Mandatory Assignment 1 - STK9900

Inger Grünbeck

2023-03-16

The assignment is performed using R markdown in RStudio. For each exercise you will find a method and discussion section. The method section contains a quick description of what was coded, the code itself, and the output and plots. The discussion contains the interpretation of the results.

Importing libraries:

```
include = FALSE
message = FALSE
warning = FALSE

library(rcompanion)
library(car)
```

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

```
library(data.table)
library(ggplot2)
library(MASS)
```

Exercise 1

Importing the pollution dataset:

```
pollution=read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/no2.txt", header=TRUE)
```

a)

Method:

The mean, median, standard deviation (sd), and the IQR are inspectecd in order to analyse variable log.no2. Further a histogram and boxplot of the variable are printed.

```
summary(pollution$log.no2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.224   3.214   3.848   3.698   4.217   6.395
```

The standard deviation:

```
sd(pollution$log.no2)
```

```
## [1] 0.7505966
```

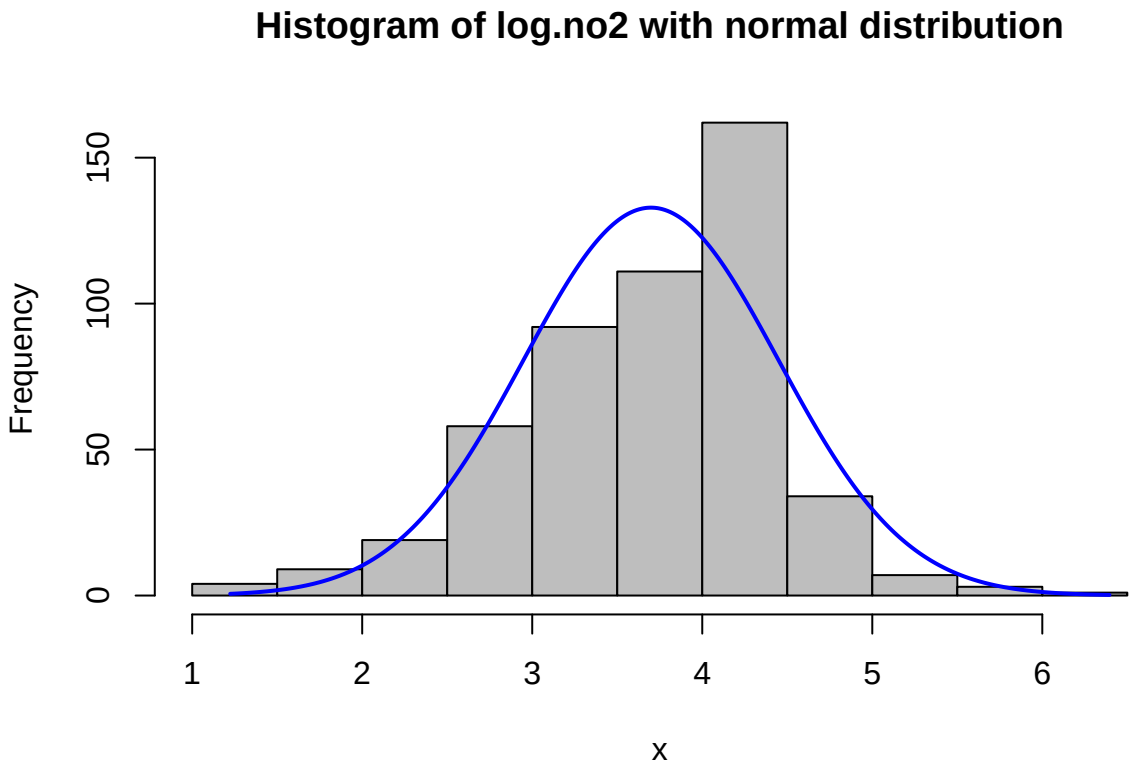
The IQR:

```
IQR(pollution$log.no2)
```

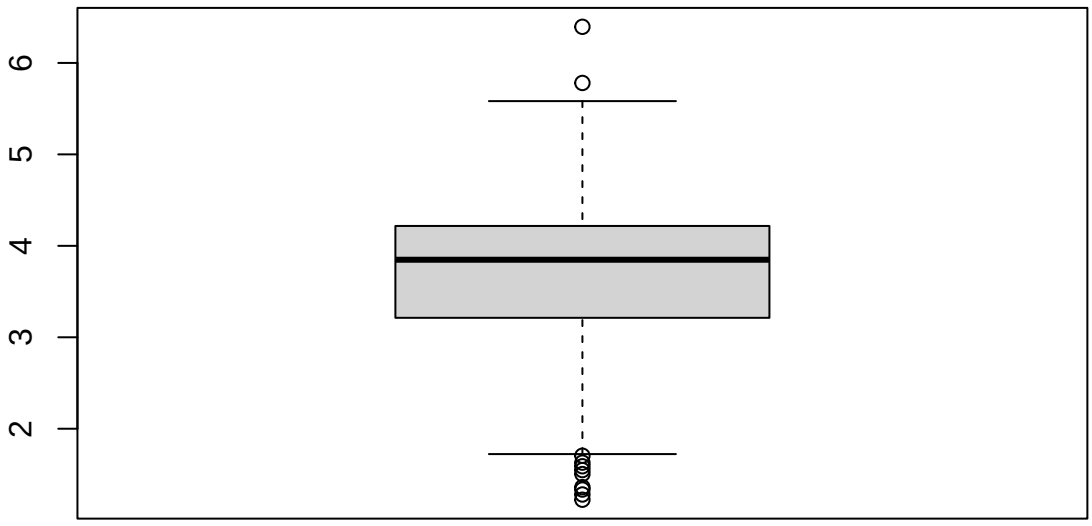
```
## [1] 1.003067
```

The descriptive plots:

```
plotNormalHistogram(pollution$log.no2, main="Histogram of log.no2 with normal distribution")
```




```
boxplot(pollution$log.no2)
```



We repeat this for the variable pollution$log.cars:

```
summary(pollution$log.cars)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.127   6.176   7.425   6.973   7.793   8.349
```

The sd:

```
sd(pollution$log.cars)
```

```
## [1] 1.087166
```

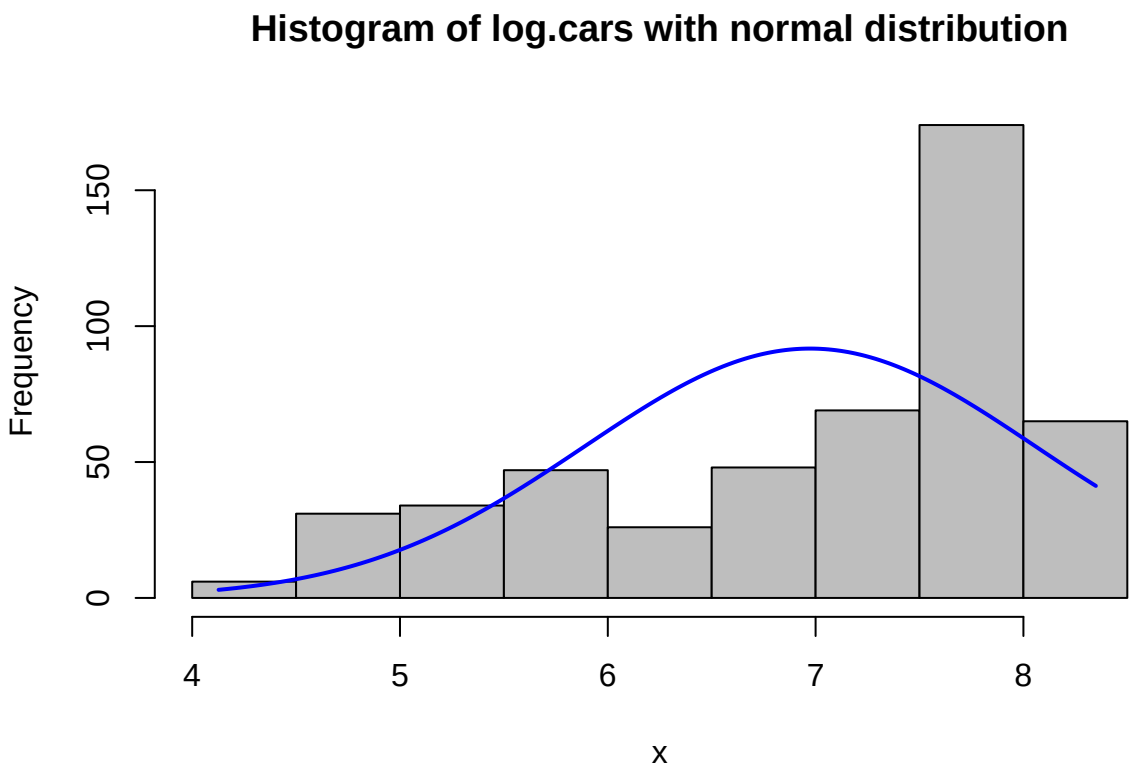
The IQR:

```
IQR(pollution$log.cars)
```

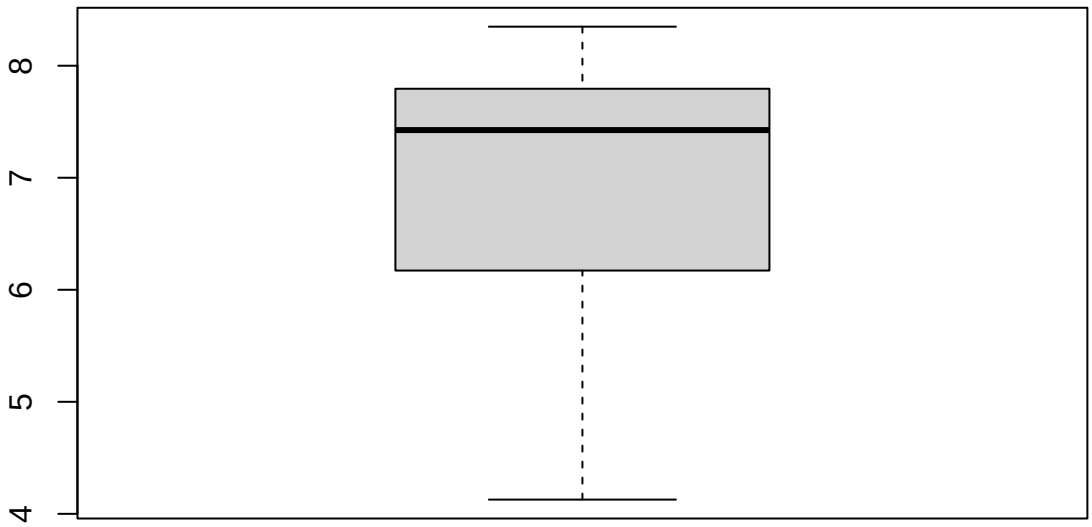
```
## [1] 1.617332
```

The descriptive plots:

```
plotNormalHistogram(pollution$log.cars, main="Histogram of log.cars with normal distribution")
```

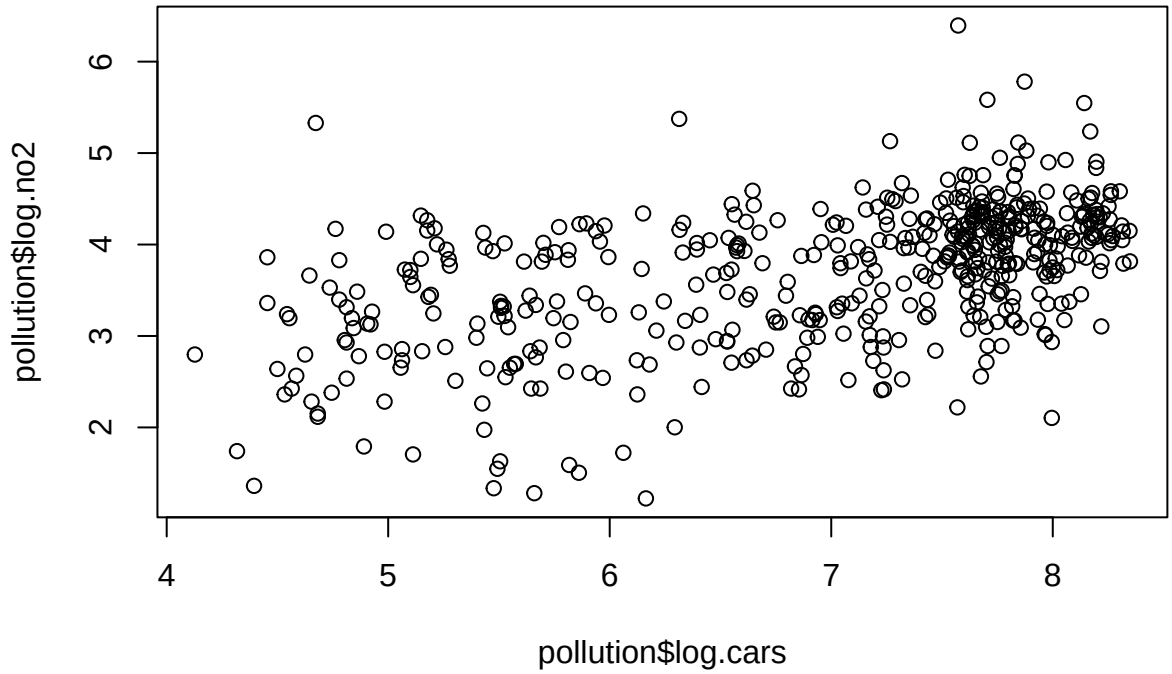



```
boxplot(pollution$log.cars)
```



Finally, a scatterplot of the two variables is plotted in order to examine the correlation between the pollution level and the number of cars per hour. Also, the correlation of the variables is calculated:

```
plot(pollution$log.cars, pollution$log.no2)
```



The correlation between no2 and cars:

```
cor(pollution$log.cars, pollution$log.no2)
```

```
## [1] 0.5120504
```

Discussion:

From the descriptive statistics for variable log.no2 we can see that the mean and median are close to similar, indicating that the observations could be close to normally distributed. We can also see that the 1. and 3. quartile are not to far removed from the median, while the min and max values are noticeably further away/ noticeably smaller and larger than the quartiles. This could also indicate some outliers in the observations.

From the histogram we can see that the suspicion was correct, and the observations are close to normal distributed based on its bell shaped curve. The blue line shows how the observations should behave to be normal distributed. The observations do not deviate to much from this template. Further, the boxplot confirms the previous statements. It looks close to symmetrical, indicating a normal distribution. We can also see some outliers, as expected.

For the log.cars variable, we can see that the median and mean value are more distant from each other, indicating that the observations do not fully follow a normal distribution. We can also see that both the sd and IQR are larger for log.cars than for log.no2. This indicates a wider bell shape/histogram. As the difference between the min value and the 1. quartile is larger than the difference between the max value and the 3. quartile, the boxplot will also not be as symmetrical as for the log.no2 variable.

The histogram and boxplot confirm all of the above assumptions. The observation are right-skewed (heavier to the high values). There are no outliers to see ion the boxplot, but both the histogram as well as the boxplot show that the observations do not follow a normal distribution.

According to the scatterplot for the two variables, the observations tend to follow a line that is slightly tilted up to the right. This indicates a positive correlation between the variables. The correlation coefficient confirms this, as it is relative large (0.5) and positive.

b)

Method:

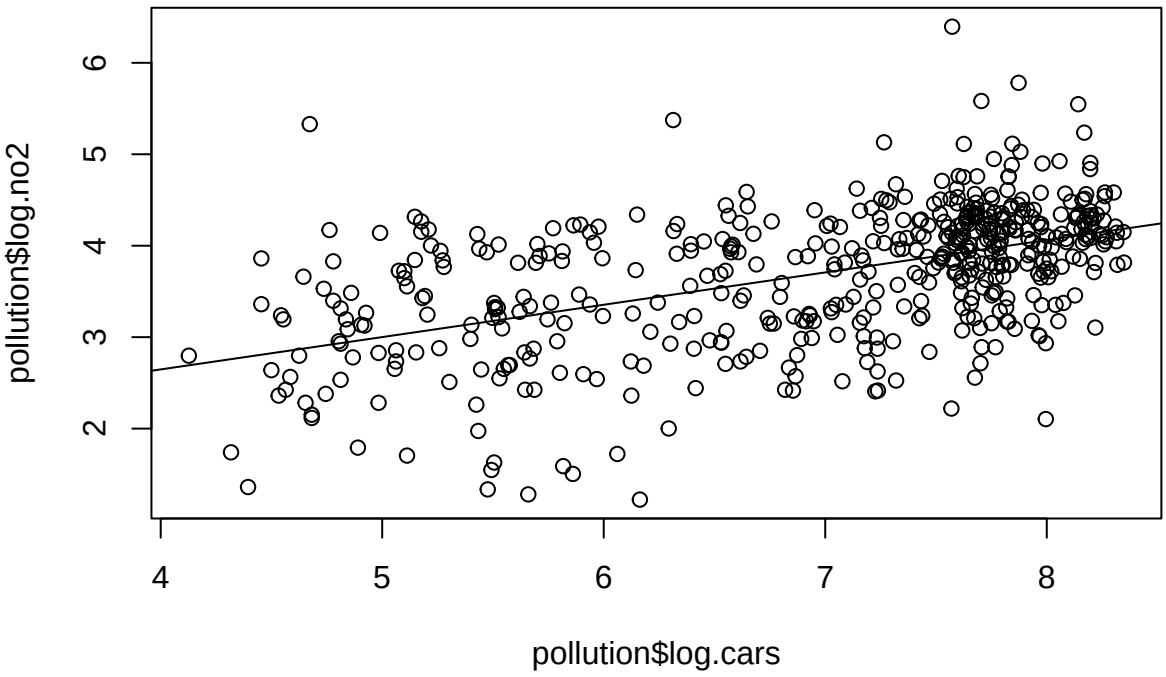
A linear model is fitted, with the NO2 levels as dependent and cars/hour as dependent variable:

```
pollution.fit = lm(log.no2~log.cars, data=pollution)
summary(pollution.fit)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18822 -0.40071  0.06428  0.40362  2.48472
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.23310    0.18755   6.575 1.23e-10 ***
## log.cars     0.35353    0.02657  13.303  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6454 on 498 degrees of freedom
## Multiple R-squared:  0.2622, Adjusted R-squared:  0.2607
## F-statistic:   177 on 1 and 498 DF,  p-value: < 2.2e-16
```

The observations are plotted together with a line representing the fitted linear regression model:

```
plot(pollution$log.cars, pollution$log.no2)
abline(pollution.fit)
```



Discussion:

According to the models' estimated intercept, the basic leve of the log concentration of no2 (when no traffic is present) corresponds to 1.233. When the traffic (log of cars per hour) increases by one, the log of the no2 concentration will increase by 0.3535 units. This can be seen based on the estimated parameter for log.cars. This also confirms what we saw in the scatterplot, that there is a positive correlation between the two variables.

R^2, the coefficient of determination, tells us how much of the of the variance present in the dependent variable can be explained by the model/the independent variables. A high R^2 value indicates that a high amount of the dependent variable's variance can be explained by the model, and therefore is a good representation. In this case the R^2 = 0.266, indicating that only a low emount of variance is explained at that the model has improvement potential.

From the plotted scatterplot together with the fitted regression line, we can see that the line fits the observations. So the model at least represents the relationship between the dependent and independent variable. But it would be interesting to see how the second degree term of log.cars would relate to the log.no2 variable. Maybe it would fit better?

c)

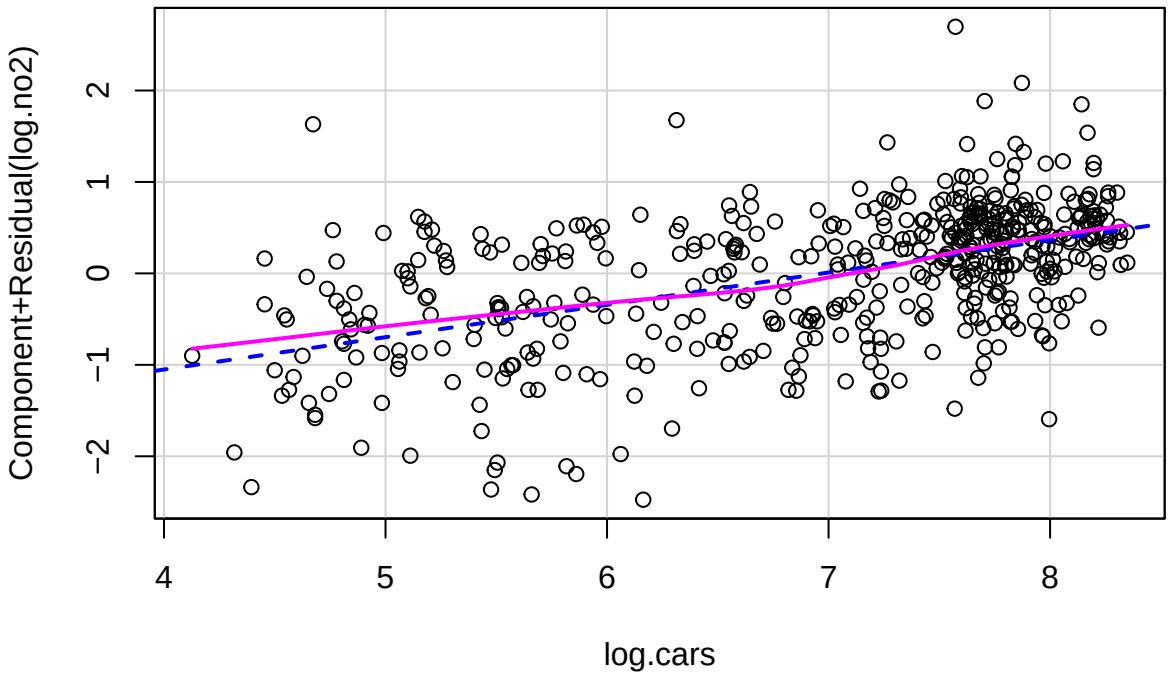
For a linear regression we assume:

1. linearity between the dependent and independent variable
2. Homoscedasticity - constant variance in the residuals
3. Normally distributed residuals
4. Uncorrelated errors (This assumptions I have not testes for, as this not has been part of the curriculum yet)

Method:

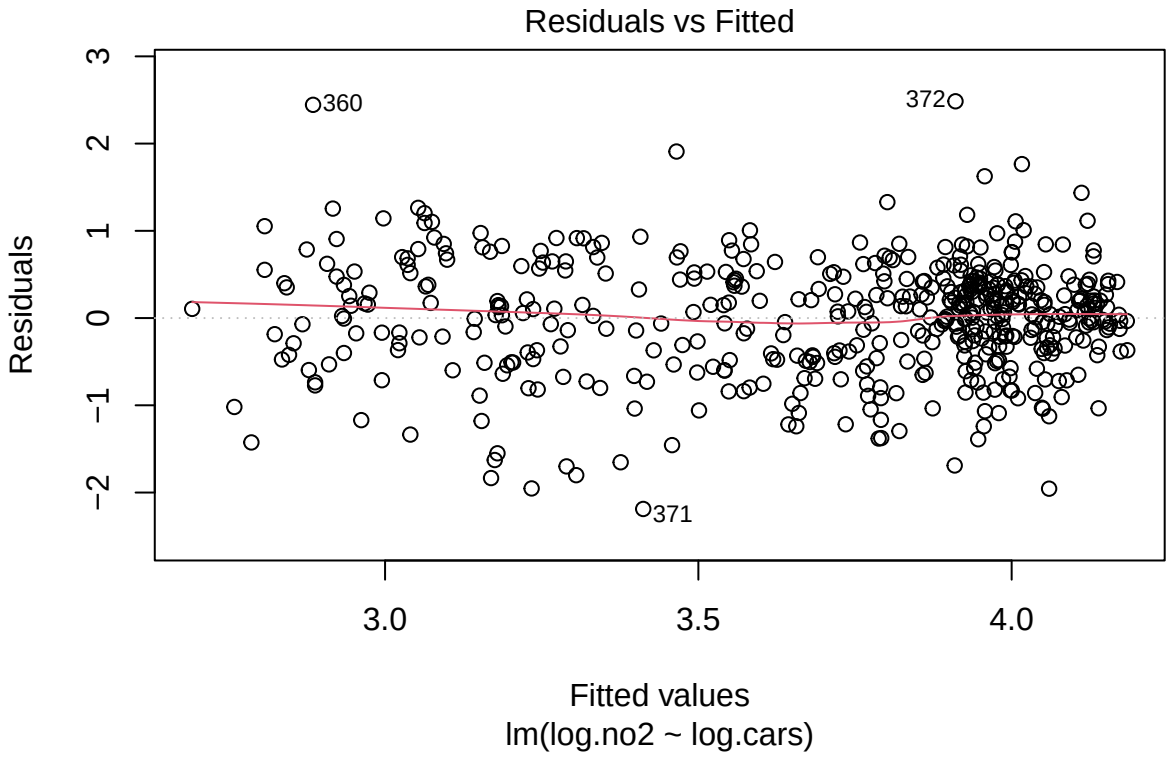
For the first assumption, a CPR (component-plus-residual) plot was plotted:

```
crPlots(pollution.fit)
```

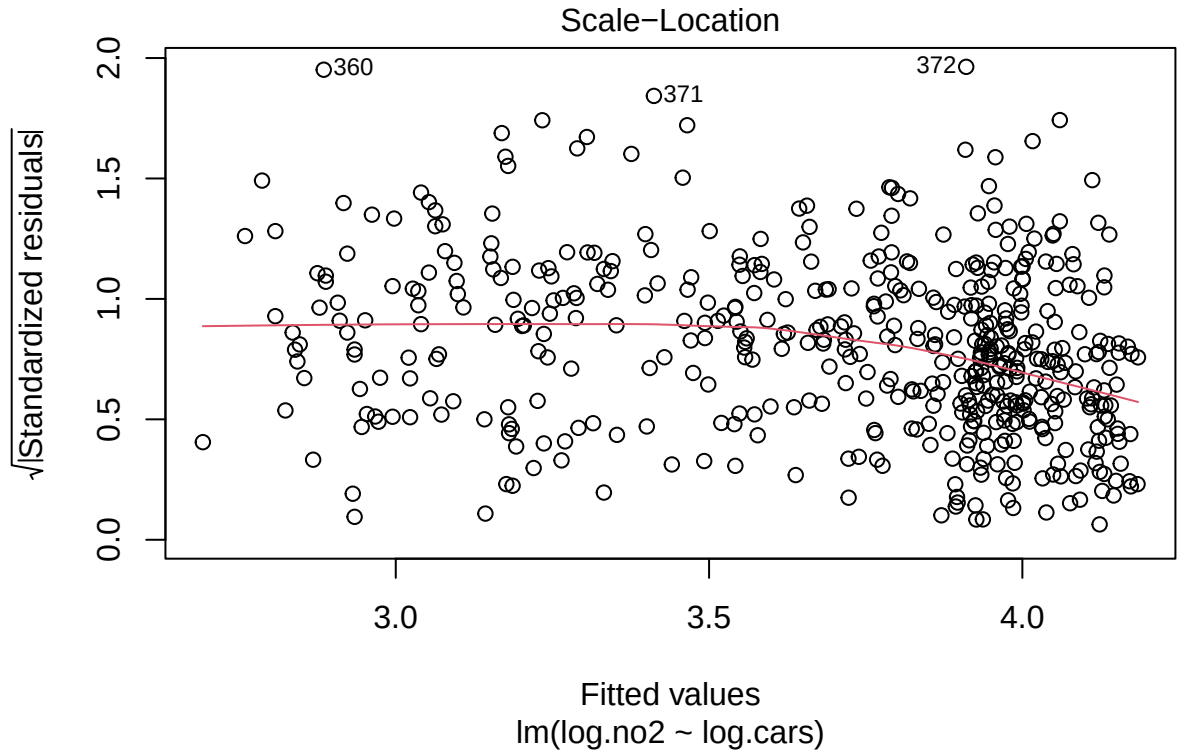


For the homoscedasticity check, the residuals vs fitted values are plotted. If there is no systematic pattern in the residuals, the assumption is fulfilled.

```
plot(pollution.fit, 1)
```

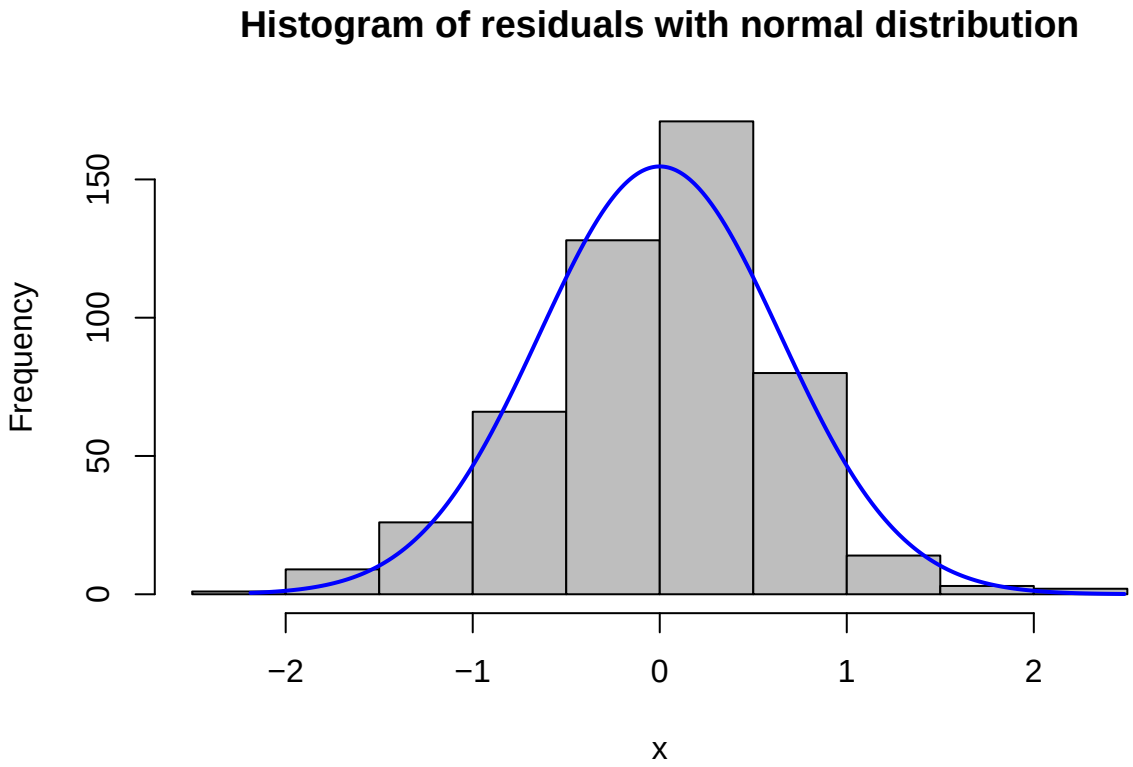



```
plot(pollution.fit, 3)
```

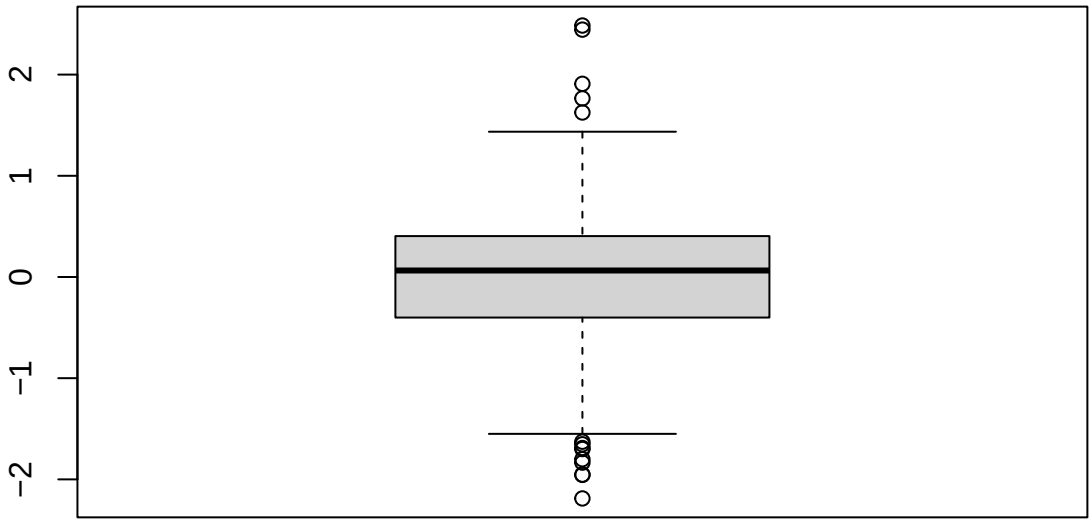


For the linearity assumption, one can use histograms, boxplots and Q-Q plots:

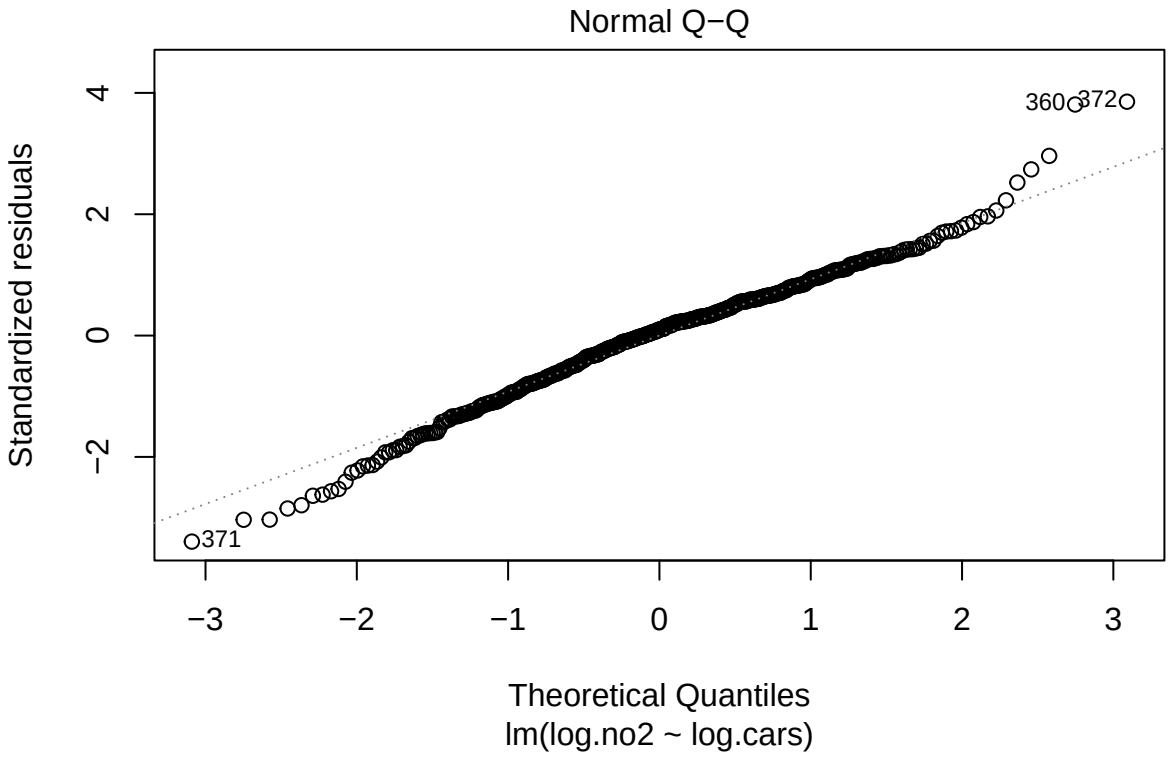
```
plotNormalHistogram(pollution.fit$residuals, main="Histogram of residuals with normal distribution")
```




```
boxplot(pollution.fit$residuals)
```




```
plot(pollution.fit, 2)
```



Discussion:

The CPR plot that is used to check the model for linearity between the depndent ans independent variable shows that the model may be appropriate for the observations present, and that the linearity assumption is reasonable. This can be seen as the pink line follows the template line close. It might be that a second degree polynomial would fit better, as the pink line is slightly curved, but this is not sure.

From the residual vs. fitted values pot, we can see that the observations seem symmetrically distributed along the template line. They also do not stray to far away from the template line in the y-direction. All over, the plot looks good. In the standardized residual plot we can see that the red line/the variance decreases in the end of the end of the plot. This might indicate that the predicted/fitted values are predicted to high. This should maybe be looked at closer. But all in all, this plot also looks ok. I would therefore comclude that the second assumption of homoscedasticity is also fulfilled.

The histogram of the residuals follows a bell shape, the boxplot looks symmetrical, ans the qq-plot follows mostly the template line. I therefore conclude that also the third assumption of normally distributed residuals is fulfilled.

d)

Method:

I'm applying the backwards-method in order to exclude predictors in possible multiple models. I start by including all variables and exclude one and one based on their significance/p-value. The models are being evaluated based on the Rˆ2. I test both applying a log-transformation to the variables, as well as including their the second degree terms.

Model 1: -> all variables

```
pollution.fit.1=lm(log.no2~log.cars+temp+wind.speed+hour.of.day, data=pollution)
summary(pollution.fit.1)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + wind.speed + hour.of.day,
##     data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.24876 -0.32070  0.03084  0.33860  1.96057
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.152131   0.175045   6.582 1.19e-10 ***
## log.cars     0.456974   0.028411  16.084  < 2e-16 ***
## temp        -0.026855   0.003905  -6.877 1.85e-11 ***
## wind.speed  -0.149334   0.014076 -10.609  < 2e-16 ***
## hour.of.day -0.013025   0.004452  -2.926   0.0036 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5508 on 495 degrees of freedom
## Multiple R-squared:  0.4658, Adjusted R-squared:  0.4615
## F-statistic: 107.9 on 4 and 495 DF,  p-value: < 2.2e-16
```

Model 3: -> remove hour.of.day

```
pollution.fit.3=lm(log.no2~log.cars+temp+wind.speed, data=pollution)
summary(pollution.fit.3)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + wind.speed, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.13980 -0.33142  0.04882  0.35257  1.97666
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.316491   0.167043   7.881 2.07e-14 ***
## log.cars     0.409026   0.023384  17.492  < 2e-16 ***
## temp        -0.026447   0.003932  -6.725 4.83e-11 ***
## wind.speed  -0.146594   0.014152 -10.359  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.555 on 496 degrees of freedom
## Multiple R-squared:  0.4566, Adjusted R-squared:  0.4533
## F-statistic: 138.9 on 3 and 496 DF,  p-value: < 2.2e-16
```

Model 4: -> remove temp

```
pollution.fit.4=lm(log.no2~log.cars+wind.speed, data=pollution)
summary(pollution.fit.4)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + wind.speed, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.03158 -0.34600  0.01792  0.39106  2.10633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.54518    0.17067   9.054   <2e-16 ***
## log.cars     0.37928    0.02396  15.828   <2e-16 ***
## wind.speed  -0.16088    0.01460 -11.019   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5791 on 497 degrees of freedom
## Multiple R-squared:  0.407,  Adjusted R-squared:  0.4047
## F-statistic: 170.6 on 2 and 497 DF,  p-value: < 2.2e-16
```

In order to see if a log transformation might affect the relationship between the dependent and independent variables, the not transformed variables are also transformed. Temperature is not transformed, as it has negative values.

Model 5: -> all variables

```
pollution.fit.5=lm(log.no2~log.cars+temp+log(wind.speed)+log(hour.of.day), data=pollution)
summary(pollution.fit.5)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + log(wind.speed) + log(hour.of.day),
##     data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.26406 -0.31784  0.04366  0.34532  1.83379
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.111953   0.169410   6.564 1.33e-10 ***
## log.cars          0.461040   0.031041  14.852  < 2e-16 ***
## temp             -0.026922   0.003853  -6.988 9.07e-12 ***
## log(wind.speed)  -0.415333   0.036410 -11.407  < 2e-16 ***
## log(hour.of.day) -0.098007   0.041880  -2.340   0.0197 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5431 on 495 degrees of freedom
## Multiple R-squared:  0.4807, Adjusted R-squared:  0.4766
## F-statistic: 114.6 on 4 and 495 DF,  p-value: < 2.2e-16
```

Model 6: -> remove log(hour.of.day)

```
pollution.fit.6=lm(log.no2~log.cars+temp+log(wind.speed), data=pollution)
summary(pollution.fit.6)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + log(wind.speed), data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07759 -0.33892  0.05458  0.36666  1.83266
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.229009   0.162586   7.559 1.98e-13 ***
## log.cars         0.411979   0.022995  17.916  < 2e-16 ***
## temp            -0.026304   0.003861  -6.813 2.79e-11 ***
## log(wind.speed) -0.414496   0.036572 -11.334  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5455 on 496 degrees of freedom
## Multiple R-squared:  0.475,  Adjusted R-squared:  0.4718
## F-statistic: 149.6 on 3 and 496 DF,  p-value: < 2.2e-16
```

As model 6 has a lower R^2 than model 5 and all variables in model 6 are marked as significant with a 99% significance level, I don't continue excluding predictors. Instead I test if including the second degree term of the variables instead of log transforming them will improve the R^2.

Model 7: -> all variables + second degree terms

```
pollution.fit.7=lm(log.no2~log.cars+I(log.cars^2)+temp+I(temp^2)+wind.speed+I(wind.speed^2)+hour.of.day
summary(pollution.fit.7)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + I(log.cars^2) + temp + I(temp^2) +
##     wind.speed + I(wind.speed^2) + hour.of.day + I(hour.of.day^2),
##     data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.07791 -0.32530  0.02189  0.36192  1.85442
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       4.7410924  1.0699553   4.431 1.16e-05 ***
## log.cars         -0.5744937  0.3322757  -1.729  0.08444 .
## I(log.cars^2)     0.0799072  0.0253074   3.157  0.00169 **
## temp             -0.0281999  0.0039124  -7.208 2.16e-12 ***
## I(temp^2)         0.0002639  0.0003775   0.699  0.48480
## wind.speed       -0.3749046  0.0436121  -8.596  < 2e-16 ***
## I(wind.speed^2)   0.0297334  0.0054763   5.430 8.90e-08 ***
## hour.of.day      -0.0329595  0.0244929  -1.346  0.17903
```

```
## I(hour.of.day^2)  0.0008904  0.0008744   1.018  0.30899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5331 on 491 degrees of freedom
## Multiple R-squared:  0.5036, Adjusted R-squared:  0.4955
## F-statistic: 62.26 on 8 and 491 DF,  p-value: < 2.2e-16
```

Model 8: -> remove hour.of.day^2

```
pollution.fit.8=lm(log.no2~log.cars+I(log.cars^2)+temp+I(temp^2)+wind.speed+I(wind.speed^2)+hour.of.day
summary(pollution.fit.8)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + I(log.cars^2) + temp + I(temp^2) +
##     wind.speed + I(wind.speed^2) + hour.of.day, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00476 -0.33674  0.02924  0.35946  1.82597
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.6066558  1.0618207   4.338 1.74e-05 ***
## log.cars        -0.5365889  0.3301970  -1.625  0.10479
## I(log.cars^2)    0.0748395  0.0248142   3.016  0.00269 **
## temp            -0.0278030  0.0038931  -7.142 3.34e-12 ***
## I(temp^2)        0.0002075  0.0003734   0.556  0.57869
## wind.speed      -0.3734395  0.0435900  -8.567  < 2e-16 ***
## I(wind.speed^2)  0.0292781  0.0054582   5.364 1.25e-07 ***
## hour.of.day     -0.0084495  0.0045468  -1.858  0.06372 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5332 on 492 degrees of freedom
## Multiple R-squared:  0.5025, Adjusted R-squared:  0.4954
## F-statistic:    71 on 7 and 492 DF,  p-value: < 2.2e-16
```

Model 9: -> remove temp^2

```
pollution.fit.9=lm(log.no2~log.cars+I(log.cars^2)+temp+wind.speed+I(wind.speed^2)+hour.of.day, data=poll
summary(pollution.fit.9)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + I(log.cars^2) + temp + wind.speed +
##     I(wind.speed^2) + hour.of.day, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.00854 -0.33453  0.02454  0.35820  1.82492
##
## Coefficients:
```

```
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.620939   1.060765   4.356 1.61e-05 ***
## log.cars        -0.539959   0.329910  -1.637  0.10233
## I(log.cars^2)    0.075157   0.024790   3.032  0.00256 **
## temp            -0.027290   0.003779  -7.221 1.97e-12 ***
## wind.speed      -0.371854   0.043466  -8.555  < 2e-16 ***
## I(wind.speed^2)  0.029067   0.005441   5.342 1.41e-07 ***
## hour.of.day     -0.008471   0.004543  -1.864  0.06286 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5328 on 493 degrees of freedom
## Multiple R-squared:  0.5022, Adjusted R-squared:  0.4961
## F-statistic: 82.89 on 6 and 493 DF,  p-value: < 2.2e-16
```

Model 10: -> remove log.cars

```
pollution.fit.10=lm(log.no2~I(log.cars^2)+temp+wind.speed+I(wind.speed^2)+hour.of.day, data=pollution)
summary(pollution.fit.10)
```

```
##
## Call:
## lm(formula = log.no2 ~ I(log.cars^2) + temp + wind.speed + I(wind.speed^2) +
##     hour.of.day, data = pollution)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.08213 -0.32288  0.02327  0.34721  1.81603
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.894267   0.110817  26.117  < 2e-16 ***
## I(log.cars^2)    0.034724   0.002069  16.786  < 2e-16 ***
## temp            -0.027190   0.003785  -7.183 2.52e-12 ***
## wind.speed      -0.367001   0.043438  -8.449 3.32e-16 ***
## I(wind.speed^2)  0.028565   0.005442   5.249 2.27e-07 ***
## hour.of.day     -0.011171   0.004240  -2.634  0.00869 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5337 on 494 degrees of freedom
## Multiple R-squared:  0.4995, Adjusted R-squared:  0.4944
## F-statistic:  98.6 on 5 and 494 DF,  p-value: < 2.2e-16
```

Model 11: -> remove hour.of.day

```
pollution.fit.11=lm(log.no2~I(log.cars^2)+temp+wind.speed+I(wind.speed^2), data=pollution)
summary(pollution.fit.11)
```

```
##
## Call:
## lm(formula = log.no2 ~ I(log.cars^2) + temp + wind.speed + I(wind.speed^2),
##     data = pollution)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.99114 -0.33111  0.04251  0.36930  1.82765
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       2.900016   0.111459  26.019  < 2e-16 ***
## I(log.cars^2)     0.031692   0.001729  18.327  < 2e-16 ***
## temp             -0.026870   0.003806  -7.060 5.66e-12 ***
## wind.speed       -0.365108   0.043692  -8.356 6.57e-16 ***
## I(wind.speed^2)   0.028634   0.005474   5.231 2.50e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5369 on 495 degrees of freedom
## Multiple R-squared:  0.4925, Adjusted R-squared:  0.4884
## F-statistic: 120.1 on 4 and 495 DF,  p-value: < 2.2e-16
```

I finish excluding predictors as the R^2 keeps decreasing, and all variables in model 6 are marked as significant with a 99% significance level.

Discussion:

Based on the models, log-transforming the variables will slightly improve the R^2 of the best non-transformed model (model 1: 0.4658 vs. model 5: 0.4807). Both models included all variables, but we can see that the models' R^2 does not decrease a lot when removing hour of day. Model 7 includes the second degree term of all variables, and results in a R^2=0.5036, improving the amount of explained variance relative to model 5. But in model 11 we can see that we can remove some of the variables in model 7 and still achieve a relative high R^2 value, 0.4925. I choose to continue with model 11, as this model is less complex/includes less variables than model 7 as well as their R^2 values are similar.

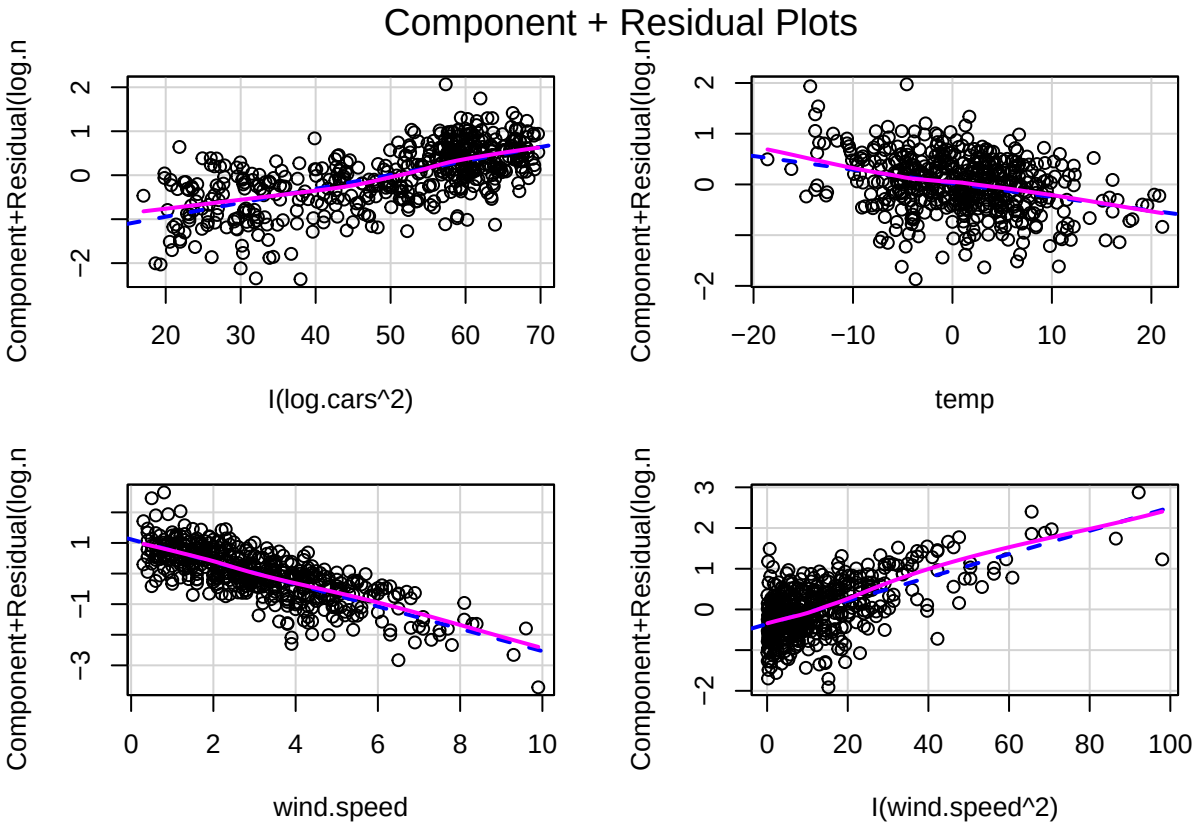
e)

Method:

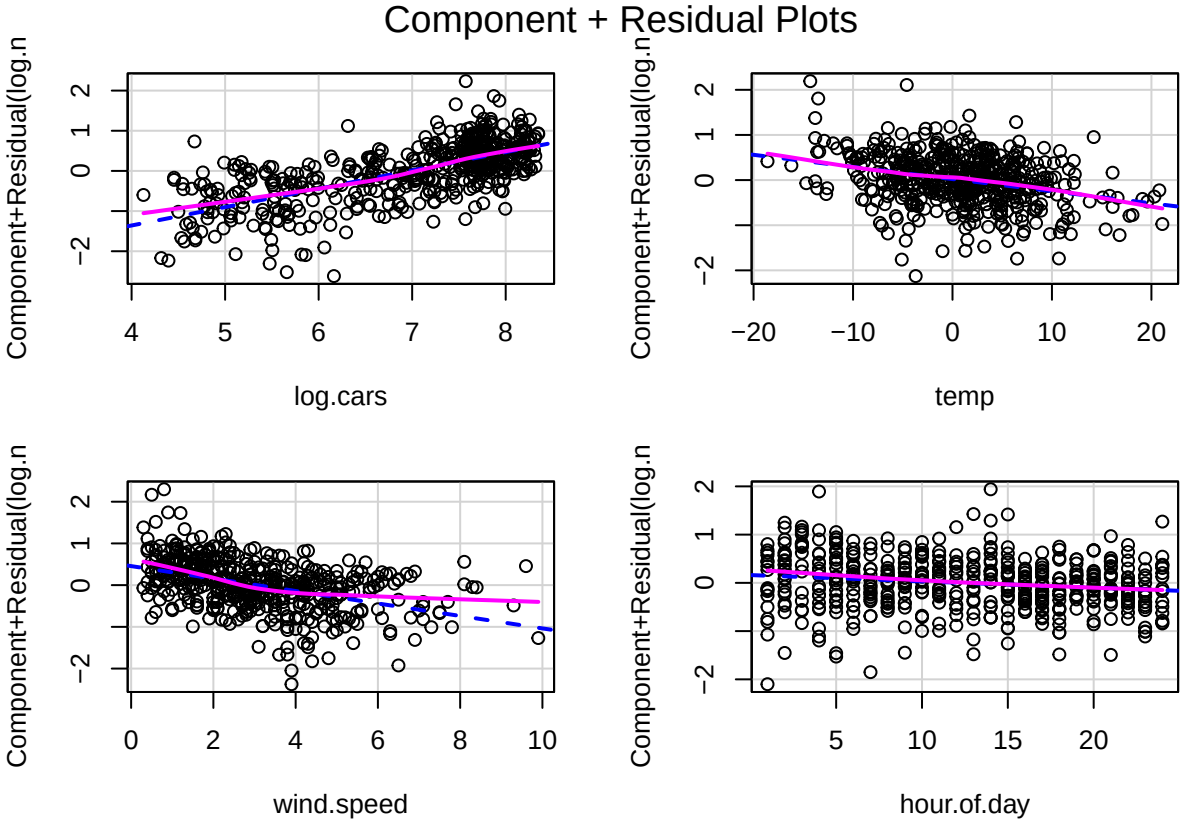
In order to check the assumptions for model 11, I use the same methods as in Excersice 1c) -> plotting CPR plots for the independent variables to check for linear relationships between dependent and independent variables, plotting the residuals vs fitted values to check for constant variance and plotting the histogram, boxplot and qq-plot of the residuals to examine whether they are normally distributed.

The CPR plots follows below. I have also included model 1's CPR plots to compare the effect of including the second degree terms. Also model 9's plot is included to compare how the linearity for log.cars behave when both log.cars and log.cars^2 are included.

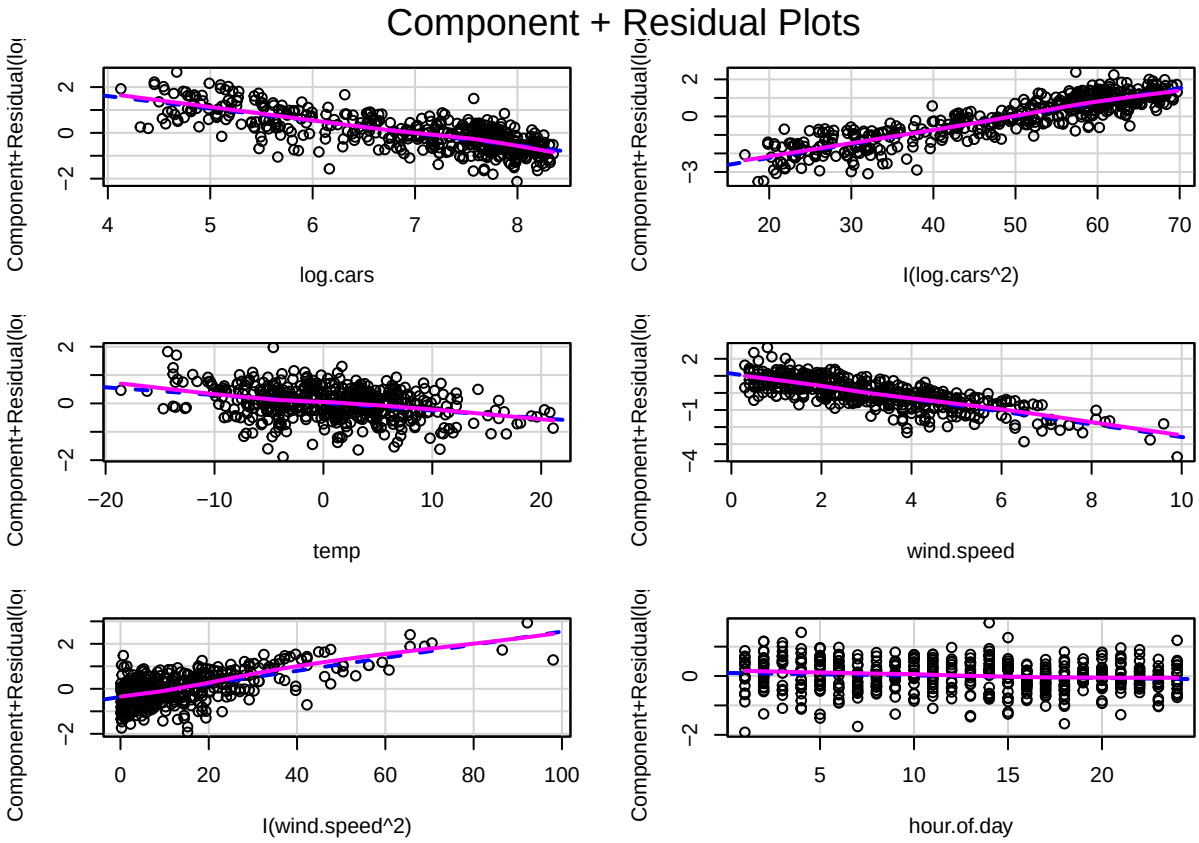
```
crPlots(pollution.fit.11)
```




```
crPlots(pollution.fit.1)
```

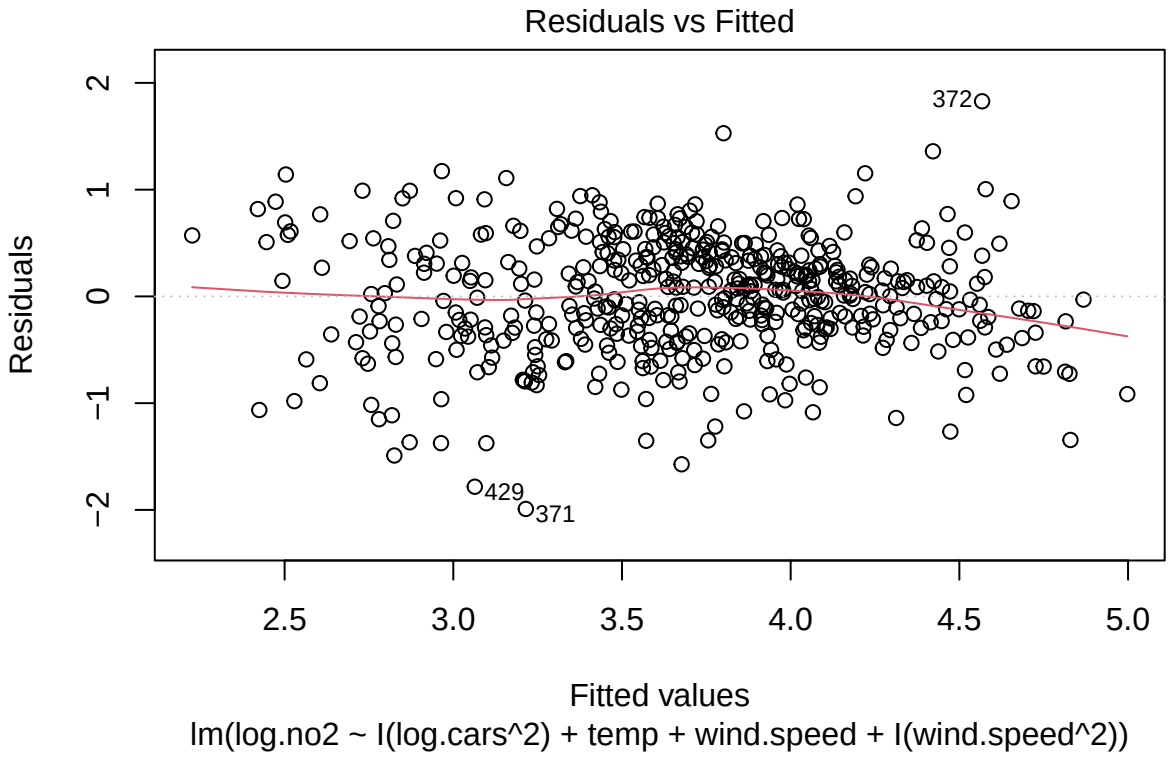


```
crPlots(pollution.fit.9)
```

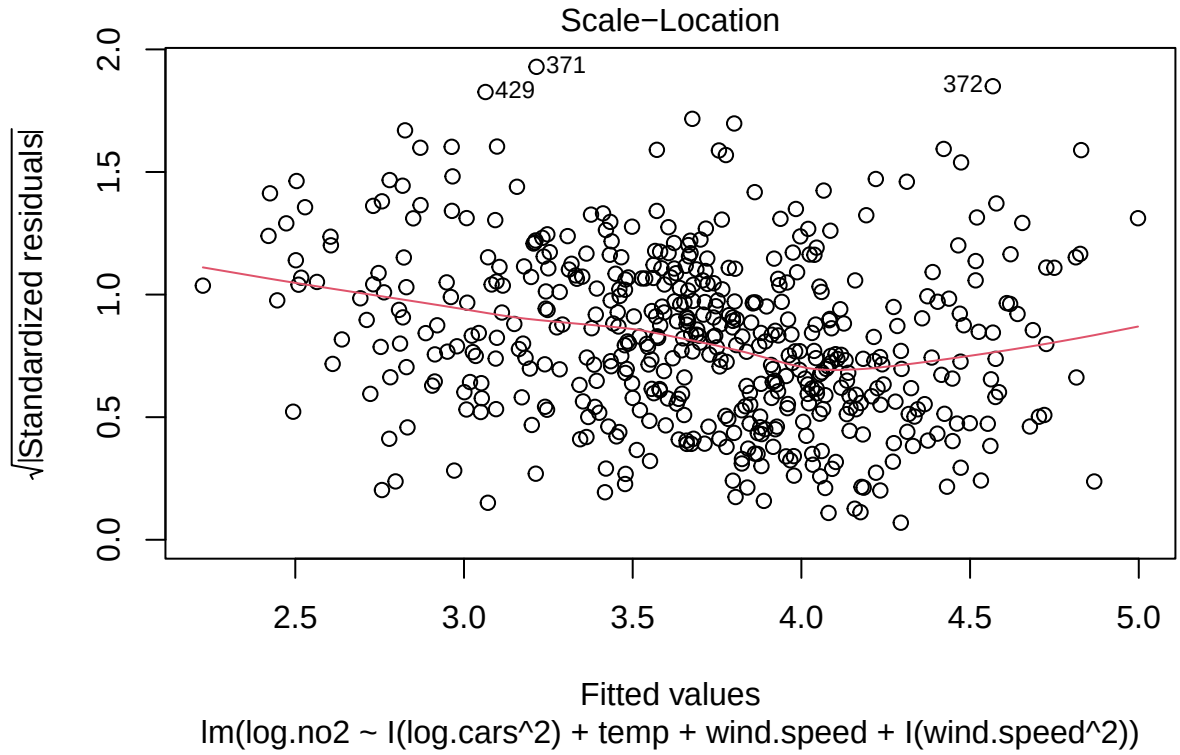


The residual vs. fitted plots:

```
plot(pollution.fit.11, 1)
```

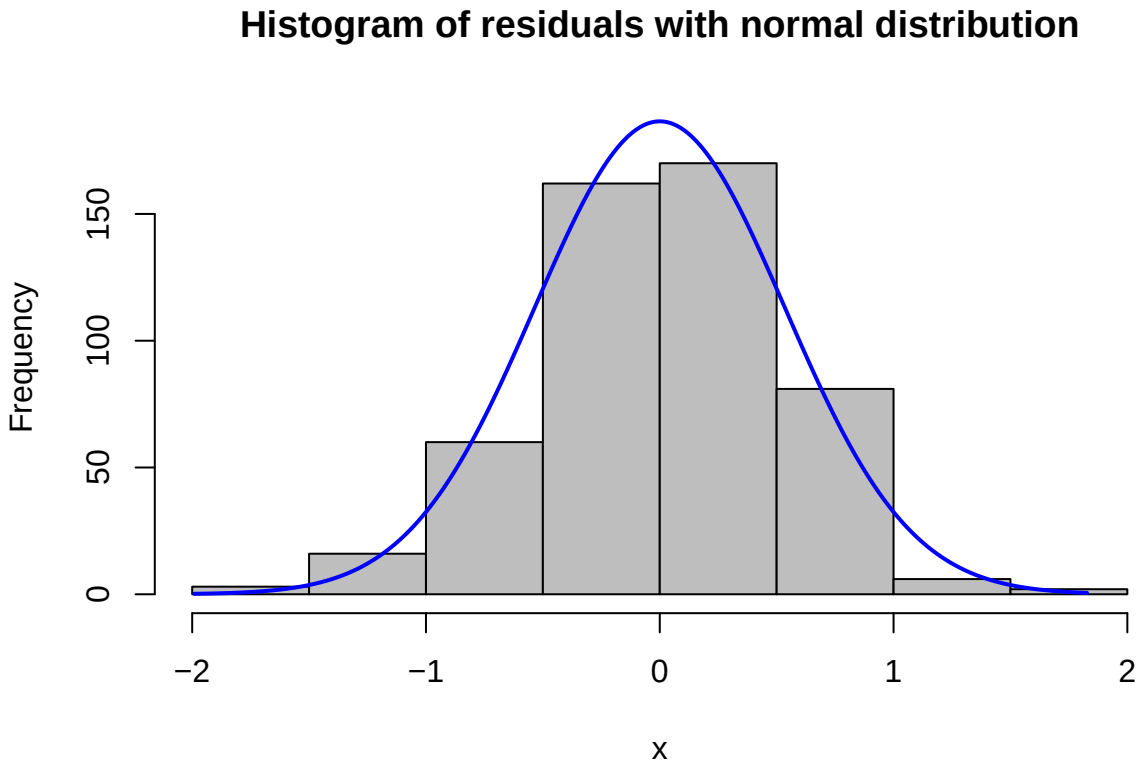



```
plot(pollution.fit.11, 3)
```

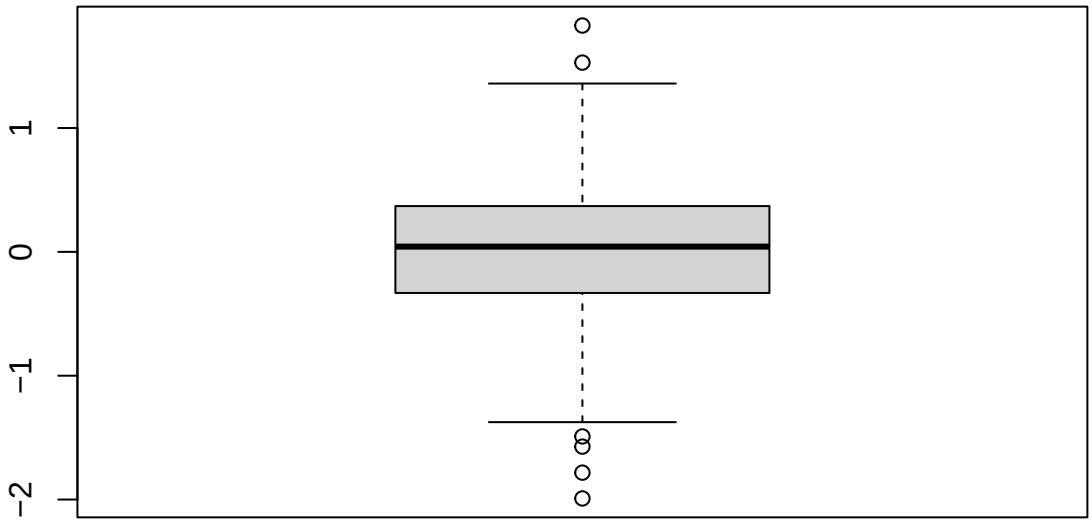


The histogram, boxplot and qq-plot of the residuals:

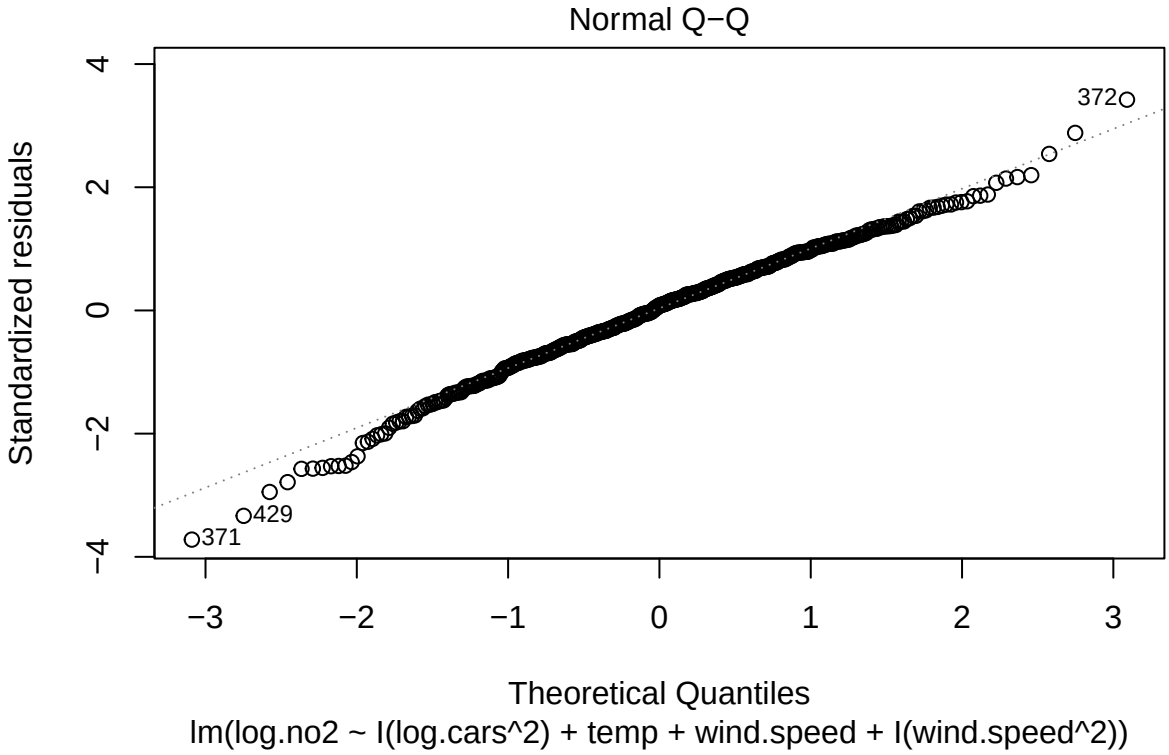
```
plotNormalHistogram(pollution.fit.11$residuals, main="Histogram of residuals with normal distribution")
```




```
boxplot(pollution.fit.11$residuals)
```




```
plot(pollution.fit.11, 2)
```



Discussion:

Interpreting the model:
We can see that the estimated intercept now is higher than in the previous, simple, regression model. This is because more independent variables have been included. We can see that the estimated parameter for the traffic (log.cars^2) and the second degree term of the wind speed both affect the no2 concentration, increasing it slightly if either the traffic or wind speed increases (estimated parameters: 0.032 and 0.029). The variable influencing the concentration the most is the first degree term of wind speed. It effects the no2 concentration negatively, and the estimated parameter is larger then its second degree equivalent (-0.37 vs 0.029). So the total effect of the wind speed will cause the no2 concentration to decrease when the wind speed increases. This seems reasonable, as the wind will redistribute the no2 particles.
The last estimated parameter belongs to the temperature variable. An increase in temperature will also decrease the no2 concentration, although only slightly (estimated parameter = -0.027).

All the above mentioned estimated parameters are estimated with a 99% significance level.
We can see that the models R^2 is 0.4925, which is considerably higher than for the simple model, when we only included the traffic. The model explains approximately half of the variance found in the dependent variable, meaning that there is still some improvements to be done to the model.

Comment on the assumption checks:
From the CPR plots for model we can see that all variables fit to the template line, meaning that there exists a linearity in the parameters. The second degree term for log.cars did not fit better to the model than log.cars did in the previous CPR plot. In model 9, both of the variables are included as predictors. We can see that this fits better for both of the variables, as both behave more as expected when looking for linearity in the parameters. The same behaviour can be observed for the wind speed, when comparing models 11's CPR plot to model 1's plots.
To sum up, the CPR plots look fine, and the linearity assumption is considered as fulfilled.

The same goes for the homoscedasticity assumption, based on the residual vs. fitted value plot, which looks fine. Although, one should maybe take a closer look at the standardized residual plot, as it seems like the variance decreases, which could indicate a pattern in the variance.

The normal distribution assumption seems also fulfilled, as the histogram of the residuals follows a bell shape, the boxplot looks symmetrical, ans the qq-plot follows mostly the template line.

Excersice 2:

Importing the blood-dataset, and defining the age-variable as categorical:

```
blood=read.table("https://www.uio.no/studier/emner/matnat/math/STK4900/data/blood.txt", header=TRUE, co
```

a)

Method:

I first inspect the mean, sd, median, IQR, min, max, 1.Qr and 3. Qr per age group. Also a histogram and boxplot of the groups blood pressure are inspected.

```
setDT(blood)
blood[, as.list(summary(Bloodpr)), by = age]
```

```
##    age Min. 1st Qu. Median     Mean 3rd Qu. Max.
## 1:   1  104   112.0    117 122.1667  129.00  160
## 2:   2  108   121.5    137 139.0833  157.75  174
## 3:   3  110   138.0    148 155.1667  164.00  214
```

The sd:

```
blood[ ,list(sd=sd(Bloodpr)), by=age]
```

```
##    age       sd
## 1:   1 15.33761
## 2:   2 22.62524
## 3:   3 27.71883
```

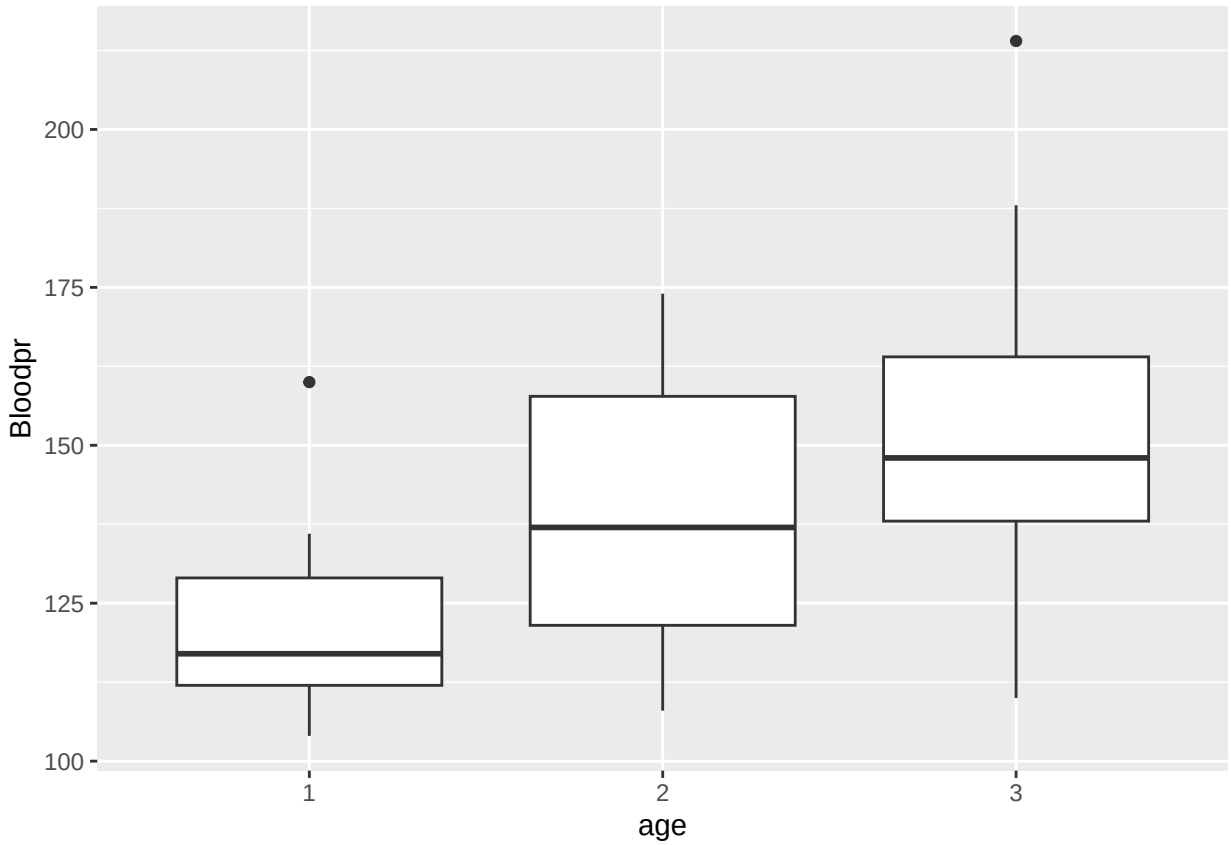
The IQR:

```
blood[ ,list(IQR=IQR(Bloodpr)), by=age]
```

```
##    age   IQR
## 1:   1 17.00
## 2:   2 36.25
## 3:   3 26.00
```

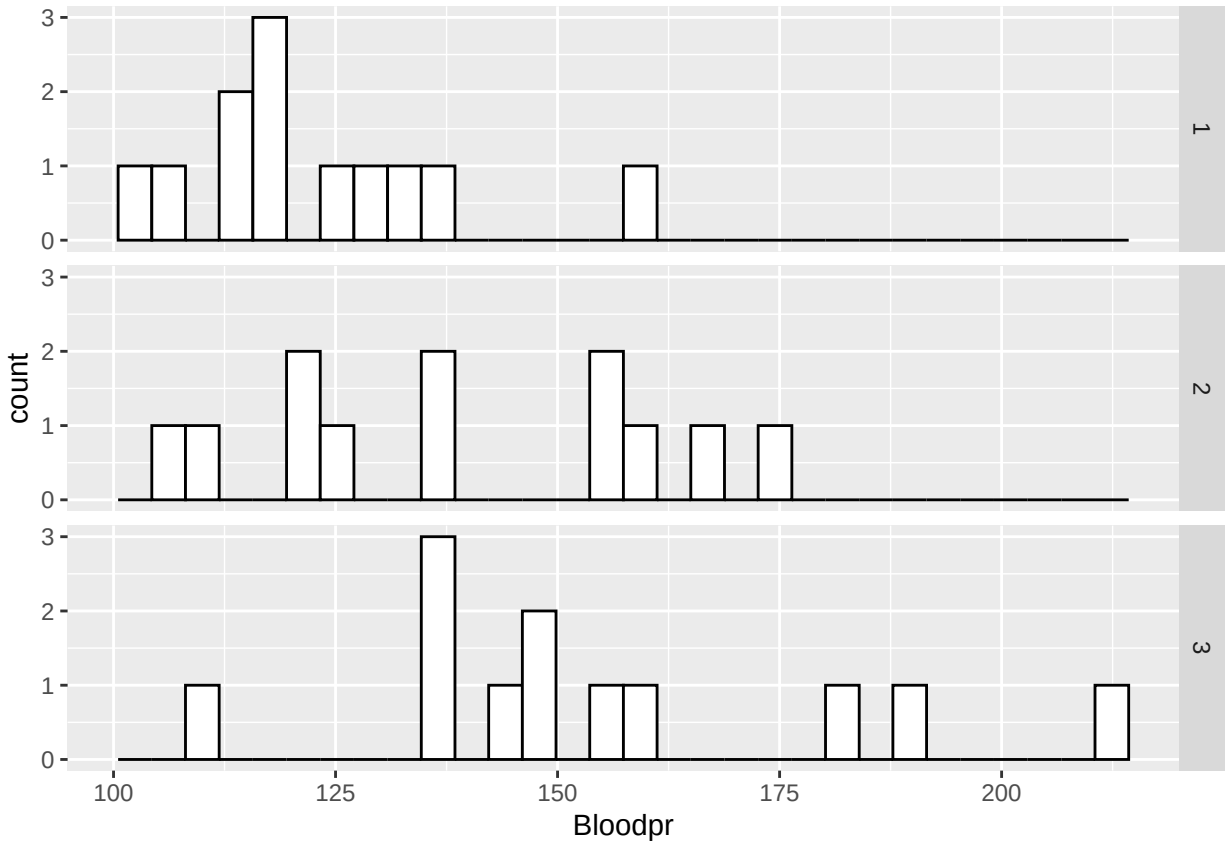
Plots by groups:

```
ggplot(blood, aes(x=age, y=Bloodpr)) +
    geom_boxplot()
```




```
ggplot(blood, aes(x = Bloodpr)) +
  geom_histogram(fill = "white", colour = "black") +
  facet_grid(age ~ .)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Discussion:

From the descriptive statistic, we can see that both the mean and median vary between the groups, indicating that the blood pressure could differ between groups. Also their 1. and 3. quartile are differently distributed. If we take a closer look at the groups histograms, we can see they are all differently distributed. None of them look very close to a bell shape, but this could also be because of the small sample size (12 per group). There are indications that if there were more samples, the group observations would look more than a bell shape. It looks like group 1 might have lower expected blood pressure, but it is hard to say if there is a difference between group 2 and 3.
When looking at the boxplot, it definitely looks like there could be a significant difference between group 1 and 3's expected blood pressure, as their boxes representing the IQR do not overlap. It is harder to say something about group 2. It could be that there is a sign. difference between group 1 and 2, as group 2's median is outside of group 1's IQR. I cannot say if there is a difference between group 2 and 3.

b)

Method:

A two-way ANOVA is performed to examine whether there is a difference in the expected blood pressure across the groups:

```
aov.blood = aov(Bloodpr~age, data=blood)
summary(aov.blood)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## age          2   6535    3268   6.469 0.00426 **
## Residuals   33  16670     505
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Discussion:

The assumptions that are done for a two-way ANOVA, are:

1. The observations are independent of each other. As the samples are random, and one person cannot be in multiple age groups, I assume that the assumption is fulfilled.
2. The dependent variable should be near-to-normally distributed for each group. This is hard to examine with such a small sample size. The histogram of the groups can be examined, but will not give a conclusive answer. From the histogram in 2a), one can see that the observations in the groups is very spread, only group 1 looks like something that could resemble a normal distribution. I am therefore nut fully sure if the assumption is fulfilled for this specific sample of people, and believe one would need more samples to determine this. But the test is robust to some deviation from the normal distribution, so I continue examining the ANOVA.
3. The variance for the groups are similar. One can test for this assumption by using for example Levenes test. But it needs to be mentioned that the sample size of the dataset is relative small (12 obs. in each group), and it can therefore be expected that the group's variance is unequal. We instead consider the groups' samplesize. As these are the same, I may assume the assumption as fulfilled.

In the hypothesis test, the null hypothesis states that the average expected blood pressure of all groups is the same.

The alternative hypothesis states that one or more groups might have a significantly different expected average blood pressure compared to the other groups.

A high F-value for 2 and 33 degrees of freedom might indicate that we can reject the null hypothesis, and that one or more groups' expected average outcome differs from the others. According to the performed ANOVA test, a low p-value is returned (p=0.00426), and we can reject the null hypothesis with a 99% significance level.

In order to evaluate which of the groups differs from the others, we need to perform a regression analysis and evaluate the expected outcome. This is performed in 2c). As mentioned earlier, based on the boxplot of the groups, it can be expected that at least group 1 and group 3 will differ from each other, but I am less confident whether group 2's expected blood pressure will be significantly different relative to the other groups.

c)

Method:

In order to compare the expected blood pressure for the groups a regression model is fitted, using the blood pressure as dependent variable and age as independent variable. A treatment-contrast is used with age group 1 as reference, meaning the intercept returns the expected outcome in the reference group, while for group 2 and 3 the expected outcome relative to group 1 is returned (expected blood pressure in group 2 - expected blood pressure in group 1,and expected blood pressure in group 3 - expected blood pressure in group 1).

```
blood.fit = lm(Bloodpr~age, data = blood)
summary(blood.fit)
```

```
##
## Call:
## lm(formula = Bloodpr ~ age, data = blood)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -45.167 -15.583  -5.167  14.104  58.833
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  122.167      6.488  18.829  < 2e-16 ***
## age2          16.917      9.176   1.844  0.07423 .
## age3          33.000      9.176   3.596  0.00104 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.48 on 33 degrees of freedom
## Multiple R-squared:  0.2816, Adjusted R-squared:  0.2381
## F-statistic: 6.469 on 2 and 33 DF,  p-value: 0.004263
```

Discussion:

From the estimated values, we can see that group 3' expected average blood pressure differs significantly from group 1, the reference, as previously suspected both in 2a) and 2b). It confirms that we could reject the null hypothesis, as there are at least 2 groups that differ significantly. Further, we can see that group 3 is estimated to have a 33 units higher average blood pressure than group 1. This corresponds to the descriptive statistics examined in 2a).

As expected based on 2a), it is hard to say whether group 1 and 2 differ. It depends on the required significance level. But when operating with a standard 95% sign. level, the two groups do not differ significantly.

In order to see if there is a significant difference between group 2 and 3, the regression would need to be repeated, using either group 2 or 3 as a reference. But based on the previous results and the descriptive statistics and plots, I sont expect them to differ significantly.