen zur Übung 8

1. a)
```
>senic <- read.table("/afs/ethz.ch/users/s/sfs/datasets/senic2.dat",
            header=T)
> senic$school_factor(senic$school)
> senic$region_factor(senic$region)
```

b)
```
> summary(senic)
       id            length           age              inf
 Min.   :  1   Min.    : 6.700   Min.   :38.80   Min.   :1.300
 1st Qu.: 29   1st Qu.: 8.340   1st Qu.:50.90   1st Qu.:3.700
 Median : 57   Median : 9.420   Median :53.20   Median :4.400
 Mean   : 57   Mean    : 9.648   Mean   :53.23   Mean   :4.342
 3rd Qu.: 85   3rd Qu.:10.470   3rd Qu.:56.20   3rd Qu.:5.200
 Max.   :113   Max.    :19.560   Max.   :65.90   Max.   :7.800



    cult              xray           beds          school   region
Min.   :1.60    Min.   : 39.60   Min.   : 29.0     1:17     1:28
1st Qu.:8.40    1st Qu.: 69.50   1st Qu.:106.0     2:96     2:32
Median :14.10   Median : 82.30   Median :186.0              3:37
Mean   :15.79   Mean   : 81.63   Mean   :252.2              4:16
3rd Qu :20.30   3rd Qu.: 94.10   3rd Qu.:312.0
Max.   ::60.50  Max.   :133.50   Max.   :835.0


      pat              nurs             serv
 Min.   : 20.0   Min.   : 14.0    Min.   : 5.70
 1st Qu.: 68.0   1st Qu.: 66.0    1st Qu.:31.40
 Median :143.0   Median :132.0    Median :42.90
 Mean   :191.4   Mean   :173.2    Mean   :43.16
 3rd Qu.:252.0   3rd Qu.:218.0    3rd Qu.:54.30
 Max.   :791.0   Max.   :656.0    Max.   :80.00
```

Es hat zwei extreme Beobachtungen in `length`. Die Variablen `cult, beds, pat, nurs` (und `length`) sind rechtsschief. `beds, pat, nurs` (und etwas weniger stark `serv`) sind hochkorreliert, alles Variablen für die Grösse des Spitals.

c)
```
>
mod1_lm(length~inf+age+cult+xray+beds+school+region+pat+nurs+serv,data=senic)

   > summary(mod1,corr=F)

Call: lm(formula = length ~ inf + age + cult + xray + beds + school +
region + pat +
      nurs + serv, data = senic)
```

```
Residuals:
   Min     1Q   Median     3Q   Max
 -2.01 -0.652 -0.02861 0.5862 6.268

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.1962  1.6401     1.3391  0.1836
        inf  0.4580  0.1280     3.5791  0.0005
        age  0.0804  0.0281     2.8583  0.0052
       cult  0.0051  0.0159     0.3206  0.7492
       xray  0.0123  0.0071     1.7266  0.0873
       beds -0.0048  0.0036    -1.3236  0.1886
     school -0.1391  0.2197    -0.6329  0.5283
    region1 -0.4142  0.1752    -2.3644  0.0200
    region2 -0.2505  0.0940    -2.6651  0.0090
    region3 -0.3108  0.0909    -3.4184  0.0009
        pat  0.0150  0.0044     3.3999  0.0010
       nurs -0.0059  0.0022    -2.6752  0.0087
       serv -0.0125  0.0137    -0.9095  0.3653

Residual standard error: 1.226 on 100 degrees of freedom
Multiple R-Squared: 0.6328
F-statistic: 14.36 on 12 and 100 degrees of freedom, the p-value is
1.11e-016
```

```
d) > mod6_lm(length~inf+age+region+pat+nurs,data=senic)
   > summary(mod6,corr=F)

Call: lm(formula = length ~ inf + age + region + pat + nurs, data =
senic)
Residuals:
    Min      1Q  Median     3Q  Max
 -2.141 -0.7741 -0.1274 0.6764 6.41

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  2.7818  1.5053     1.8481  0.0674
        inf  0.5602  0.1000     5.6043  0.0000
        age  0.0712  0.0270     2.6348  0.0097
    region1 -0.5065  0.1654    -3.0614  0.0028
    region2 -0.2856  0.0895    -3.1909  0.0019
    region3 -0.3499  0.0869    -4.0276  0.0001
        pat  0.0094  0.0019     4.9543  0.0000
       nurs -0.0072  0.0021    -3.4829  0.0007

Residual standard error: 1.242 on 105 degrees of freedom
Multiple R-Squared: 0.6043
F-statistic: 22.91 on 7 and 105 degrees of freedom, the p-value is 0
```

```
e) > step1_step(mod1)
   Start:  AIC= 189.3379
   length ~ inf + age + cult + xray + beds + school + region + pat +
nurs + serv

Single term deletions

Model:
length ~ inf + age + cult + xray + beds + school + region + pat + nurs +
serv

scale:  1.502681

       Df Sum of Sq      RSS       Cp
<none>              150.2681 189.3379
   inf  1  19.24937 169.5175 205.5819
   age  1  12.27701 162.5452 198.6095
  cult  1   0.15441 150.4226 186.4869
  xray  1   4.47961 154.7477 190.8121
  beds  1   2.63264 152.9008 188.9651
school  1   0.60185 150.8700 186.9343
region  3  30.53510 180.8032 210.8569
   pat  1  17.37017 167.6383 203.7027
  nurs  1  10.75387 161.0220 197.0864
  serv  1   1.24309 151.5112 187.5756

Step:  AIC= 186.4869
 length ~ inf + age + xray + beds + school + region + pat + nurs + serv

Single term deletions

Model:
length ~ inf + age + xray + beds + school + region + pat + nurs + serv

scale:  1.502681

       Df Sum of Sq      RSS       Cp
<none>              150.4226 186.4869
   inf  1  27.68983 178.1124 211.1714
   age  1  12.53376 162.9563 196.0153
  xray  1   4.85660 155.2792 188.3381
  beds  1   2.49819 152.9207 185.9797
school  1   0.80636 151.2289 184.2879
region  3  35.26382 185.6864 212.7346
   pat  1  17.56658 167.9891 201.0481
  nurs  1  10.60853 161.0311 194.0901
  serv  1   1.34440 151.7670 184.8259

Step:  AIC= 184.2879
 length ~ inf + age + xray + beds + region + pat + nurs + serv
```

```
Single term deletions

Model:
length ~ inf + age + xray + beds + region + pat + nurs + serv

scale:  1.502681

       Df Sum of Sq      RSS       Cp
<none>              151.2289 184.2879
   inf  1  27.07358 178.3025 208.3561
   age  1  11.96327 163.1922 193.2458
  xray  1   5.13676 156.3657 186.4193
  beds  1   3.00103 154.2299 184.2836
region  3  34.72115 185.9501 209.9930
   pat  1  20.73927 171.9682 202.0218
  nurs  1  10.21999 161.4489 191.5025
  serv  1   1.17508 152.4040 182.4576

Step:  AIC= 182.4576
 length ~ inf + age + xray + beds + region + pat + nurs

Single term deletions

Model:
length ~ inf + age + xray + beds + region + pat + nurs

scale:  1.502681

       Df Sum of Sq      RSS       Cp
<none>              152.4040 182.4576
   inf  1  25.99921 178.4032 205.4515
   age  1  11.80512 164.2091 191.2574
  xray  1   5.32318 157.7272 184.7754
  beds  1   4.23969 156.6437 183.6919
region  3  33.54746 185.9515 206.9890
   pat  1  21.98165 174.3856 201.4339
  nurs  1  12.03075 164.4347 191.4830
```

Die Funktion `step()` minimiert AIC, resp Cp. Das resultierende Modell enthält im Gegensatz zum Vorhergehenden noch die Variablen `xray` und `beds`.

```
f) > sum(lmdiag(mod6)[,7])  [1] 195.8386
   > sum(lmdiag(step1)[,7]  [1] 189.9139
```

Die PRESS–Statistik vom Modell step1 ist kleiner.

g) Die Beobachtung-Nr. 47 ist ein Ausreisser, die Beobachtung-Nr. 112 ein Hebelpunkt. Das sind die zwei Spitäler mit den beiden längsten Hospitalisationsdauern.

```
> senic[c(47,112),]
     id length  age inf cult  xray beds school region pat nurs serv
 47  47  19.56 59.9 6.5 17.2 113.7  306      2      1 273  172 51.4
112 112  17.94 56.2 5.9 26.4  91.8  835      1      1 791  407 62.9
```

```
h) > mod17_lm(length~inf+age+region+pat,data=senic[-c(47,112),])
   > summary(mod17,corr=F)

Call: lm(formula = length ~ inf + age + region +
pat,data=senic[-c(47,112),])
Residuals:
    Min      1Q  Median     3Q Max
 -1.958 -0.7069 -0.1273 0.5646 3.1

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  4.1386  1.2311     3.3616  0.0011
        inf  0.4874  0.0813     5.9953  0.0000
        age  0.0513  0.0220     2.3283  0.0218
    region1 -0.2985  0.1359    -2.1971  0.0302
    region2 -0.2025  0.0727    -2.7831  0.0064
    region3 -0.3907  0.0692    -5.6488  0.0000
        pat  0.0022  0.0007     2.9184  0.0043

Residual standard error: 1.005 on 104 degrees of freedom
Multiple R-Squared: 0.5613
F-statistic: 22.18 on 6 and 104 degrees of freedom, the p-value is
1.11e-016
```

Die Variable `nurs` ist nicht mehr signifikant.

Die automatisierte Variante mit der `step()`- Funktion liefert ebenfalls ein Modell ohne `nurs`, und ohne `beds`.

```
 > step2_step(mod11)
 > summary(step2,corr=F)

Call: lm(formula = length ~ inf + age + xray + region + pat,
data=senic[-c(47,112),])
Residuals:
    Min      1Q   Median     3Q   Max
 -1.886 -0.6974 -0.07325 0.5115 3.043

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  3.5440  1.2800     2.7687  0.0067
        inf  0.4279  0.0892     4.7965  0.0000
        age  0.0530  0.0219     2.4174  0.0174
       xray  0.0090  0.0057     1.5684  0.1199
    region1 -0.2772  0.1356    -2.0440  0.0435
    region2 -0.1804  0.0736    -2.4514  0.0159
    region3 -0.3744  0.0695    -5.3889  0.0000
        pat  0.0023  0.0007     3.1514  0.0021

Residual standard error: 0.9976 on 103 degrees of freedom
Multiple R-Squared: 0.5716
```

```
F-statistic: 19.63 on 7 and 103 degrees of freedom, the p-value is
2.22e-016
```