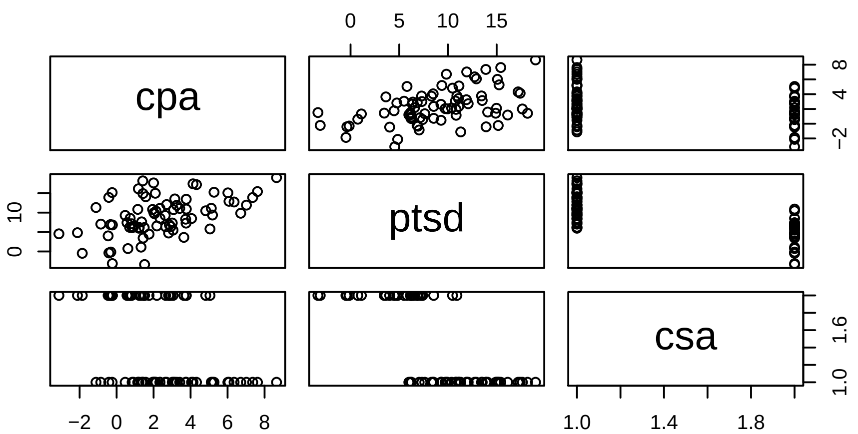
# Solution to Series 4

**1. a)** Read in the data and look at the data, do you see any problems? Make sure that all the variables are in the correct R data type.

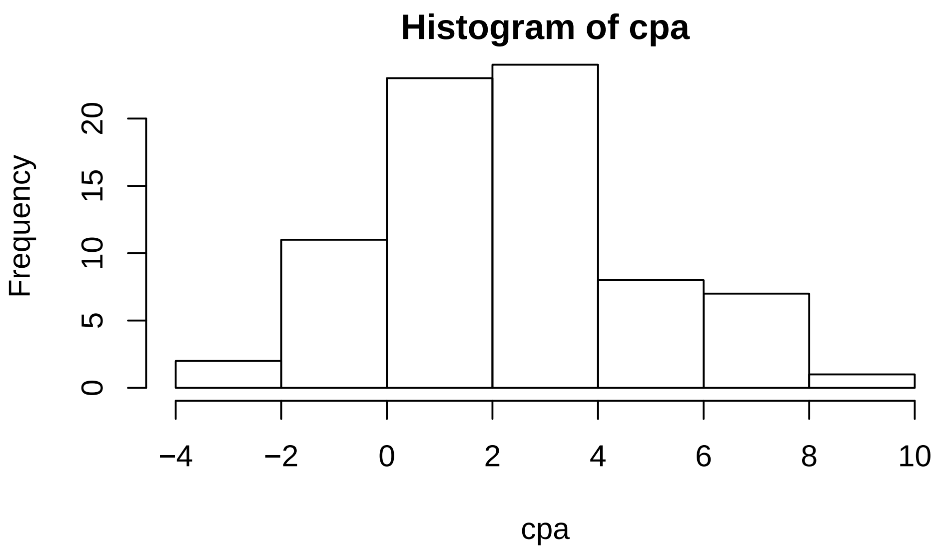
```
> sexab <- read.csv("http://stat.ethz.ch/Teaching/Datasets/abuse.csv",header=TRUE)
> attach(sexab)
```

Look at the data:

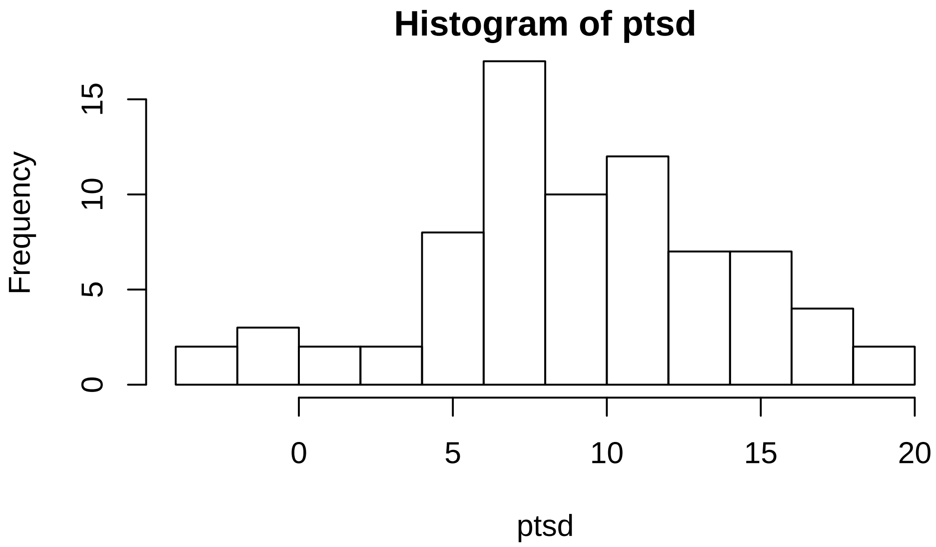
```
> pairs(sexab)
```




```
> hist(cpa)
```



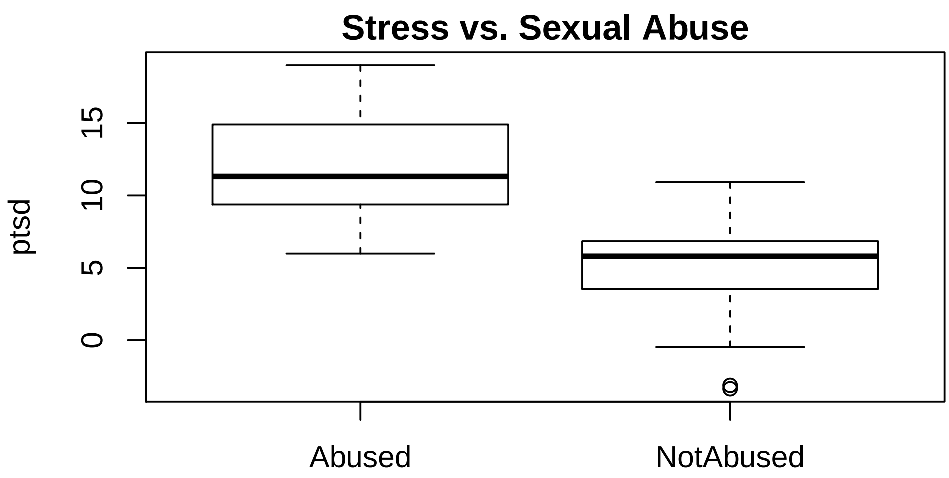

```
> hist(ptsd)
```




No data problems. No transformations necessary.

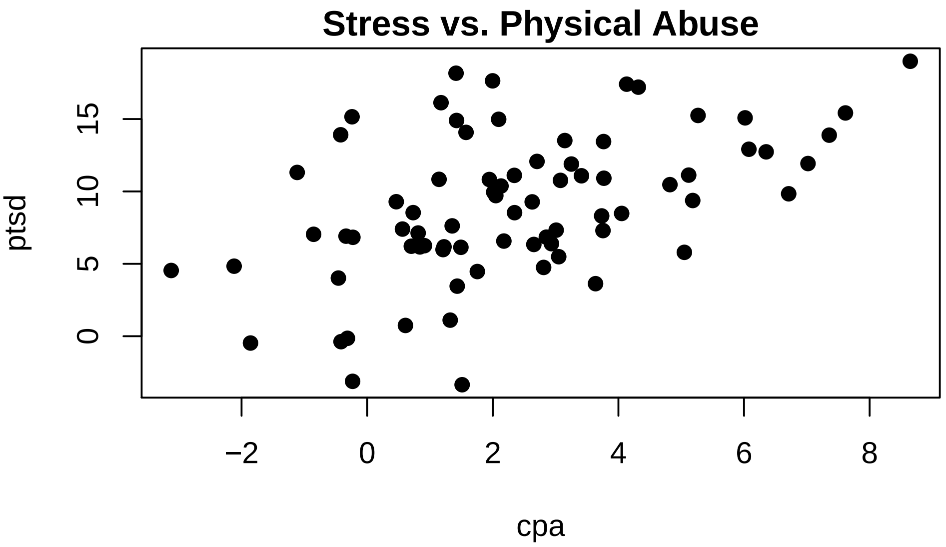
**b)** Use scatter plots and box plots to display the variable `ptsd` in dependence of the variables `csa` and `cpa`. Box plot of ptsd vs. csa:

```
> boxplot(ptsd ~ csa, ylab="ptsd", main="Stress vs. Sexual Abuse")
```




Scatter plot of ptsd vs. cpa:

```
> plot(ptsd ~ cpa, ylab="ptsd", main="Stress vs. Physical Abuse", pch=19)
```



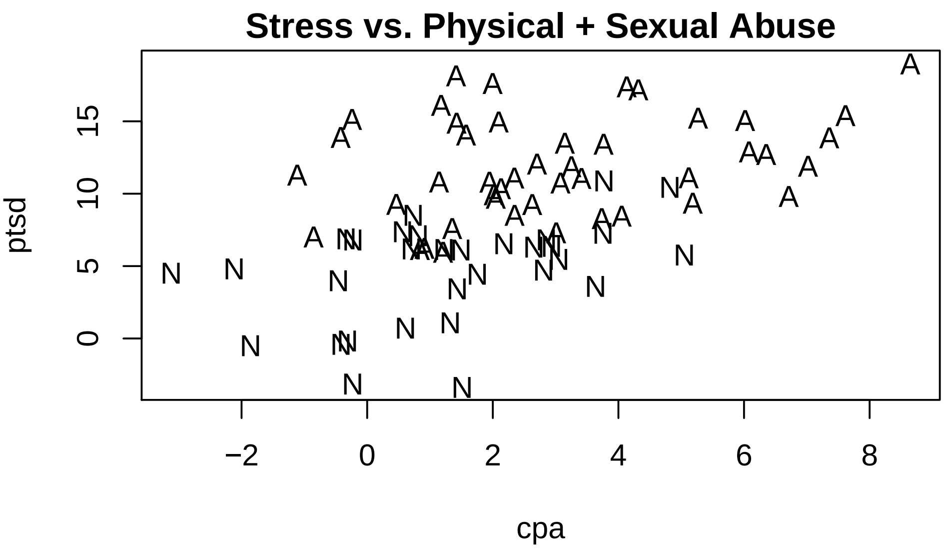

This scatter plot could be misleading. The fact that we plot both groups of woman in one plot could indicate a bigger dependence of `ptsd` and `cpa` as there really is.

**c)** Make a scatter plot of `ptsd` against `cpa`. Use different symbols for abused and non-abused woman. R-hint:

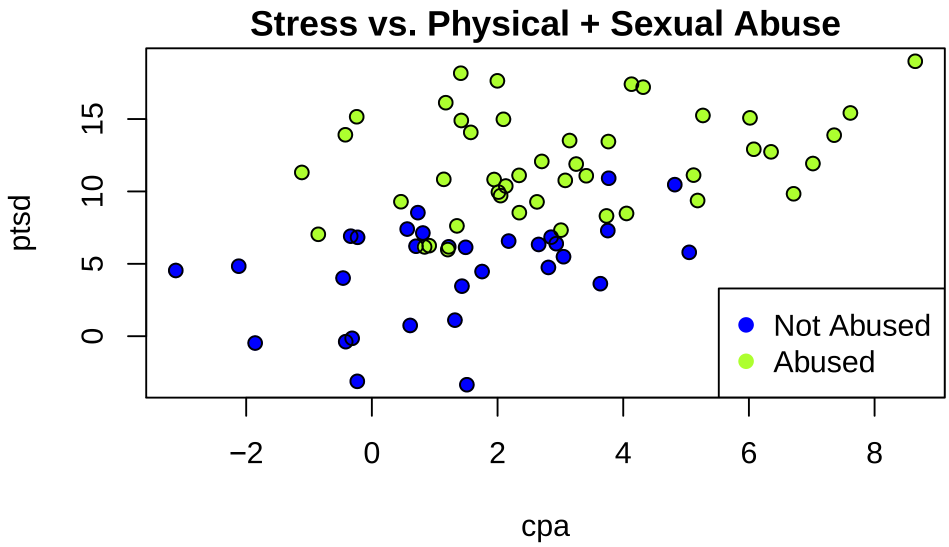
```
plot(cpa, ptsd, type="n")
text(cpa, ptsd, labels=substring(csa,1,1))
```

Scatter plot with different symbols for the different groups.

```
> plot(ptsd ~ cpa, ylab="ptsd", main="Stress vs. Physical + Sexual Abuse", type="n")
> text(cpa, ptsd, labels=substring(csa, 1, 1))
```




```
> plot(ptsd ~ cpa, pch=19, col="blue", main="Stress vs. Physical + Sexual Abuse")
> points(ptsd ~ cpa, pch=19, col="greenyellow", subset=(csa=="Abused"))
> points(ptsd ~ cpa)
> legend("bottomright", legend=c("Not Abused", "Abused"),
        pch=19, col=c("blue", "greenyellow"))
```




From this plots we see that the dependence between `ptsd` and `cpa` is not that big. But the two groups differ much concerning the stress-level. We now do a coherent analysis via quantitative methods.

**d)** Carry out a test in order to see if sexual abused woman have a higher PTSD-score. Why doesn't this test give you a complete answer? Hint: Look at the scatter plot from part c.).

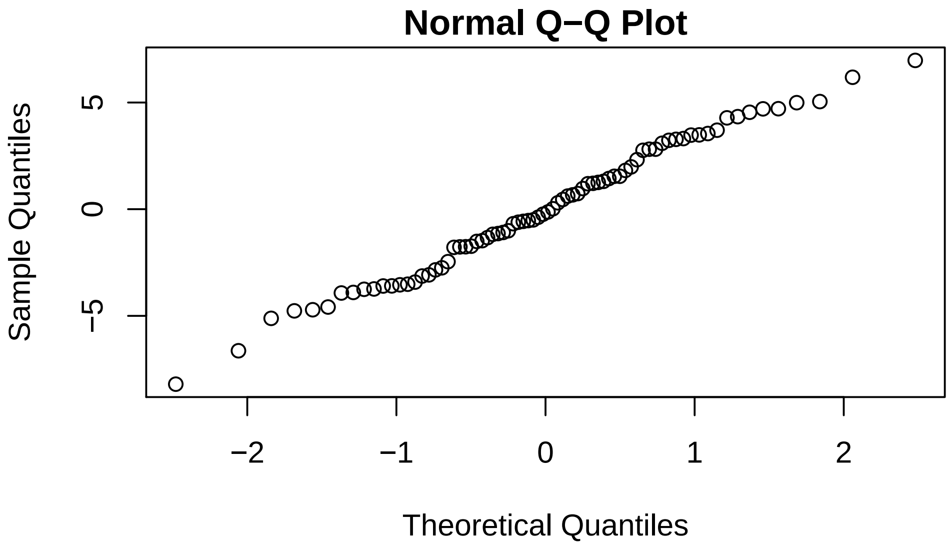
```
> t.test(ptsd ~ csa, paired=FALSE, var.equal=TRUE)

        Two Sample t-test

data:  ptsd by csa
t = 8.9387, df = 74, p-value = 2.172e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 5.630165 8.860273
sample estimates:
   mean in group Abused mean in group NotAbused
              11.941093                4.695874
```

The null-hypothesis gets rejected. This shows us that there is a statistically significant difference in stress-level between the two groups of woman. But what's with the factor physical abuse. We suggest that also the factor physical abuse has a influence on the stress-level. That is, we have to take both variables in to account at the same time. For that we fit a linear regression model.

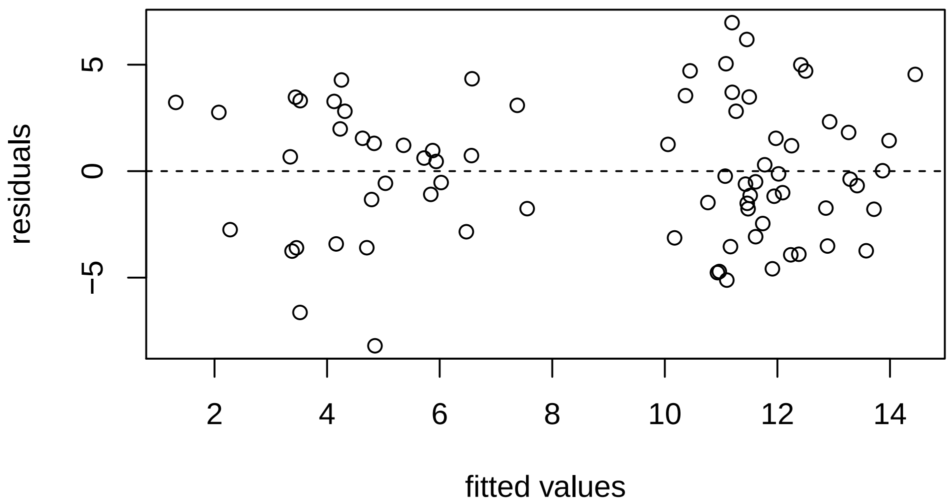
**e)** Fit a regression model to the data with both predictors and their interaction. What do the resulting coefficients mean?

```
> fit.interact <- lm(ptsd ~ cpa * csa, data=sexab)
> summary(fit.interact)
Call:
lm(formula = ptsd ~ cpa * csa, data = sexab)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1999 -2.5313 -0.1807  2.7744  6.9748

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.5571     0.8063  13.094  < 2e-16 ***
cpa                0.4500     0.2085   2.159   0.0342 *
csaNotAbused      -6.8612     1.0747  -6.384 1.48e-08 ***
cpa:csaNotAbused   0.3140     0.3685   0.852   0.3970
```

```
---
Signif. codes:  0
> qqnorm(fit.interact$resid)
```




```
> plot(fit.interact$fitted,fit.interact$resid,xlab="fitted values",ylab="residuals")
> abline(h=0,lty=2)
```




**2.** **a)** Get an overview of the data and account for possible problems. Which of the variables need to be transformed or not?

Overview over the data:

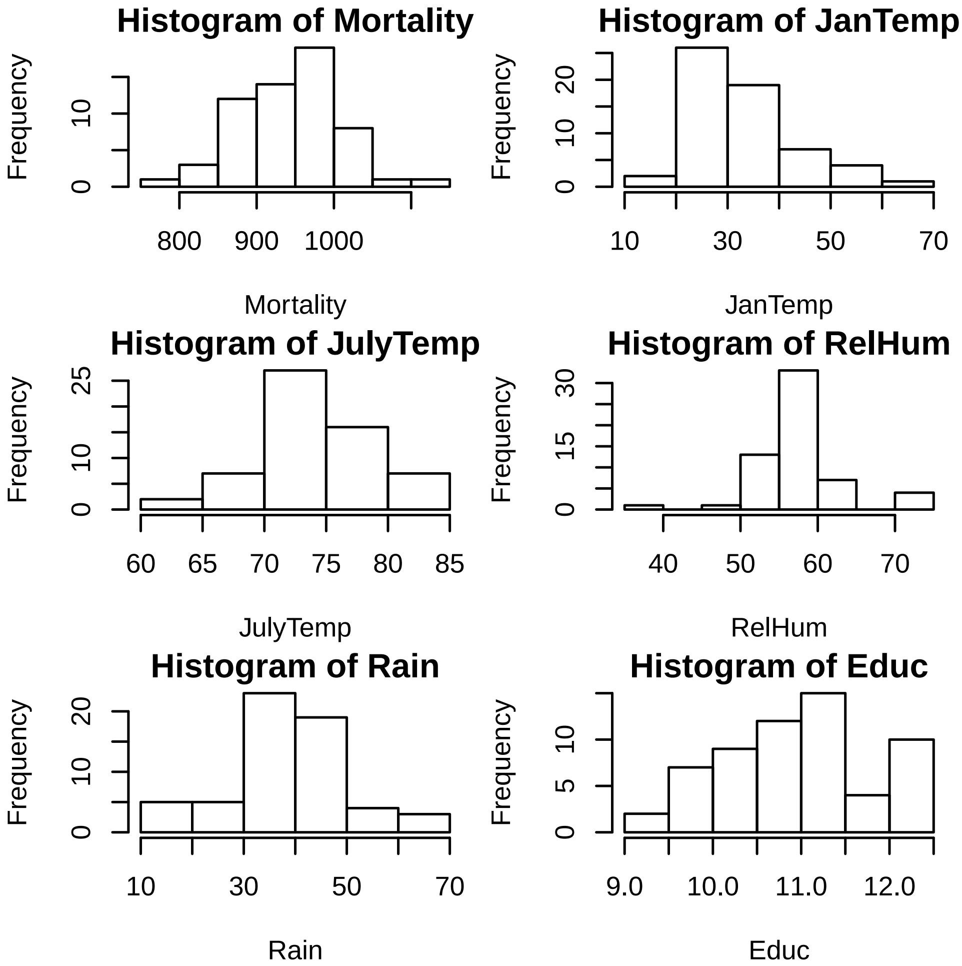
```
> mortality <- read.csv("http://stat.ethz.ch/Teaching/Datasets/mortality.csv",
                        header=TRUE)
> attach(mortality)
> summary(mortality)
                          City       Mortality
 Akron, OH                  : 1   Min.   : 790.7
 Albany-Schenectady-Troy, NY: 1   1st Qu.: 899.4
 Allentown, Bethlehem,PA-NJ : 1   Median : 946.2
 Atlanta, GA                : 1   Mean   : 941.2
 Baltimore, MD              : 1   3rd Qu.: 984.1
```

```
 Birmingham, AL             : 1   Max.   :1113.2  
 (Other)                    :53                   
    JanTemp        JulyTemp         RelHum     
 Min.   :12.0   Min.   :63.00   Min.   :38.00  
 1st Qu.:27.0   1st Qu.:72.00   1st Qu.:55.50  
 Median :31.0   Median :74.00   Median :57.00  
 Mean   :33.8   Mean   :74.41   Mean   :57.75  
 3rd Qu.:39.5   3rd Qu.:77.00   3rd Qu.:60.00  
 Max.   :67.0   Max.   :85.00   Max.   :73.00  

      Rain            Educ             Dens     
 Min.   :10.00   Min.   : 9.00   Min.   :1441  
 1st Qu.:33.50   1st Qu.:10.40   1st Qu.:3138  
 Median :38.00   Median :11.00   Median :3626  
 Mean   :38.51   Mean   :10.97   Mean   :3910  
 3rd Qu.:44.00   3rd Qu.:11.50   3rd Qu.:4566  
 Max.   :65.00   Max.   :12.30   Max.   :9699  

    NonWhite      WhiteCollar         Pop         
 Min.   : 0.80   Min.   :33.80   Min.   : 124833  
 1st Qu.: 4.90   1st Qu.:43.40   1st Qu.: 566515  
 Median : 9.50   Median :45.50   Median : 914427  
 Mean   :11.88   Mean   :46.39   Mean   :1438037  
 3rd Qu.:15.70   3rd Qu.:49.90   3rd Qu.:1717201  
 Max.   :38.50   Max.   :62.20   Max.   :8274961  

     House           Income            HC        
 Min.   :2.650   Min.   :25782   Min.   :  1.00  
 1st Qu.:3.210   1st Qu.:30004   1st Qu.:  7.00  
 Median :3.270   Median :32452   Median : 15.00  
 Mean   :3.247   Mean   :33247   Mean   : 38.47  
 3rd Qu.:3.360   3rd Qu.:35496   3rd Qu.: 30.50  
 Max.   :3.530   Max.   :47966   Max.   :648.00  

      NOx              SO2        
 Min.   :  1.00   Min.   :  1.00  
 1st Qu.:  4.00   1st Qu.: 13.00  
 Median :  9.00   Median : 32.00  
 Mean   : 22.97   Mean   : 54.66  
 3rd Qu.: 24.50   3rd Qu.: 70.00  
 Max.   :319.00   Max.   :278.00  
```

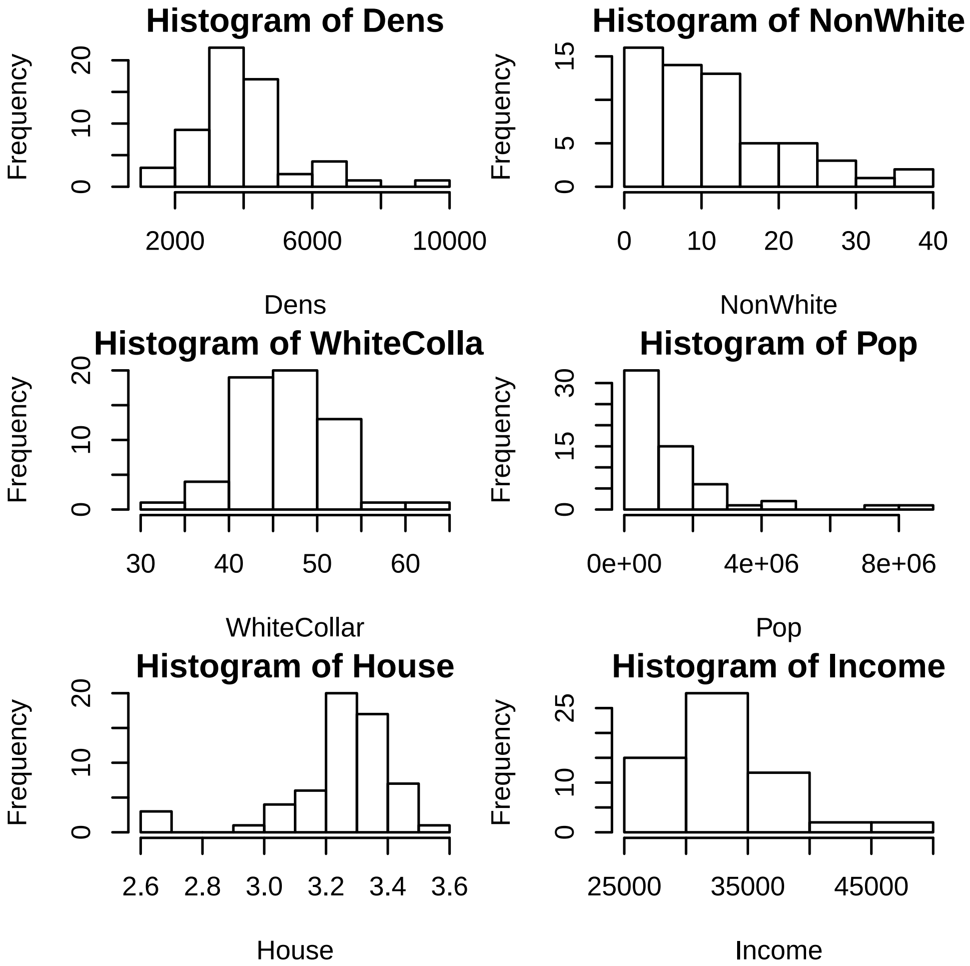
```
> rownames(mortality) <- mortality$City
> mortality <- mortality[,-1]
```

We do not see any big data problems. We set city as row names and delete the variable city. Transformationen:

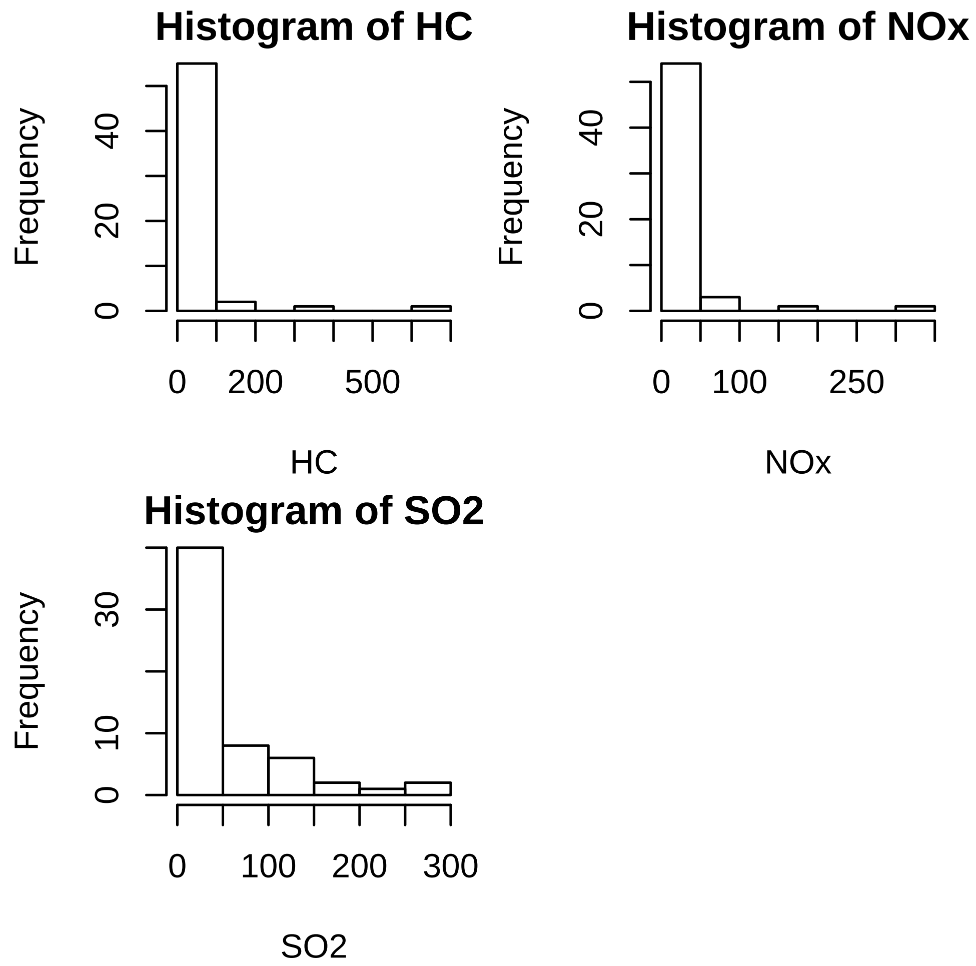
```
> par(mfrow=c(3,2))
> hist(Mortality)    ## ok, no transformation
> hist(JanTemp)      ## ok, no transformation
> hist(JulyTemp)     ## ok, no transformation
> hist(RelHum)       ## ok, no transformation
> hist(Rain)         ## ok, no transformation
> hist(Educ)         ## ok, no transformation
```




```
> par(mfrow=c(3,2))
> hist(Dens)         ## right skewed, log-tranformation recomendable
> hist(NonWhite)     ## ratio, arcsin-transformation recomendable
> hist(WhiteCollar) ## ratio, arcsin-transformation recomendable
> hist(Pop)          ## right skewed, log-tranformation recomendable
> hist(House)        ## ok, no transformation
> hist(Income)       ## right skewed, log-tranformation recomendable
```




```
> par(mfrow=c(2,2))
> hist(HC)          ## strongly right skewed, log-tranformation mandatory
> hist(NOx)         ## strongly right skewed, log-tranformation mandatory
> hist(SO2)         ## strongly right skewed, log-tranformation mandatory
```




```
> detach(mortality)
> mortality$Dens        <- log(mortality$Dens)
> mortality$NonWhite    <- asin(sqrt(mortality$NonWhite/100))
> mortality$WhiteCollar <- asin(sqrt(mortality$WhiteCollar/100))
> mortality$Pop         <- log(mortality$Pop)
> mortality$Income      <- log(mortality$Income)
> mortality$HC          <- log(mortality$HC)
> mortality$NOx         <- log(mortality$NOx)
> mortality$SO2         <- log(mortality$SO2)
> attach(mortality)
```

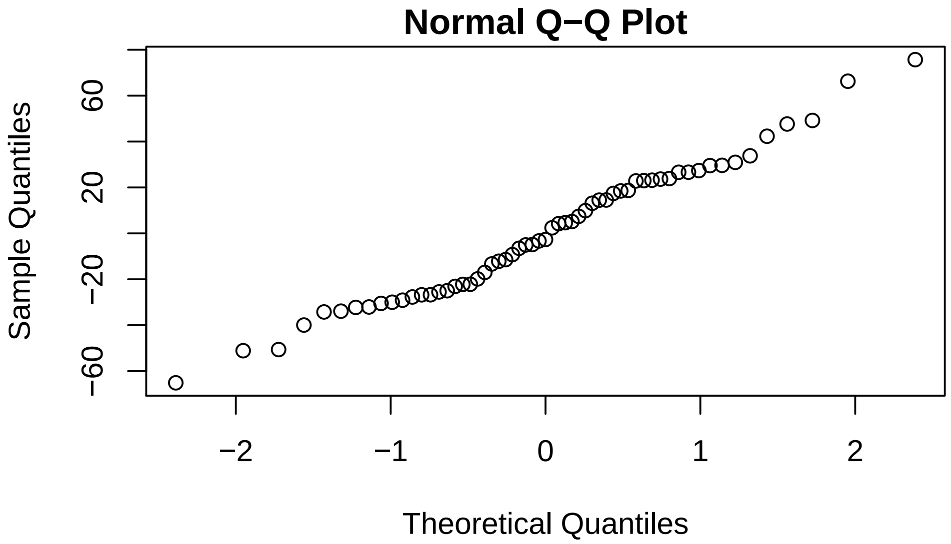
**b)** Carry out a multiple linear regression containing all variables. Does the model fit well? Check the residuals.

Full model:

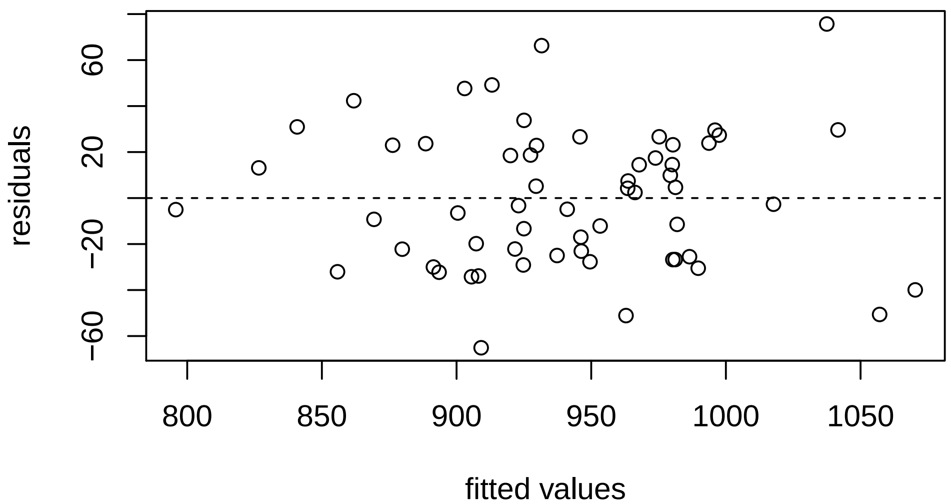
```
> fit <- lm(Mortality ~ ., data=mortality)
> summary(fit)
```

```
Call:
lm(formula = Mortality ~ ., data = mortality)

Residuals:
   Min     1Q Median     3Q    Max
-65.08 -25.23  -2.67  23.08  75.70

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1496.4915   572.7205   2.613  0.01224 *
```

```
JanTemp       -2.4479     0.8808  -2.779  0.00798 **
JulyTemp      -1.9350     2.0329  -0.952  0.34638
RelHum         0.1065     1.0614   0.100  0.92052
Rain           1.7727     0.5748   3.084  0.00352 **
Educ         -13.3849     8.7561  -1.529  0.13351
Dens          11.9490    16.1836   0.738  0.46423
NonWhite     326.6757    62.9092   5.193 5.09e-06 ***
WhiteCollar -146.3477   112.5510  -1.300  0.20028
Pop            4.8037     7.7245   0.622  0.53723
House        -43.2697    38.9460  -1.111  0.27260
Income       -27.3906    47.8041  -0.573  0.56958
HC           -21.1925    15.1050  -1.403  0.16763
NOx           35.7323    14.3143   2.496  0.01637 *
SO2           -5.3995     7.4040  -0.729  0.46970
---
Signif. codes:  0
> qqnorm(fit$resid)
```



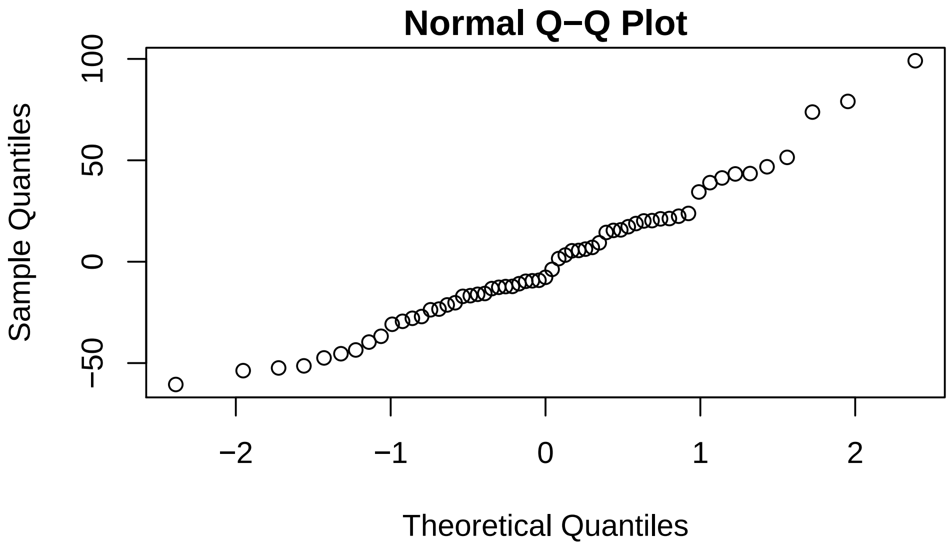

```
> plot(fit$fitted,fit$resid,xlab="fitted values",ylab="residuals")
> abline(h=0,lty=2)
```



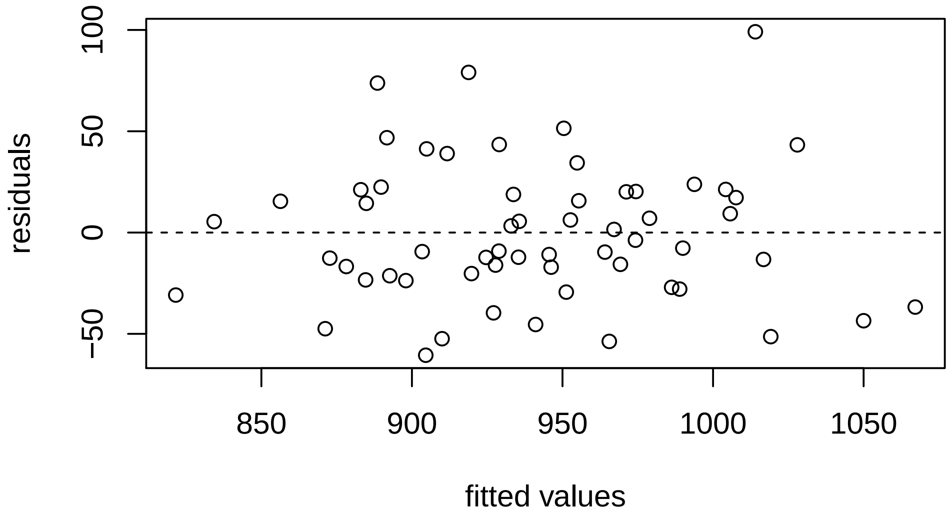

This model fits quite well, i.e. the model assumptions are fulfilled. We do not se any violation of the model assumptions.

**c)** Now take all the non-significant variables out of the model and compute the regression again. Compare your results to the one from part b.).
Now just use the significant variables:

```
> fit.sign <- lm(Mortality ~ JanTemp + Rain + NonWhite + NOx, data=mortality)
> summary(fit.sign)

Call:
lm(formula = Mortality ~ JanTemp + Rain + NonWhite + NOx, data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-60.537 -22.328  -7.677  20.186  99.117

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 788.6724    25.8034  30.565  < 2e-16 ***
JanTemp      -2.4277     0.5166  -4.699 1.84e-05 ***
Rain          2.4648     0.4692   5.254 2.59e-06 ***
NonWhite    277.1610    40.9045   6.776 9.53e-09 ***
NOx          20.6490     4.5502   4.538 3.21e-05 ***
---
Signif. codes:  0

> qqnorm(fit.sign$resid)
```




```
> plot(fit.sign$fitted,fit.sign$resid,xlab="fitted values",ylab="residuals")
> abline(h=0,lty=2)
```




Now all the variables are highly significant. The error variance is sligtly bigger, R-squared and also adjusted R-squared are smaller compared to the full model. On the oder hand is the p-value of the F-test bigger now.

Even though leaving away all of the non-significant variables worked quite well here, one sould not do that. A better strategie would be to delete the non-significant variables step by step, always deleting the one with the biggest p-value.

**d)** Start with the full multiple linear model. Remove now step by step the variable with the biggest p-value as long as it is over 0.05. Compare the result to the one from c.). R-hint: Use the R-function `update()`.

Step by step strategie: Use the function `update()`.

```
> fit.reduc <- fit
> fit.reduc <- update(fit.reduc, ~.-RelHum)     ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-Income)     ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-Pop)        ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-Dens)       ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-SO2)        ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-JulyTemp)   ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-HC)         ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-House)      ; summary(fit.reduc)
> fit.reduc <- update(fit.reduc, ~.-WhiteCollar); summary(fit.reduc)
```

```
Call:
lm(formula = Mortality ~ JanTemp + Rain + Educ + NonWhite + NOx,
    data = mortality)

Residuals:
    Min      1Q  Median      3Q     Max
-83.471 -23.987   4.444  19.880  85.943

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 992.2069    79.6994  12.449  < 2e-16 ***
JanTemp      -2.1304     0.5017  -4.246 8.80e-05 ***
Rain          1.8122     0.5066   3.577 0.000752 ***
Educ        -16.4207     6.1202  -2.683 0.009710 **
NonWhite    268.2564    38.8832   6.899 6.56e-09 ***
NOx          18.3230     4.3960   4.168 0.000114 ***
---
Signif. codes:  0
```

Now we stop because all of the remaining variables are significant. The error-variance, R-squared and p-value of the F-test look better then in the model from part c.). Also the residuals are looking good.

**e)** Again starting from the full model, carry out partial F-tests, in order to answer the question if

- all meteo-variables
- all air pollution-variables and
- all demographic-variables

can be removed from the model. Use the R-function `anova()`.

Fitting the model without the meteo-variables:

```
> fit.ohne.meteo <- lm(Mortality ~ .-JanTemp-JulyTemp-RelHum-Rain, data=mortality)
> anova(fit, fit.ohne.meteo)
```

```
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - JanTemp - JulyTemp - RelHum - Rain
  Res.Df   RSS Df Sum of Sq      F   Pr(>F)
1     44 51543
2     48 71705 -4    -20162 4.3027 0.005037 **
---
Signif. codes:  0
```

With the function `anova()` one carrys out a F-test in order to compare the two models. This test is significant, i.e. the null-hypothesis gets rejected. That is, the bigger model, the one with the meteo-variables, is better. So we can not leave away the meteo-variables.

Fitting the model without the air pollution-variables:

```
> fit.ohne.luft <- lm(Mortality ~ .-HC-NOx-SO2, data=mortality)
> anova(fit, fit.ohne.luft)
```

```
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - HC - NOx - SO2
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     44 51543
2     47 61244 -3   -9700.8 2.7604 0.0533 .
---
Signif. codes:  0
```

The partial F-test is not siginificant. Hence we can take the air pollution-variables out of the model.

Fitting the model without the demographic-variables:

```
> fit.ohne.demografie <- lm(Mortality ~ .-Educ-Dens-NonWhite-WhiteCollar-Pop-House
                              -Income, data=mortality)
> anova(fit, fit.ohne.demografie)
```

```
Analysis of Variance Table

Model 1: Mortality ~ JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
```

```
    SO2
Model 2: Mortality ~ (JanTemp + JulyTemp + RelHum + Rain + Educ + Dens +
    NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
    SO2) - Educ - Dens - NonWhite - WhiteCollar - Pop - House -
    Income
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1     44  51543
2     51 101406 -7    -49863 6.0808 5.369e-05 ***
---
Signif. codes:  0
```

The p-value of the test is very small, that is we can not leave away the demographic-variables.