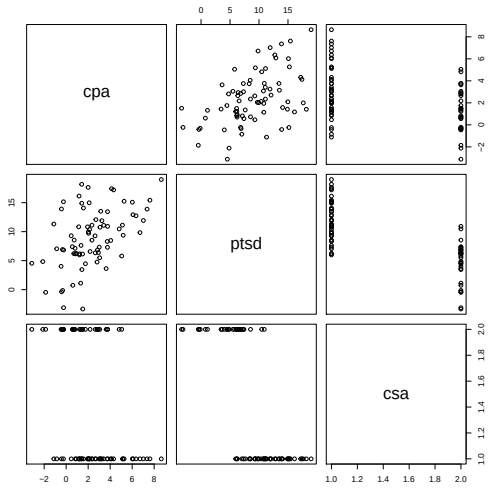
Exercise 1

The data in this example comes from a study of the effects of childhood sexual abuse on adult females reported in Rodriguez et al. ("Post-traumatic stress disorder in adult female survivors of childhood sexual abuse: a comparison study", Journal of Consulting and Clinical Psychology, 1997). 45 women who reported childhood sexual abuse (`csa`) were measured for post-traumatic stress disorder (`ptsd`) and childhood physical abuse (`cpa`), both on standardized scales. Additionally, the same quantities were recorded for 31 women who did not experience childhood sexual abuse. The dependent variable is `ptsd`. The data can be downloaded from the website. Read in the data with `read.table(...,sep=",",header=TRUE)`.

```
# reda in the data
dat = read.table(paste0(dir,"data/abuse.csv"), sep = ",", header = TRUE)

head(dat)

##        cpa     ptsd    csa
## 1  2.04786  9.71365 Abused
## 2  0.83895  6.16933 Abused
## 3 -0.24139 15.15926 Abused
## 4 -1.11461 11.31277 Abused
## 5  2.01468  9.95384 Abused
## 6  6.71131  9.83884 Abused
```

(a) Read in the data and investigate it graphically using the R function `pairs()`. Additionally, check if R reads the data correctly (i.e. ptsd and cpa as numerical variables, csa as factor variable).

```
# plot the three variables against each other
pairs(dat)
```




```
# the plots already indicate that childhood physical and sexual
# abose are associated with PTSD.
```

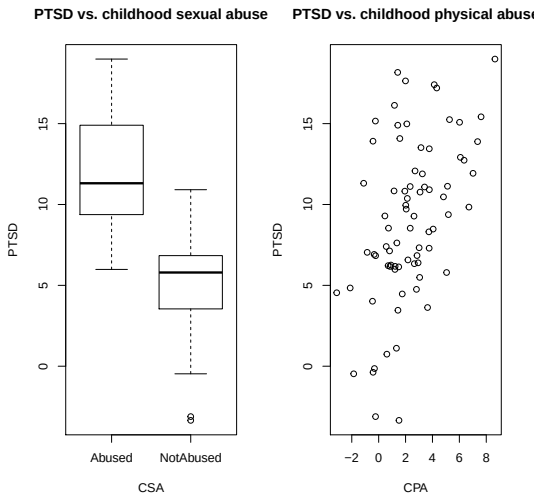
```
str(dat)
```

```
## 'data.frame': 76 obs. of  3 variables:
##  $ cpa : num  2.048 0.839 -0.241 -1.115 2.015 ...
##  $ ptsd: num  9.71 6.17 15.16 11.31 9.95 ...
##  $ csa : Factor w/ 2 levels "Abused","NotAbused": 1 1 1 1 1 1 1 1 1 1 ...
```

```
# the read.table() function automatically reads the data
# correctly into R.
```

(b) Investigate the relationship between the variable `ptsd` and `csa` respectively `ptsd` and `cpa` graphically.

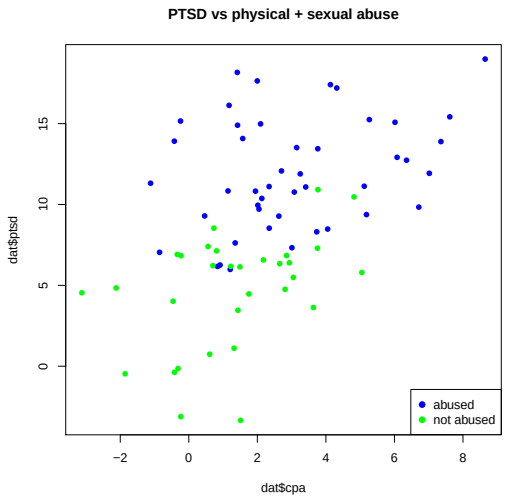
```
par(mfrow = c(1,2))

boxplot(dat$ptsd ~ dat$csa, ylab = "PTSD", xlab="CSA",
        main = "PTSD vs. childhood sexual abuse")
plot(dat$ptsd ~ dat$cpa, ylab = "PTSD", xlab="CPA",
     main = "PTSD vs. childhood physical abuse")
```




```
# We see a strong relationship between the both explanatory
# variables and the outcome ptsd.
```

(c) Now, create a scatter plot of `ptsd` against `cpa`. Use different colors for abused and non-abused women. What's the problem if we don't separate by abused and non-abused women. (**R-Hint**: First use `plot(...,type="n", pch=16)`. Then use `points(...,pch=16, col=...)` to plot the points for each subset.)

```
# plot cpa vs ptsd
plot(dat$cpa, dat$ptsd, main="PTSD vs physical + sexual abuse",
     pch=16, type="n")
points(dat$cpa[dat$csa=="Abused"], dat$ptsd[dat$csa=="Abused"],
       pch = 16, col="blue")
points(dat$cpa[dat$csa=="NotAbused"], dat$ptsd[dat$csa=="NotAbused"],
       pch = 16, col="green")
legend("bottomright", legend=c("abused", "not abused"),
       pch=19, col=c("blue", "green"))
```




```
# If we don't separate by abused and non-absued women we would
# estimate a bigger dependency between cpa and ptsd than there
# really is.
```

(d) Carry out a test in order to see if sexually abused women have a higher PTSD-score. Why does this test not give you a complete conclusion of the statistical dependence between `ptsd` and the predictors `cpa` and `csa`?

```
# We do a two sample t-test to evaluate whether or not there
# exists a significant difference between the population means
# of the two groups of women.
t.test(dat$ptsd ~ dat$csa)

##
##  Welch Two Sample t-test
##
## data:  dat$ptsd by dat$csa
## t = 8.9006, df = 63.675, p-value = 8.803e-13
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  5.618873 8.871565
## sample estimates:
##    mean in group Abused mean in group NotAbused
##               11.941093                4.695874

# The null-hypothesis, i.e., both population means are equal,
```

```
# is rejected at 5%. This shows us that there is a statistically
# significant difference in stress-level between the two groups of
# women. However, this analysis is not considering the influence
# of the variable CPA. Having a look at the graph with the two subgroups,
# we see that CPA and CSA are not independent. Thus, for a complete
# analysis we need to do a multiple regression including both predictors.
```

(e) Fit a regression model to the data with both predictors and their interaction. Check the model assumptions using appropriate plots.

```
mod.interact <- lm(ptsd ~ cpa * csa, data = dat)
summary(mod.interact)
```

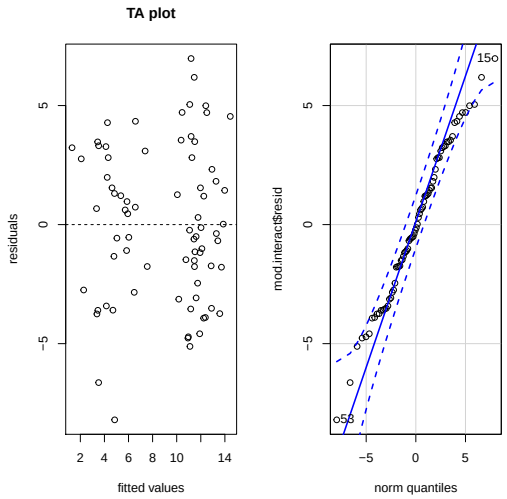
```
##
## Call:
## lm(formula = ptsd ~ cpa * csa, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1999 -2.5313 -0.1807  2.7744  6.9748
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       10.5571     0.8063  13.094  < 2e-16 ***
## 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 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.279 on 72 degrees of freedom
## Multiple R-squared:  0.5828,Adjusted R-squared:  0.5654
## F-statistic: 33.53 on 3 and 72 DF,  p-value: 1.133e-13
```

```
library(car)
```

```
## Loading required package:  carData
```

```
par(mfrow = c(1,2))
# Tukey Ascombe plot
plot(mod.interact$fitted, mod.interact$resid,
```

```
  xlab = "fitted values", ylab = "residuals", main = "TA plot")
abline(h=0,lty=2)
qqPlot(mod.interact$resid, dist = "norm",
  mean = mean(mod.interact$resid), sd = sd(mod.interact$resid))
```



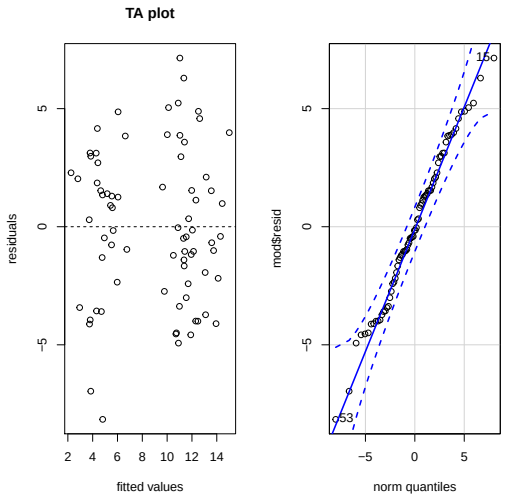

```
## [1] 53 15

# There seems to be no violation of the model assumptions
```

(f) Is it appropriate to simplify the model from the previous task, i.e. are there terms that can be left out? If so, again perform a residual analysis of the simpler model.

```
# The summary showed that the interaction term is non-significant and
# we therefore remove it from the model.
mod <- lm(ptsd ~ cpa + csa, data=dat)
summary(mod)

##
## Call:
## lm(formula = ptsd ~ cpa + csa, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1567 -2.3643 -0.1533  2.1466  7.1417
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    10.2480     0.7187  14.260  < 2e-16 ***
## cpa             0.5506     0.1716   3.209  0.00198 **
## csaNotAbused   -6.2728     0.8219  -7.632 6.91e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.273 on 73 degrees of freedom
## Multiple R-squared:  0.5786,Adjusted R-squared:  0.5671
## F-statistic: 50.12 on 2 and 73 DF,  p-value: 2.002e-14
```

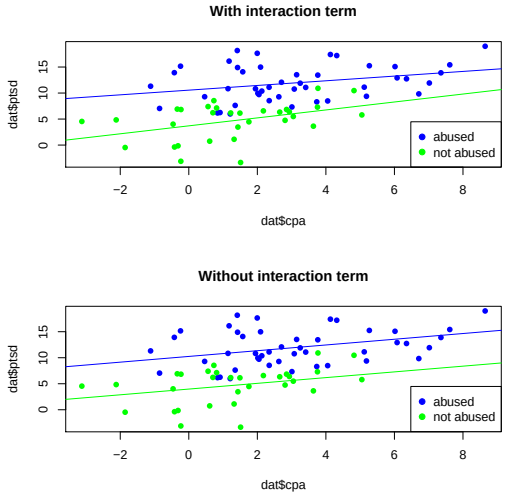
```
# Without the interaction, we can simply interpret the coefficients.
# We see especially, that the PTSD is 6 times lower for not abused
# compared to abused women if the cpa is held constant.
# All predictors are significant, we therefore keep that model as it is.

# We again check the model assumptions
par(mfrow = c(1,2))
plot(mod$fitted, mod$resid,
  xlab = "fitted values", ylab = "residuals", main = "TA plot")
abline(h=0,lty=2)
qqPlot(mod$resid, dist = "norm",
  mean = mean(mod$resid), sd = sd(mod$resid))
```



```
## [1] 53 15

# there are no violations.
```

(g) Draw two plots, one for the model with, one for the model without interaction term. As basis, you can use the plot where you differentiated abused and non abused women by color. Now, draw the regression lines on top of the plots. What's the difference. Do you think the interaction is necessary.

```
par(mfrow = c(2,1))
# model with interaction term
plot(dat$cpa, dat$ptsd, main="With interaction term",
     pch=16, type="n")
points(dat$cpa[dat$csa=="Abused"], dat$ptsd[dat$csa=="Abused"],
       pch = 16, col="blue")
points(dat$cpa[dat$csa=="NotAbused"], dat$ptsd[dat$csa=="NotAbused"],
       pch = 16, col="green")
legend("bottomright", legend=c("abused", "not abused"),
       pch=19, col=c("blue", "green"))
abline(mod.interact$coefficients[1:2], col="blue")
abline(mod.interact$coefficients[1:2]+mod.interact$coefficients[3:4],
       col="green")

# Model without interaction term
plot(dat$cpa, dat$ptsd, main="Without interaction term",
     pch=16, type="n")
points(dat$cpa[dat$csa=="Abused"], dat$ptsd[dat$csa=="Abused"],
       pch = 16, col="blue")
points(dat$cpa[dat$csa=="NotAbused"], dat$ptsd[dat$csa=="NotAbused"],
       pch = 16, col="green")
legend("bottomright", legend=c("abused", "not abused"),
       pch=19, col=c("blue", "green"))
abline(mod$coefficients[1:2], col="blue")
abline(mod$coefficients[1:2]+c(mod$coefficients[3],0), col="green")
```




```
# The interaction term allows us to fit different intercepts and slopes
# for the two groups. Without the interaction, we have different intercepts
# only. It seems to be reasonable to exclude the interaion, the slope is
# only slightly different.
```

Exercise 2

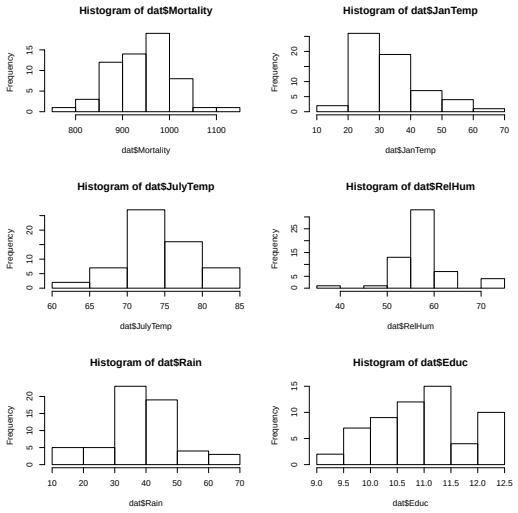
In a study on the contribution of air pollution to mortality, General Motors collected data from 60 US Standard Metropolitan Statistical Areas (SMSAs). The dependent variable is the age adjusted mortality (called `Mortality` in the data set). The data includes variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants. You can download the data from the website and read it with `read.table(...,sep=",", header=TRUE)`

```
# read in the data
dat <- read.table(paste0(dir,"data/mortality.csv"), sep = ",", header = TRUE)
```

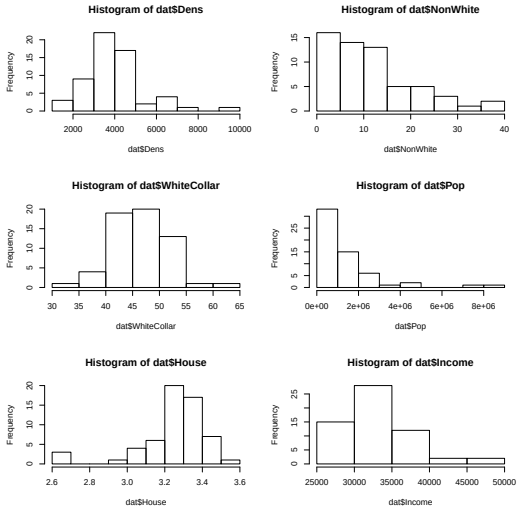
(a) First, set the city names as row names. Then, use histograms to check the distribution of the variables. If necessary, transform them. For right skewed data, use a log-transformation, for percentages, use an arcsin-transformation.

```
# we set the city names as row names
rownames(dat) <- dat$City
dat <- dat[,-1]
```

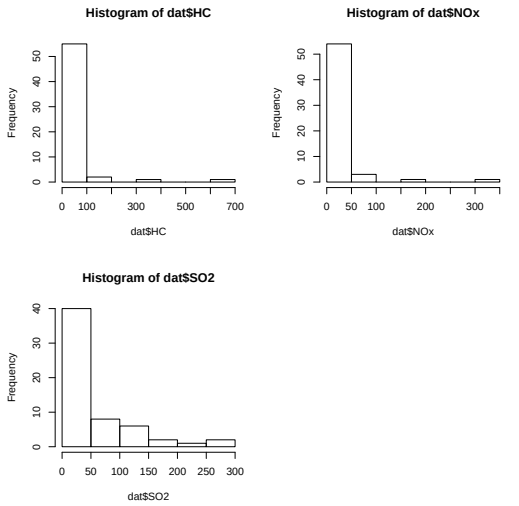
```
par(mfrow=c(3,2))
hist(dat$Mortality)    ## ok, no transformation
hist(dat$JanTemp)      ## right-skewed, log transformation recommendable
hist(dat$JulyTemp)     ## ok, no transformation
hist(dat$RelHum)       ## ok, no transformation
hist(dat$Rain)         ## ok, no transformation
hist(dat$Educ)         ## ok, no transformation
```




```
par(mfrow=c(3,2))
hist(dat$Dens)         ## right skewed, log-tranformation recommendable
hist(dat$NonWhite)     ## percentage, arcsin-transformation recommendable
hist(dat$WhiteCollar)  ## percentage, arcsin-transformation recommendable
hist(dat$Pop)          ## right skewed, log-tranformation recommendable
hist(dat$House)        ## ok, no transformation
hist(dat$Income)       ## right skewed, log-tranformation recommendable
```




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



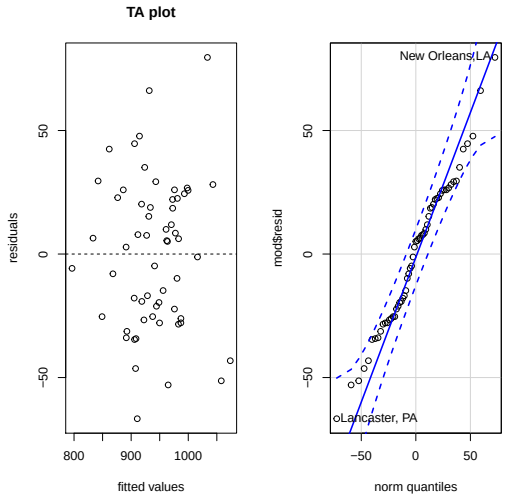

```
# We transform the following variables
dat$JanTemp     <- log(dat$JanTemp)
dat$Dens        <- log(dat$Dens)
dat$NonWhite    <- asin(sqrt(dat$NonWhite/100))
dat$WhiteCollar <- asin(sqrt(dat$WhiteCollar/100))
```

```
dat$Pop         <- log(dat$Pop)
dat$Income      <- log(dat$Income)
dat$HC          <- log(dat$HC)
dat$NOx         <- log(dat$NOx)
dat$SO2         <- log(dat$SO2)
```

(b) Carry out a multiple linear regression containing all variables. Does the model fit well? Check the residuals. (**R-Hint**: Using "." in `lm(... ~ .)` includes all variables into the model.)

```
# fit the model
mod <- lm(Mortality ~ ., data=dat)
summary(mod)

##
## Call:
## lm(formula = Mortality ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.668 -25.338   5.108  22.670  79.594
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1514.05643  592.42867   2.556  0.01413 *
## JanTemp      -65.90878   27.23547  -2.420  0.01972 *
## JulyTemp      -2.18908    2.06935  -1.058  0.29589
## RelHum         0.04771    1.08381   0.044  0.96509
## Rain           1.70646    0.58318   2.926  0.00541 **
## Educ         -12.26491    8.87953  -1.381  0.17417
## Dens          16.05653   16.29979   0.985  0.32997
## NonWhite     321.61186   64.66123   4.974 1.05e-05 ***
## WhiteCollar -154.16478  114.47231  -1.347  0.18496
## Pop            2.34899    7.79886   0.301  0.76468
## House        -28.18972   37.85883  -0.745  0.46047
## Income       -17.90976   48.47305  -0.369  0.71354
## HC           -23.84947   15.27338  -1.562  0.12557
## NOx           34.00128   14.51624   2.342  0.02375 *
## SO2           -1.35604    6.90926  -0.196  0.84531
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.86 on 44 degrees of freedom
## Multiple R-squared:  0.7634,Adjusted R-squared:  0.6881
## F-statistic: 10.14 on 14 and 44 DF,  p-value: 1.373e-09

# Even though most predictors are non significant, the model seems
# to fit the data quite well.

# Check the model assumptions
library(car)
par(mfrow = c(1,2))
plot(mod$fitted, mod$resid,
  main = "TA plot", xlab = "fitted values", ylab = "residuals")
abline(h = 0, lty = 2)
qqPlot(mod$resid, dist = "norm",
  mean = mean(mod$resid), sd = sd(mod$resid))
```



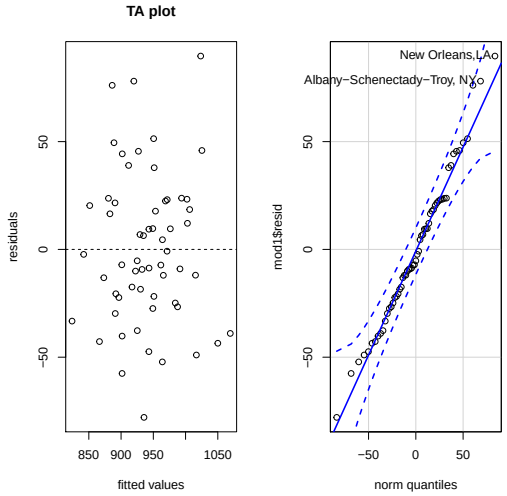

```
## New Orleans,LA  Lancaster, PA
##             36             27

# The model assumptions are not violated.
```

(c) Now take all the non-significant variables out of the model and compute the regression again. Do you think this is a good strategie to simplify the model? Compare your simplified model to the full model using an anova.

```r
# Build a new model based on the significant predictors
mod1 <- lm(Mortality ~ JanTemp + Rain + NonWhite + NOx, data=dat)
summary(mod1)
```

```
##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + NonWhite + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -77.919 -23.592  -5.281  22.011  89.691
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 980.8357    62.7178  15.639  < 2e-16 ***
## JanTemp     -79.8471    18.8162  -4.244 8.70e-05 ***
## Rain          2.5434     0.4822   5.275 2.40e-06 ***
## NonWhite    276.2770    42.5363   6.495 2.72e-08 ***
## NOx          20.9886     4.6856   4.479 3.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36.32 on 54 degrees of freedom
## Multiple R-squared:  0.6847,Adjusted R-squared:  0.6614
## F-statistic: 29.32 on 4 and 54 DF,  p-value: 5.674e-13
```

```r
# Check model assumptions again
par(mfrow = c(1, 2))
plot(mod1$fitted, mod1$resid,
  main = "TA plot", xlab = "fitted values", ylab = "residuals")
abline(h = 0, lty = 2)
qqPlot(mod1$resid, dist = "norm",
  mean = mean(mod1$resid), sd = sd(mod1$resid))
```




```
##               New Orleans,LA Albany-Schenectady-Troy, NY
##                           36                           2
```

```
# All predictors are now highly significant. As expected with fewer
# variables, the residuals are a little bigger now and R^2 decreased
# slightly. However, the difference in adjusted R^2 is very small,
# indicating that we have not lost much explanatory power.

# Even though leaving out all of the non-significant variable at
# once worked quite well here, this is not a good strategy in
# general. If the predictors are not mutually independent, leaving
# out one can have a huge effect on the significance of the others.
# A better way of pruning the model thus is to leave out predictors
# step by step, one at a time.

# We use an anova to compare the models
anova(mod, mod1)
```

```
## 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 + Rain + NonWhite + NOx
##   Res.Df   RSS  Df Sum of Sq      F Pr(>F)
## 1     44 53474
## 2     54 71247 -10    -17773 1.4624  0.186
```

```
# Using ANOVA, the above observation can be corroborated by comparing
# the two models: we test H0 (the smaller, nested model is already good)
# against HA (we need a bigger model with additional predictors) and as
# indicated by a non-significant p-value (e.g. at the 5% level), we
# can not reject H0.
```

(d) Start with the full model. Remove now step by step the variable with the biggest p-value as long as it is over 0.05. Use again an anova to compare the full model to the reduced one. Compare the result to the result of the previous subtask. (**R-Hint**: Use the function `update()`)

```
# We reduce the model as long as we are left with significant predictors
# only .
mod.reduc <- mod
mod.reduc <- update(mod.reduc, ~.-RelHum)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
##     NonWhite + WhiteCollar + Pop + House + Income + HC + NOx +
##     SO2, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.738 -25.325   5.229  22.785  79.521
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1522.5940   553.5340   2.751  0.00854 **
## JanTemp      -66.0256    26.8036  -2.463  0.01766 *
## JulyTemp      -2.2342     1.7771  -1.257  0.21516
## Rain           1.7110     0.5678   3.014  0.00423 **
## Educ         -12.2876     8.7657  -1.402  0.16784
## Dens          16.0014    16.0704   0.996  0.32472
## NonWhite     322.3336    61.8501   5.212 4.53e-06 ***
## WhiteCollar -154.1022   113.1870  -1.361  0.18014
## Pop            2.3599     7.7080   0.306  0.76089
## House        -28.3888    37.1684  -0.764  0.44898
## Income       -18.0148    47.8743  -0.376  0.70847
```

```
## HC           -23.8440    15.1026  -1.579  0.12138
## NOx           34.0558    14.3021   2.381  0.02155 *
## SO2           -1.4567     6.4474  -0.226  0.82228
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.47 on 45 degrees of freedom
## Multiple R-squared:  0.7634,Adjusted R-squared:  0.695
## F-statistic: 11.17 on 13 and 45 DF,  p-value: 3.976e-10

mod.reduc <- update(mod.reduc, ~.-SO2)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
##     NonWhite + WhiteCollar + Pop + House + Income + HC + NOx,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.414 -24.501   3.764  22.349  84.136
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1476.3654   508.9942   2.901  0.00570 **
## JanTemp      -62.6563    22.0407  -2.843  0.00665 **
## JulyTemp      -2.1685     1.7349  -1.250  0.21766
## Rain           1.6932     0.5565   3.043  0.00387 **
## Educ         -11.7713     8.3749  -1.406  0.16658
## Dens          15.3827    15.6712   0.982  0.33143
## NonWhite     319.5287    59.9631   5.329 2.89e-06 ***
## WhiteCollar -155.2406   111.9024  -1.387  0.17204
## Pop            2.1424     7.5683   0.283  0.77839
## House        -26.6033    35.9420  -0.740  0.46296
## Income       -15.4399    46.0158  -0.336  0.73875
## HC           -23.8494    14.9459  -1.596  0.11740
## NOx           32.8564    13.1427   2.500  0.01605 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 34.12 on 46 degrees of freedom
## Multiple R-squared:  0.7631,Adjusted R-squared:  0.7013
## F-statistic: 12.35 on 12 and 46 DF,  p-value: 1.119e-10

mod.reduc <- update(mod.reduc, ~.-Pop)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
##     NonWhite + WhiteCollar + House + Income + HC + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.002 -25.180   3.806  23.184  84.056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1464.677    502.328   2.916  0.00542 **
## JanTemp      -63.036     21.784  -2.894  0.00575 **
## JulyTemp      -2.074      1.686  -1.230  0.22471
## Rain           1.677      0.548   3.060  0.00365 **
## Educ         -11.567      8.262  -1.400  0.16806
## Dens          15.518     15.510   1.000  0.32219
## NonWhite     321.751     58.862   5.466 1.71e-06 ***
## WhiteCollar -154.170    110.739  -1.392  0.17042
## House        -28.564     34.922  -0.818  0.41752
## Income       -11.935     43.883  -0.272  0.78683
## HC           -24.039     14.784  -1.626  0.11063
## NOx           33.618     12.738   2.639  0.01124 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.78 on 47 degrees of freedom
## Multiple R-squared:  0.7627,Adjusted R-squared:  0.7071
## F-statistic: 13.73 on 11 and 47 DF,  p-value: 3.024e-11

mod.reduc <- update(mod.reduc, ~.-Income)
summary(mod.reduc)

##
## Call:
```

```
## lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
##     NonWhite + WhiteCollar + House + HC + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -68.184 -25.120   4.127  22.528  83.274
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1351.8460   280.5051   4.819 1.49e-05 ***
## JanTemp      -63.7347    21.4218  -2.975  0.00457 **
## JulyTemp      -2.0778     1.6695  -1.245  0.21934
## Rain           1.6935     0.5392   3.141  0.00288 **
## Educ         -12.2927     7.7434  -1.588  0.11896
## Dens          15.5653    15.3586   1.013  0.31592
## NonWhite     322.5924    58.2112   5.542 1.25e-06 ***
## WhiteCollar -157.8965   108.8227  -1.451  0.15330
## House        -28.2564    34.5651  -0.817  0.41769
## HC           -23.6377    14.5676  -1.623  0.11122
## NOx           33.0513    12.4445   2.656  0.01070 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.45 on 48 degrees of freedom
## Multiple R-squared:  0.7623,Adjusted R-squared:  0.7128
## F-statistic: 15.39 on 10 and 48 DF,  p-value: 7.686e-12

mod.reduc <- update(mod.reduc, ~.-House)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + JulyTemp + Rain + Educ + Dens +
##     NonWhite + WhiteCollar + HC + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -72.137 -25.144   4.209  24.152  83.480
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) 1176.7896   180.5674   6.517 3.71e-08 ***
## JanTemp      -55.2844    18.6991  -2.957  0.00477 **
## JulyTemp      -1.9777     1.6593  -1.192  0.23906
## Rain           1.7423     0.5341   3.262  0.00202 **
## Educ         -10.4655     7.3886  -1.416  0.16298
## Dens          18.9748    14.7313   1.288  0.20378
## NonWhite     299.6942    50.8559   5.893 3.42e-07 ***
## WhiteCollar -156.1713   108.4334  -1.440  0.15616
## HC           -21.5406    14.2914  -1.507  0.13817
## NOx           31.7474    12.3000   2.581  0.01289 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.34 on 49 degrees of freedom
## Multiple R-squared:  0.759,Adjusted R-squared:  0.7147
## F-statistic: 17.15 on 9 and 49 DF,  p-value: 2.444e-12

mod.reduc <- update(mod.reduc, ~.-JulyTemp)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + Educ + Dens + NonWhite +
##     WhiteCollar + HC + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -74.697 -26.160   0.063  20.863  83.863
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1056.2316   150.2029   7.032 5.35e-09 ***
## JanTemp      -60.2590    18.3038  -3.292  0.00183 **
## Rain           1.7576     0.5361   3.278  0.00190 **
## Educ          -9.3189     7.3565  -1.267  0.21111
## Dens          18.3262    14.7830   1.240  0.22088
## NonWhite     261.7294    39.8105   6.574 2.78e-08 ***
## WhiteCollar -180.9759   106.8639  -1.694  0.09658 .
## HC           -14.3194    12.9978  -1.102  0.27588
## NOx           29.0735    12.1444   2.394  0.02046 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.48 on 50 degrees of freedom
## Multiple R-squared:  0.752,Adjusted R-squared:  0.7123
## F-statistic: 18.95 on 8 and 50 DF,  p-value: 1.05e-12
```

```
mod.reduc <- update(mod.reduc, ~.-HC)
summary(mod.reduc)
```

```
##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + Educ + Dens + NonWhite +
##     WhiteCollar + NOx, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -76.495 -25.543   4.253  19.846  84.672
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1067.5033   150.1677   7.109 3.66e-09 ***
## JanTemp      -64.0371    18.0173  -3.554 0.000828 ***
## Rain           1.8825     0.5251   3.585 0.000754 ***
## Educ         -11.1702     7.1770  -1.556 0.125799
## Dens          18.7825    14.8081   1.268 0.210418
## NonWhite     264.7197    39.8010   6.651 1.94e-08 ***
## WhiteCollar -179.4981   107.0791  -1.676 0.099797 .
## NOx           16.8616     4.9716   3.392 0.001350 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.55 on 51 degrees of freedom
## Multiple R-squared:  0.746,Adjusted R-squared:  0.7111
## F-statistic:  21.4 on 7 and 51 DF,  p-value: 3.851e-13
```

```
mod.reduc <- update(mod.reduc, ~.-Dens)
summary(mod.reduc)
```

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

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -80.854 -26.449   3.159  18.654  84.961
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1217.1646    93.4291  13.028  < 2e-16 ***
## JanTemp      -66.8959    17.9801  -3.721 0.000489 ***
## Rain           1.9731     0.5233   3.771 0.000418 ***
## Educ         -13.1443     7.0471  -1.865 0.067797 .
## NonWhite     261.3019    39.9414   6.542 2.66e-08 ***
## WhiteCollar -142.8799   103.7157  -1.378 0.174224
## NOx           19.5735     4.5146   4.336 6.69e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 33.74 on 52 degrees of freedom
## Multiple R-squared:  0.738,Adjusted R-squared:  0.7078
## F-statistic: 24.41 on 6 and 52 DF,  p-value: 1.59e-13

mod.reduc <- update(mod.reduc, ~.-WhiteCollar)
summary(mod.reduc)

##
## Call:
## lm(formula = Mortality ~ JanTemp + Rain + Educ + NonWhite + NOx,
##     data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -82.794 -25.435   6.366  20.410  77.977
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1183.4856    90.9344  13.015  < 2e-16 ***
## JanTemp      -70.9168    17.8912  -3.964 0.000222 ***
## Rain           1.8185     0.5154   3.528 0.000874 ***
## Educ         -17.9858     6.1597  -2.920 0.005131 **
## NonWhite     268.4084    39.9410   6.720 1.27e-08 ***
## NOx           18.4360     4.4759   4.119 0.000134 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 34.03 on 53 degrees of freedom
## Multiple R-squared:  0.7284,Adjusted R-squared:  0.7028
## F-statistic: 28.43 on 5 and 53 DF,  p-value: 6.945e-14
```

```
# We stop here, we see that in the previous model mod1,
# we missed one significant variable.

# We use an anova to compare the reduced model to the full one:
anova(mod, mod.reduc)
```

```
## 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 + Rain + Educ + NonWhite + NOx
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1     44 53474
## 2     53 61374 -9   -7899.6 0.7222 0.6859
```

```
# We see, that the p-value is higher 0.05, i.e. we don't reject H0,
# that the nested model is already good enough. As could be expected,
# the p-value of this test is even larger (when compared to the anova in
# the previous task), indicating an even reduced need for the full model.
```