STK4900 - Assignment 1

Simen Nyhus Bastnes

2. March 2017

Problem 1

In this task, we want to study how the air pollution at a measuring station at Alnabru in Oslo is related to explanatory variables such as traffic volume, temperature, wind speed, and the hour of day. The different variables in the data set are

no2	The logarithm of the concentration of NO_2
log.cars	The logarithm of the number of cars per hour
wind.speed	Wind speed (meters/second)
hour.of.day	Hour of the day measurements were collected (1-24)

a) First of all, we want to look at the the NO_2 concentration, and the number of cars per hour. After loading in the data set, we do a summary of **no2** and **log.cars**

```
summary(no2data)
```

Which gives us the full summary of all the variables

```
      no2           log.cars          temp            wind.speed
 Min.   :1.224   Min.   :4.127   Min.   :-18.6000   Min.   :0.300
 1st Qu.:3.214   1st Qu.:6.176   1st Qu.: -3.9000   1st Qu.:1.675
 Median :3.848   Median :7.425   Median :  1.1000   Median :2.800
 Mean   :3.698   Mean   :6.973   Mean   :  0.8474   Mean   :3.056
 3rd Qu.:4.217   3rd Qu.:7.793   3rd Qu.:  4.9000   3rd Qu.:4.200
 Max.   :6.395   Max.   :8.349   Max.   : 21.1000   Max.   :9.900
  hour.of.day
 Min.   : 1.00
 1st Qu.: 6.00
 Median :12.50
 Mean   :12.38
 3rd Qu.:18.00
 Max.   :24.00
```

From this, we see the minimum, maximum, median, mean, as well as the 1st and 3rd quartile for each variable. The 1st quartile is the value where 75% of the measurements are above, and for the 3rd quartile 75% are below, meaning that between the 1st and 3rd quartile lies half of all the measurements.

Looking at **no2**, we see that the 1st and 3rd quartile is quite close to the mean compared to min/max. For **log.cars**, the mean is closer to the maximum value, and the 1st and 3rd quartile deviates a bit more with respect to the mean than it did for **no2**.

We also note that the **hour.of.day** variable shows that the measurements seem to be evenly spread out over the day. The temperature **temp** varies quite a lot over the day, which might be expected with how the measurements are spread out.

In order to get a better view over how the **no2** and **log.cars** variables are distributed, we create a scatterplot.

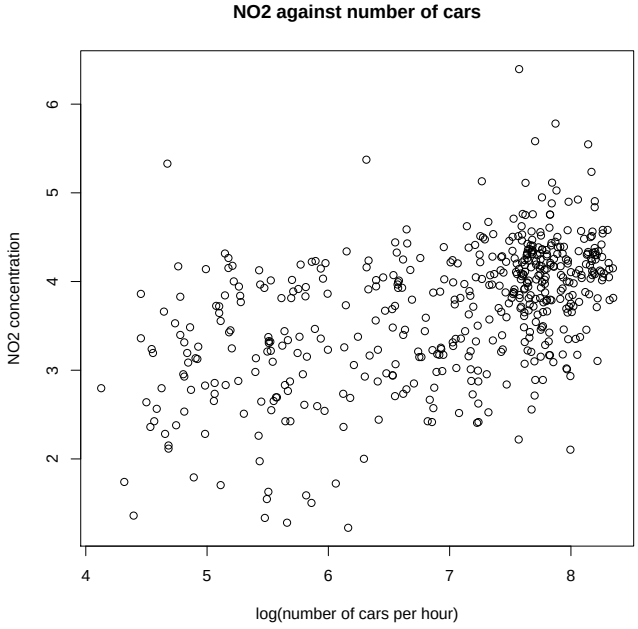

Figure 1: Scatterplot of the **no2** concentration against **log.cars**

From figure 1, we see that the density of measurement points seems significantly higher for higher values of **log.cars**, which is what we saw in the summary. The **no2** values also behave mostly like we saw earlier, with most values falling within a smaller interval. One thing to note, is that there seems to be some correlation between the variables, as the "average" concentration seems to increase with increasing traffic.

b) We fit a simple linear model where the log concentration of NO_2 is explained by the amount of traffic.

```
logcar.fit = lm(no2~log.cars, data=no2data)
```

A summary of this gives us the following

```
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
```

Looking at this summary, the most interesting for now is the coefficients underneath of *Estimate*. A linear curve can be described by

$$y(x) = ax + b$$

where a is the rate, and b represents where the curve intersects the y-axis at $x = 0$. For our linear regression model, the log.cars $= 0.35353$ is the a, the rate, while the (Intercept) $= 1.23310$ is b, such that it describes the curve

$$\text{no2}(\text{log.cars}) = 0.35353\,\text{log.cars} + 1.23310$$

We add the linear fit to the scatterplot shown earlier in figure 1

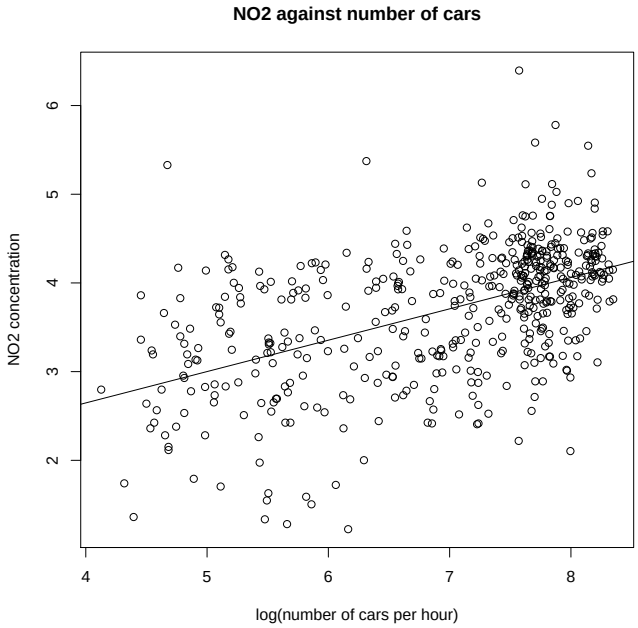

Figure 2: Scatter plot of NO_2 concentration plotted against log(number of cars per hour), with the fitted line added.

As figure 2 shows, the regression line follows the trend we noticed earlier when looking at the first scatterplot. This might indicate that there is indeed a linear correlation between the concentration of NO_2 and the number of cars, but there is also some outliers in the dataset. This can be seen in the R^2 measure from the summary, which is $R^2 = 0.2622$. R^2 is a measure of how well the linear regression line fits the data, and can be linked with the correlation factor. This R^2 factor ranges from 0 and 1, where 1 is a "good" fit. In our case, we then see that just a linear dependence on the number of cars might not be the best model we can make.

c) To further see if our model assumptions are reasonable, we want to check various residual plots. First we check the linearity by a CPR plot.

```
library(car)
crPlots(logcar.fit,terms=~log.cars)
```

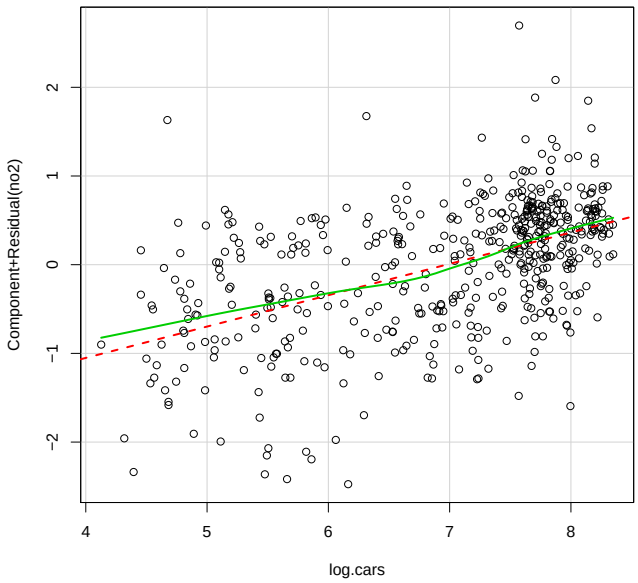

Figure 3: CPR plot for **log.cars**. The red dashed line is the fitted line, while the green solid line is the residuals.

We see in the plot that the green line deviates quite strongly from the fitted line, especially at the start, which indicates that there is non-linearity, and that our assumption does not hold very well.

Next, we want to check for constant variance (homoscedasticity). If the model is specified correctly, there should not be any systematic patterns in the residuals. We plot the residuals against the fitted values to get

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

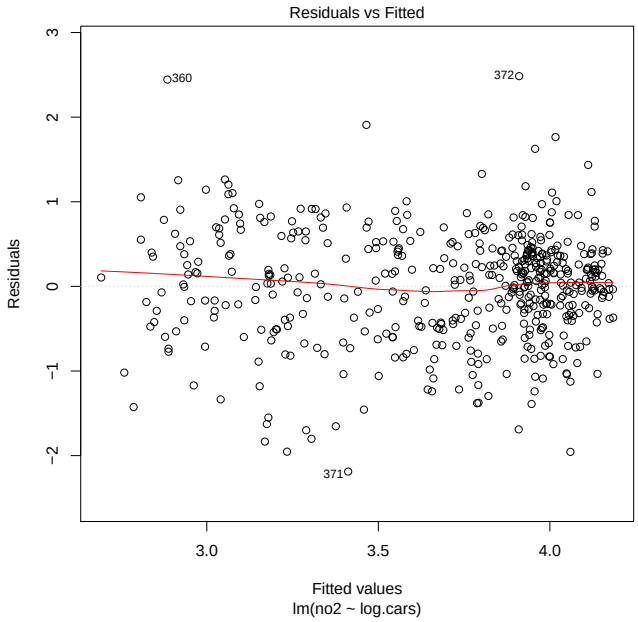

Figure 4: Residuals plotted against the fitted values. Red line is added to help see if there is a pattern in the residuals.

This does indicate that there is a pattern in the residuals, though not fully certain how to interpret it. We can also plot the standardized residuals against fitted values.

```
1 plot(logcar.fit,3)
```

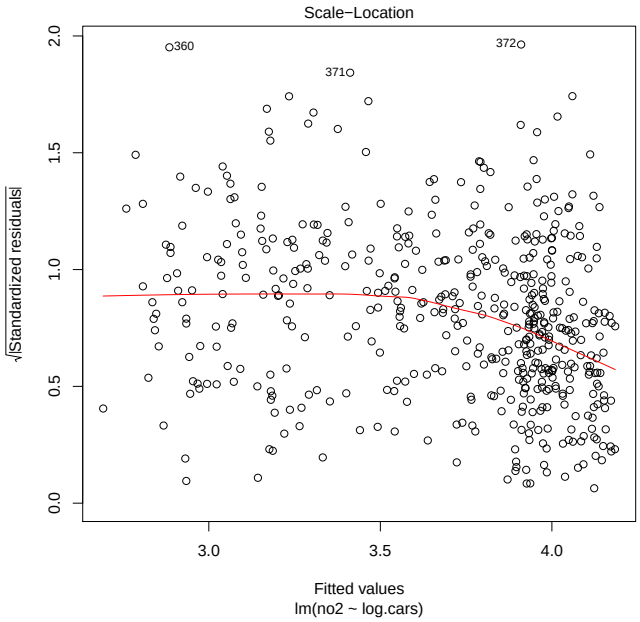

Figure 5: Standardized residuals plotted against the fitted values. Red line added to help see for patterns in the residuals.

We see that as the fitted values get higher, the red line decreases towards 0, so the variance decreases with increasing fitted values.

Finally, we want to check for normality, using that the residuals should behave as a sample from a normal distribution with mean zero if we have specified the model correctly.

```
hist(logcar.fit$res) # Histogram
boxplot(logcar.fit$res) # Boxplot
qqnorm(logcar.fit$res); qqline(logcar.fit$res) # Normal Q-Q plot
```

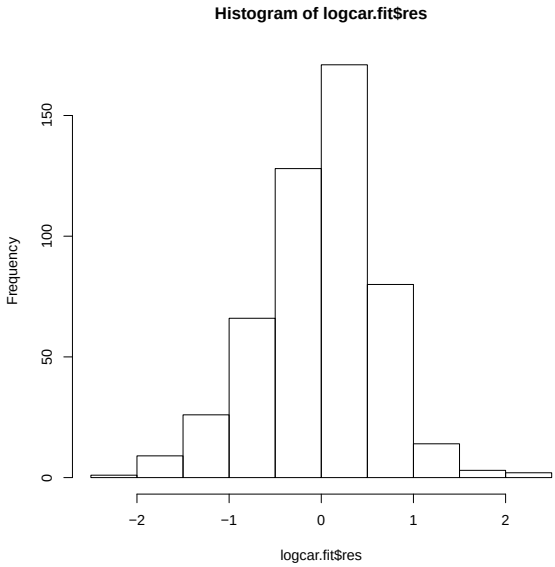

Figure 6: Histogram plot of the residuals

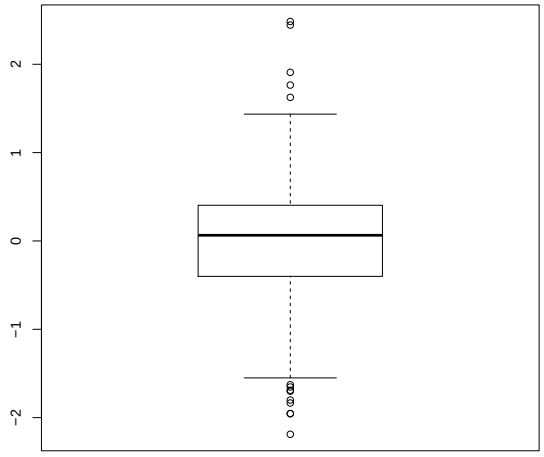

Figure 7: Boxplot plot of the residuals

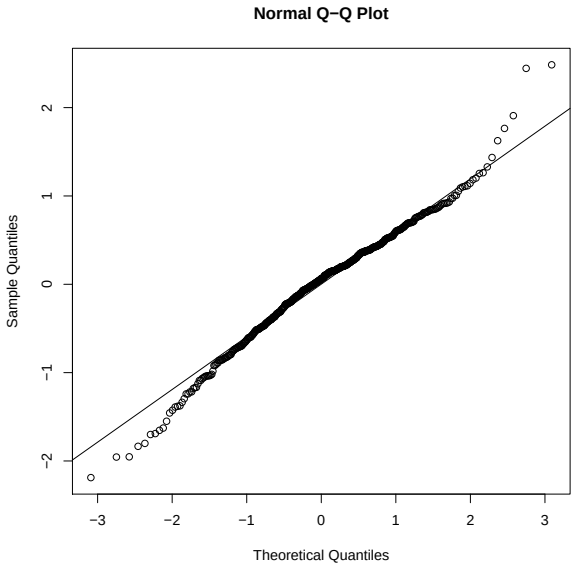

Figure 8: Normal Q-Q plot of the residuals

The histogram and boxplot look relatively close to a normal distribution, though maybe slightly shifted towards higher values. The Q-Q plot should be close to a straight line if the residuals are normally distributed, which is only true for the middle section, as in both ends, the residuals curve away from the straight line.

All of these tests make it obvious that our model assumptions are not reasonable, and therefore the NO_2 concentration does not just depend linearly on the traffic.

d) Using multiple regression, we want to study the simultaneous effect of the various covariates on the log concentration of NO_2. We attempt to find a better model than the one earlier by qualified guessing (plus trial and error). First of all, we add some dependence on the wind speed and temperature

```
logcars.fit1 = lm(no2~log.cars+log(wind.speed)+temp,data=no2data)
summary(logcars.fit1)
```

Which gives us the following summary

```
                 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 ***
log(wind.speed) -0.414496   0.036572 -11.334  < 2e-16 ***
```

```
temp            -0.026304   0.003861  -6.813 2.79e-11 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Multiple R-squared:  0.475,    Adjusted R-squared:  0.4718
```

Looking at both the significance of the P-values and R^2, we can determine how good the model is. The significance of the P-values can be seen by the number of asterisks (or potentially lack thereof), as explained in line 6. Although this is a significant improvement over what we did earlier, we can probably do better. Recall figure 3, where we concluded that there seems to be some non-linearity in the **log.cars** variable, so we could try to add a second-order term. In the end, we find the following "best" model

```
logcars.best = lm(no2~log.cars+I(log.cars^2)+wind.speed+I(wind.speed^2)+temp,data=
    no2data)
```

Which gives us the summary

```
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.338979   0.990857   5.388 1.10e-07 ***
log.cars        -0.763318   0.308159  -2.477 0.013582 *
I(log.cars^2)    0.089887   0.023556   3.816 0.000153 ***
wind.speed      -0.372617   0.043573  -8.552  < 2e-16 ***
I(wind.speed^2)  0.029321   0.005453   5.377 1.17e-07 ***
temp            -0.027120   0.003788  -7.160 2.94e-12 ***
Multiple R-squared:  0.4987,   Adjusted R-squared:  0.4936
```

e) The model coefficients found **d)** behave in similar fashion to how it did for the simple linear regression. The fitted values can be expressed as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_p x_{pi}$$

where $\hat{\beta}_0$ is the intercept, and $\hat{\beta}_1 \ldots \hat{\beta}_p$ are the coefficients of the covariates.

To verify that the model assumptions we found in **d)** are reasonable, we will use some of the plotting methods used earlier in **c)**. First, we check for linearity by making a CPR plot

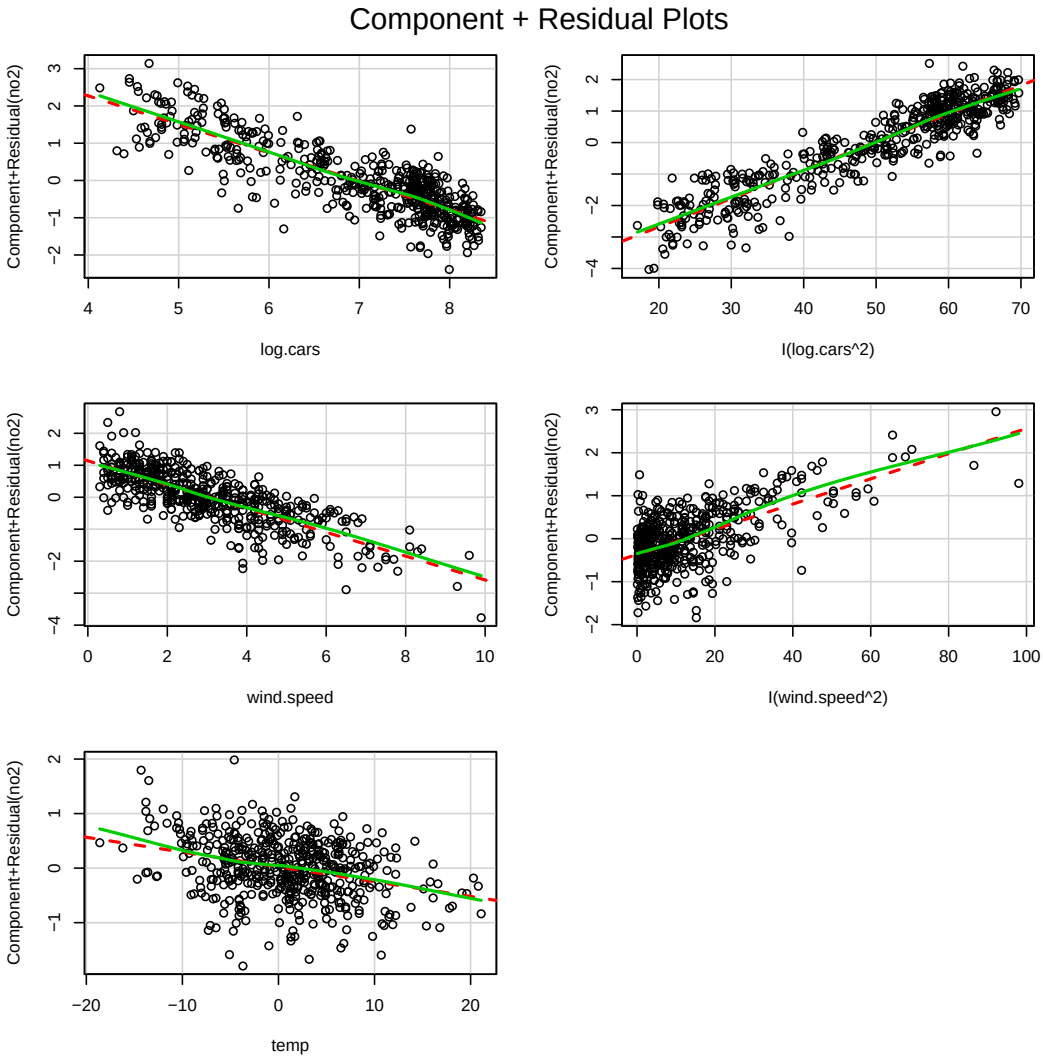

Figure 9: CPR plot showing the different components of our model.

As we see, in general the residuals are fairly close to the fitted lines, though the second order wind speed term might be somewhat problematic. So we can fairly safely say that the linearity assumption holds.

Problem 2

In this task, we will look at a data set from `blood.txt`, which contains measurements of the blood pressure of random samples of 12 men in each of three age groups. The age group is a categorical variable, and is coded with values 1, 2 and 3 in the table.

a) We are interested in checking whether or not the blood pressure is varying across age groups, so we start by looking at a simple summary of the data set.

```
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 104.0   117.5   136.0   138.8   156.2   214.0
```

This summary shows features of the whole data set, and shows that there seems to be more outliers with blood pressure above the mean. While we could create summaries for the specific age groups, we will instead create a boxplot containing all the age groups, hopefully allowing us to see the features more clearly.

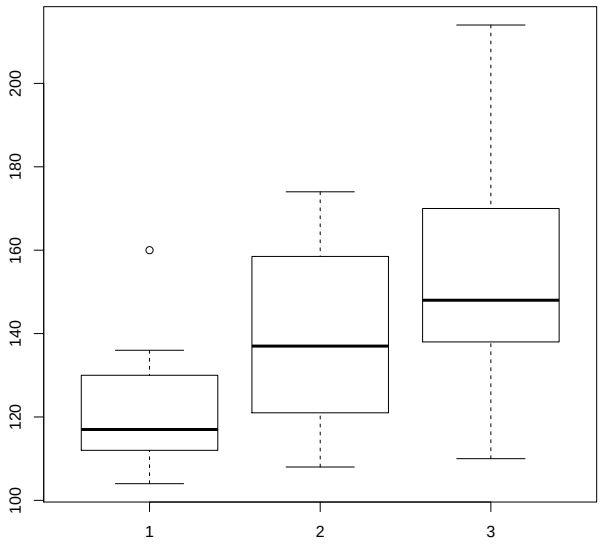

Figure 10: Boxplot of the blood pressure of the different age groups.

The figure above illustrates how the blood pressure is distributed in the different age groups. At a first glance, looking at the mean, 1st and 3rd quartile, we see that for increasing age groups, they increase respectively. While this seems like there is a connection between age and blood pressure, its worth noting that the variance of group 3 especially is very large, and we should study it a bit more carefully.

b) In order to determine whether the differences between the groups are statistically significant, we will use the one-way ANOVA.

```r
attach(blood)
anova(aov(blodtr~alder))
```

which gives us the ANOVA table

```
Analysis of Variance Table

Response: blodtr
          Df  Sum Sq Mean Sq F value   Pr(>F)
alder      2  6535.4  3267.7  6.4686 0.004263 **
Residuals 33 16670.2   505.2
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The assumptions behind the one-way ANOVA test is that we are looking at normal random samples from different independent groups. In the one-way ANOVA, we reject the null hypothesis for large values of the test statistic

$$F = \frac{MSS/(K-1)}{RSS/(n-K)}$$

which can be used to compute the P-value. Looking at our table, we see that we have a P-value of 0.004263, which is small enough so that we can safely reject the null hypothesis, confirming that there is a correlation between the blood pressure and the different age groups.

c) Finally, we want to formulate this problem using a regression model with age group as a categorical predictor variable, using treatment-contrast and the youngest group as reference.

```r
blood$alder = factor(blood$alder)
options(contrasts=c('contr.treatment','contr.poly'))
lin.fit = lm(blodtr~alder, data=blood)
summary(lin.fit)
```

Running this gives us the table

```
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 ***
alder2        16.917      9.176   1.844  0.07423 .
```

```
alder3        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
```

Looking at the table, we can interpret the coefficients **alder2** and **alder3** as the difference between the group and the reference group (in our case **alder1**). This tells us in more detail how the blood pressure varies depending on group than the one-way ANOVA, as we now are able to tell the differences between groups, and their significance to the result. It does seem as though the result from age group 2 is not significant, while the age group 3 is.

Code

Task 1a) and 1b)

```r
no2data <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/no2.txt",sep="
    \t",header=TRUE)

# Task 1a)
summary(no2data,1)
# Scatter plot of log.cars and no2
pdf('logcars_no2.pdf')
plot(no2data$log.cars, no2data$no2,main='NO2 against number of cars', xlab='log(number of
    cars per hour)', ylab='NO2 concentration')
dev.off()

# Task 1b)
# Linear fit log.cars and no2
no2data.fit.a = lm(no2~log.cars, data=no2data)
summary(no2data.fit.a)

pdf('logcars_no2_fit.pdf')
plot(no2data$log.cars, no2data$no2,main='NO2 against number of cars', xlab='log(number of
    cars per hour)', ylab='NO2 concentration')
abline(no2data.fit.a)
dev.off()
```

Task 1c)

```r
no2data <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/no2.txt",sep="
    \t",header=TRUE)

# Check various residual plots to judge if the model assumptions are reasonable
library(car)
logcar.fit = lm(no2~log.cars, data=no2data)

# Plot CPR plot
pdf('logcars_no2_crPlot.pdf')
crPlots(logcar.fit,terms=~log.cars)
dev.off()

# Check homoscedasticity
pdf('logcars_no2_1.pdf')
plot(logcar.fit,1)
dev.off()

pdf('logcars_no2_3.pdf')
plot(logcar.fit,3)
dev.off()

pdf('logcar_hist.pdf')
hist(logcar.fit$res)
dev.off()

pdf('logcar_box.pdf')
boxplot(logcar.fit$res)
dev.off()

pdf('logcar_qq.pdf')
qqnorm(logcar.fit$res); qqline(logcar.fit$res)
#plot(logcar.fit,2)
dev.off()
```

Task 1d) and 1e)

```r
no2data <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/no2.txt",sep="
    \t",header=TRUE)
library(car)

# Task 1d)
logcars.fit1 = lm(no2~log.cars+log(wind.speed)+temp,data=no2data)
summary(logcars.fit1)

logcars.fit2 = lm(no2~log.cars+I(log.cars^2)+log(wind.speed)+temp,data=no2data)
summary(logcars.fit2)
```

```
10
11 logcars.best = lm(no2~log.cars+I(log.cars^2)+wind.speed+I(wind.speed^2)+temp,data=no2data)
12 summary(logcars.best)
13
14 # Task 1e)
15 # Plot CPR plot
16 pdf('no2_1d_crPlot.pdf')
17 crPlots(logcars.best,terms=~.)
18 dev.off()
```

Task 2

```
1 blood <- read.table("http://www.uio.no/studier/emner/matnat/math/STK4900/v17/blood.txt",sep="
      ,",header=TRUE)
2
3 # Task 2a)  Summaries and boxplot
4 summary(blood$blodtr)
5 pdf('blood_box.pdf')
6 boxplot(blodtr~alder,data=blood)
7 dev.off()
8
9 # Task 2b)  One-way ANOVA
10 attach(blood)
11 anova(aov(blodtr~alder))
12
13 # Task 2c)  Regression model
14 blood$alder = factor(blood$alder)
15 options(contrasts=c('contr.treatment','contr.poly'))
16 lin.fit = lm(blodtr~alder, data=blood)
17 summary(lin.fit)
```