COMPULSORY ASSIGNMENT 1
STK4900: STATISTICAL METHODS AND APPLICATIONS

ELISABETH CHRISTENSEN

Problem 1:

a) By creating a scatter plot of the numerical data with the logarithm of the number of cars as an independent covariant, we see clearly that the concentration of NO_2 increases as the amount of cars increases as depicted in figure 1. By analyzing a summary from the data, as shown in table 1 with table 2 depicting the summary of the amount of cars during the day, we can see that the p-value for the logarithm of cars is $P < 2.2e-16$. In other words, it is significant on all significant levels. However, this does not imply that it is the only covariant present. This can be seen from the Pearson correlation coefficient where $r = 0.5120504$, meaning there is some correlation between the two variables, however the amount of cars cannot solely be the only factor affecting the concentration of NO_2 in the air. But this does imply that we can reject the null hypothesis as there is an increasing linear relationship between the independent variable log(cars) and the dependent variable $\log(NO_2)$.

b) The fitted model of the relationship between the concentration of NO_2 and the logarithm of number of cars per hour can be seen from figure 1, with the summary shown in table 3. Here, we see that during the start of the observations, at the intercept when time $t = 01{:}00$, the concentration of NO_2 in the air is approximately $\rho = 1.23310 \mu \text{gm}^{-3}$. If the amount of cars increase by 10, then the concentration will increase by $\rho = 0.35353 \mu \text{gm}^{-3}$. The R-squared measure, tells us how much the variance in one of the covariants is responsible for the variance in one or more of the other variables. It is defined as

$$R^2 = \frac{MSS}{TSS} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

In this case (from the summary of the fitted linear model) we are dealing with multiple predictors from the data, and must thus rely on multiple R-squared. However, this value always increases if we keep adding more and more predictors to our model, and can thus give a misleading result. This implies that we would get a better and better fit even though it doesn't necessarily tell us if the predictors we are adding are improving the model or not. To compensate for this effect we must look at the adjusted R-squared, which only increases if the included predictor improves the model, and decreases if otherwise. In this case, we have only taken into account the predictor of the number of cars, and thus the multiple R-squared and the adjusted R-squared are fairly similar. Here, $R^2 \sim 0.26$, implying that the number of cars per hour is not solely responsible for the variances in our model. There must be other predictors affecting the model to allow for these variances which we have not included in our fit.

TABLE 1. Correlation test of the logarithm of the number of cars per hour vs. the logarithm of the concentration of NO_2 in the air

t-value	df	p-value.	r	[95% Conf.	Interval]
13.303	498	$< 2.2e-16$	0.5120504	0.4443103	0.5739687

TABLE 2. Summary of the independent variable log(cars per hour)

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

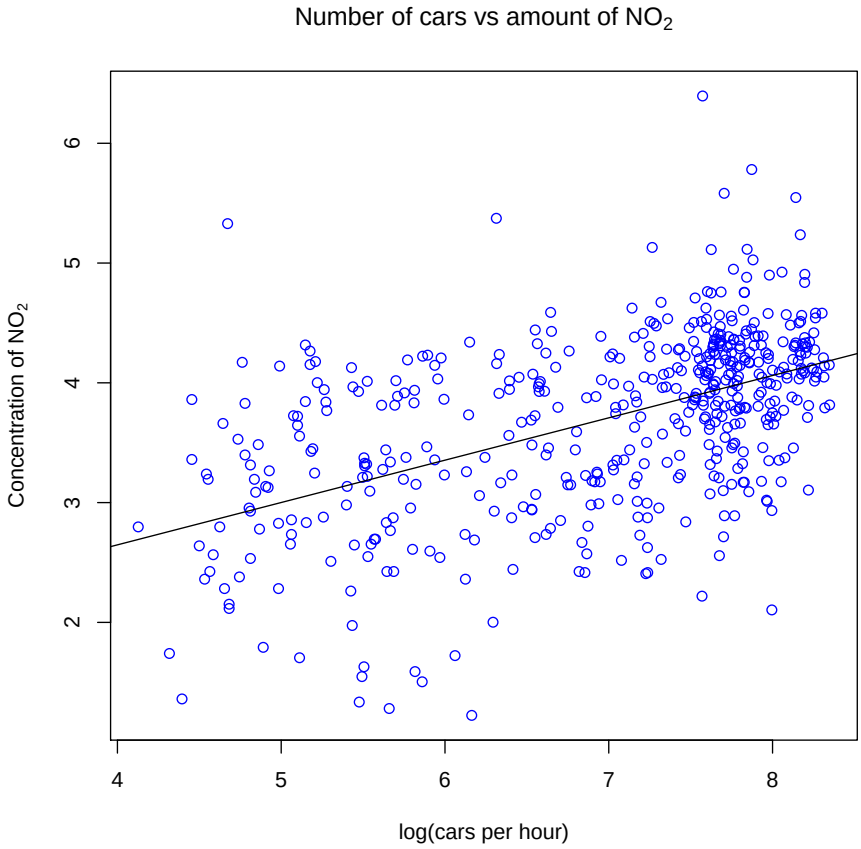


FIGURE 1. Scatter plot of the logarithm of the number of cars per hour vs. the concentration of nitrogen dioxide

c) We can check for constant variance by looking at the scatter plot between the residuals and the fitted values, as can be shown in figure 2a. Here, we see that there is a clear sign of homoscedasticity, where the variance is roughly the same across all observations of the variable log(cars). This also suggests that our assumption of the relationship between log(cars) and log(NO_2) to be linear seems reasonable. However, we do have some outliers which are not taken into account. This can be verified by our Q-Q plot in figure 3b, showing a heavy-tailed distribution. I.e. there are more observations in the middle of the distribution and fewer as we move further away from the

TABLE 3. Summary of the fitted model between the logarithm of the number of cars per hour and the logarithm of NO_2.

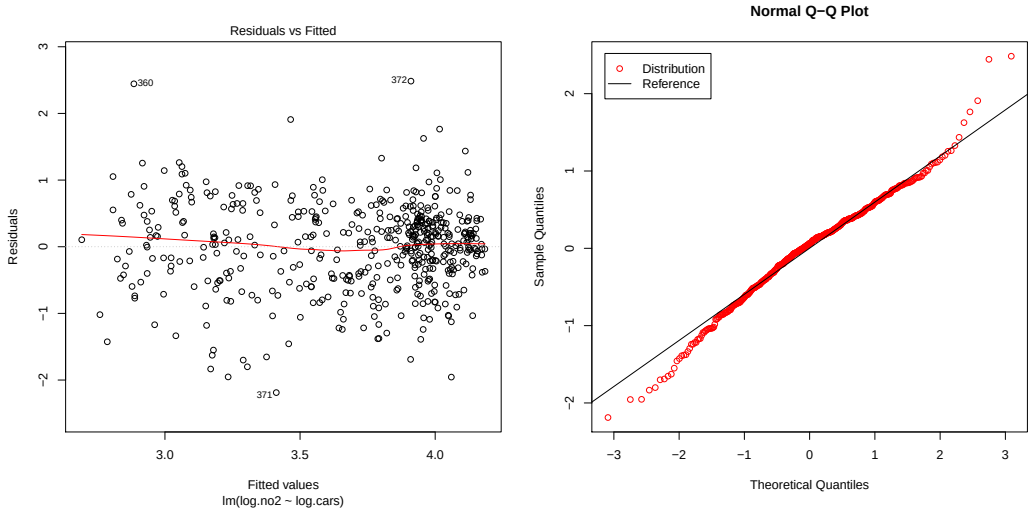
Coefficients:	Estimate	Std. error	t-value	$\Pr(>\|t\|)$
(Intercept)	1.23310	0.18755	6.575	$1.23e-10$
log.cars	0.35353	0.02657	13.303	$< 2e-16$

R^2_{multiple}	R^2_{adj}	F-statistic
0.2622	0.2607	177

TABLE 4. Summary of residuals

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-2.18822	-0.40071	0.06428	0.00000	0.40362	2.48472

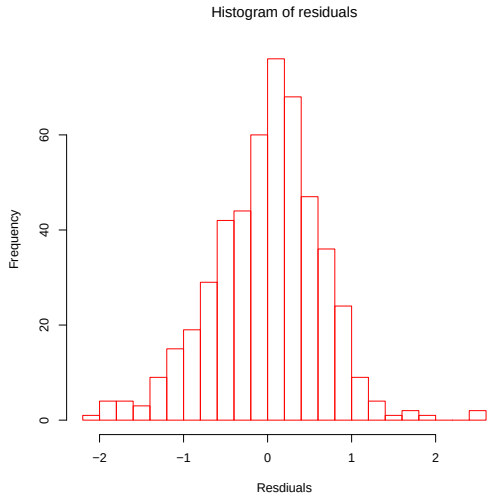
middle. This also means that it is more plausible to have outliers, as we clearly can see from the point in the far-left bottom corner and the two points in the far-right upper corner.
We can use the histogram, as shown in figure 3a, to check for normality of the residuals. The histogram clearly shows a Gaussian distribution, with mean $\mu = 0$, implying the residuals to have a normal distribution. The boxplot corresponds to the summary as seen in table 4, where we have a mean at $\mu = 0$, with median at 0.06428, and a minimum at -2.18822, and maximum at 2.48472. The median of the lower half of the data set is $Q_1 = -0.40071$, meaning about 75% of the data lies above Q_1, whereas the median of the upper half of the data set is $Q_3 = 0.40362$. The outliers can also be seen from the boxplot.



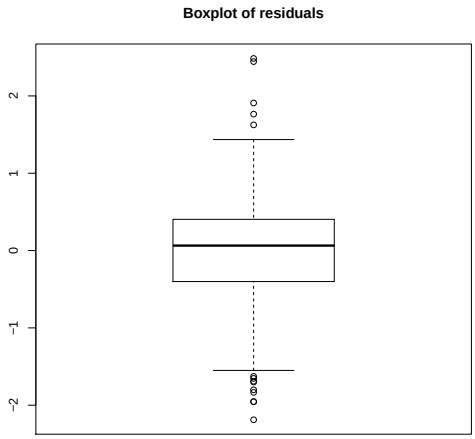
(A) Plot of residuals vs fitted values for the independent variable log(cars)

(B) Q-Q plot of residuals for the independent variable log(cars)

FIGURE 2. Plot (A) shows the plot of the residuals vs. fitted values when only comparing the logarithm of NO_2 and the logarithm of the number of cars per hour, while plot (B) shows the normal Q-Q plot from the distribution.



(A) Histogram of the residuals with 20 intervals



(B) Boxplot of residuals

FIGURE 3. Plot (A) shows the histogram of the residuals with 20 intervals, while plot (B) shows the boxplot of residuals

TABLE 5. Possible best fit models with adjusted R-squared (R^2_{adj})

Model	R^2_{adj}
log(cars) + wind.speed+ hour.of.day + temp	0.4165
log(cars) + log(wind.speed) + hour.of.day + temp	0.4791
log(cars) + log(wind.speed) + log(hour.of.day) + temp	0.4766
log(cars) + wind.speed + log(hour.of.day) + temp	0.582

d) When trying to find the best fit for our model, we can look at the adjusted R^2 in order to find the best model which can account for the variances in our data. In doing so, we can firstly look at the different combinations to first order, where we also log transform some of the variables. We then get the values as described in table 5 for different fits. Here, we see that the best fit is that of $\log(NO_2) \sim \log(\text{cars}) + \log(\text{wind.speed}) + \text{hour.of.day} + \text{temp}$ which gives an adjusted $R^2 = 0.4791$. Building on this fit we can attempt to increase the order of the polynomials in order to get a better fit. I.e. we increase the polynomials of the variables given in our current best fit and continue to add higher-order polynomials until the R^2-measure stops increasing or starts to decrease. In doing so, we find that the best model is

$$\begin{aligned}\log(NO_2) \sim & \log .\text{cars} + I(\log(\text{wind.speed})^2) + I(\log(\text{wind.speed})^3) \\ & + I(\log(\text{wind.speed})^4) + I(\log(\text{wind.speed})^5) \\ & + I(\text{hour.of.day}^2) + I(\text{hour.of.day}^3) + \text{temp}\end{aligned}$$

This leaves us with an adjusted R-squared measure as $R^2_{\text{adj}} = 0.491$. The plot for the CPR plots for different variables can be seen in figure 4.

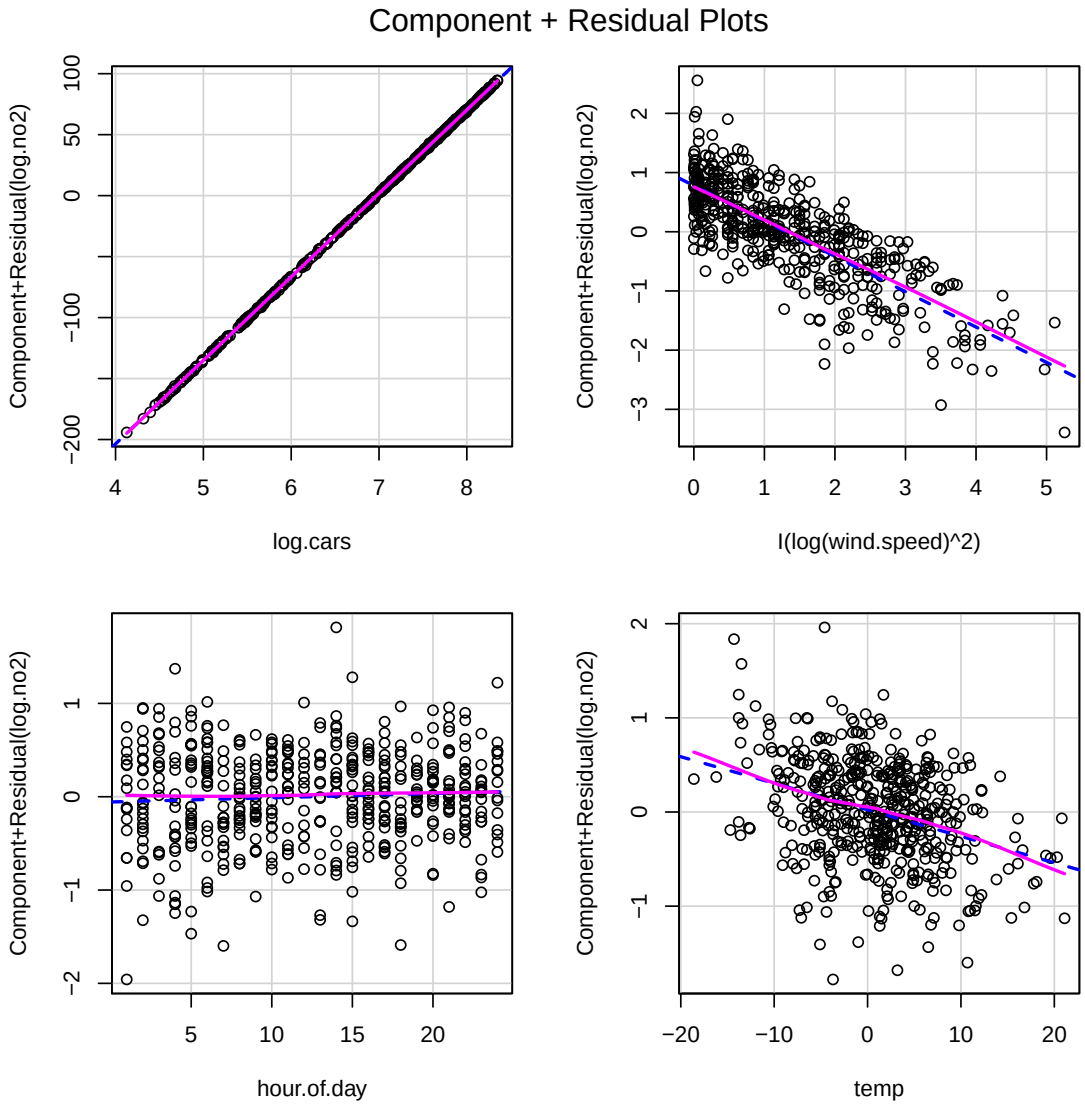


FIGURE 4. CPR plots for the variables log.cars, I(log(wind.speed)^2), hour.of.day and temp from the fit $\log(NO_2) \sim \log .\text{cars} + I(\log(\text{wind.speed})^2) + I(\log(\text{wind.speed})^3) + I(\log(\text{wind.speed})^4) + I(\log(\text{wind.speed})^5) + I(\text{hour.of.day}^2) + I(\text{hour.of.day}^3) + \text{temp}$

e) By looking at the different covariates we see from table 6 that the intercept lies at 1.070, meaning that we start out with a concentration of $\log(NO_2)$ at $\rho = 1.070 \mu\text{gm}^{-3}$ in the air, at the very start of the observations. As the logarithm of cars increase by 1, then the concentration of $\log(NO_2)$ increases by 4.72e-01. This has an obvious significance on the results as stated in the previous tasks. Looking at the second degree term of the log(wind.speed) we see that it too also plays a significant role in the density of $\log(NO_2)$ in the air, where it transports away about 6.655e-01 μgm^{-3},

TABLE 6. Summary of the best fit model

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)	
(Intercept)	1.070e+00	2.240e-01	4.776	2.36e-06	***
log.cars	4.724e-01	3.961e-02	11.926	<2e-16	***
I(log(wind.speed)^2)	-6.655e-01	1.777e-01	-3.746	0.000201	***
I(log(wind.speed)^3)	-1.691e-01	8.471e-02	-1.997	0.046388	*
I(log(wind.speed)^4)	4.051e-01	1.171e-01	3.460	0.000587	***
I(log(wind.speed)^5)	-1.021e-01	4.554e-02	-2.242	0.025414	*
I(hour.of.day^2)	-1.936e-03	1.427e-03	-1.356	0.175580	
I(hour.of.day^3)	6.116e-05	5.868e-05	1.042	0.297777	
temp	-2.692e-02	3.796e-03	-7.091	4.66e-12	***

TABLE 7. Summaries of blood pressure with respect to each age group

Age group:	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
30-45 years	104.0	112.0	117.0	122.2	129.0	160.0
46-59 years	108.0	121.5	137.0	139.1	157.8	174.0
60-75 years	110.0	138.0	148.0	155.2	164.0	214.0

as it increases by 1. The same goes for the third and fifth degree term of log(wind.speed), where they each decrease the density by 1.691e-01 and 1.021e-03 respectively. However, for the fourth degree term we see that the density increases which contradicts with our intuition. This might be a result of overfitting the model, although the term seems to be significant with a p-value p = 0.000587. The second polynomial of the hour of day does not have the greatest significance on the concentration of $\log(NO_2)$, but it does decrease for every hour by approximately 1.936e-03. Lastly, the temperature is highly significant and decreases the density of $\log(NO_2)$ by 2.692e-02 if it increases by 1 degree Celsius.

Problem 2:

a) By looking at the summaries of blood pressure with respect to each individual age group we get the values as shown in table 7. Here, we can clearly see that the mean increases with each age group. This also goes for the minimum and maximum values, where the youngest age group has a minimum value of 104.0 and a maximum value of 160.0, while the oldest age group has a minimum value of 110.0 and a maximum value of 214.0. It is also clear to see from the 1st and 3rd quartiles that in age group 1, 75% of the data lie above 112.0, and below 129.0, while in age group 2 it is between 121.5 and 157.8, and lastly in age group 3 the majority of the data lie within 138.0 and 164.0.
We can verify this with a boxplot of the blood pressures with respect to each age group as shown in figure 5. Here, we can clearly see the pattern in blood pressures as the age increases. It also appears that the variance, or the spread of data is increasing as well. From the boxplot for age group 1 we also have a clear outlier which has not been taken into account when calculating the different values from table 7.

b) By looking at the one-way ANOVA, shown in table 8, we see that the age group, with two degrees of freedom, is significant in the results of the blood

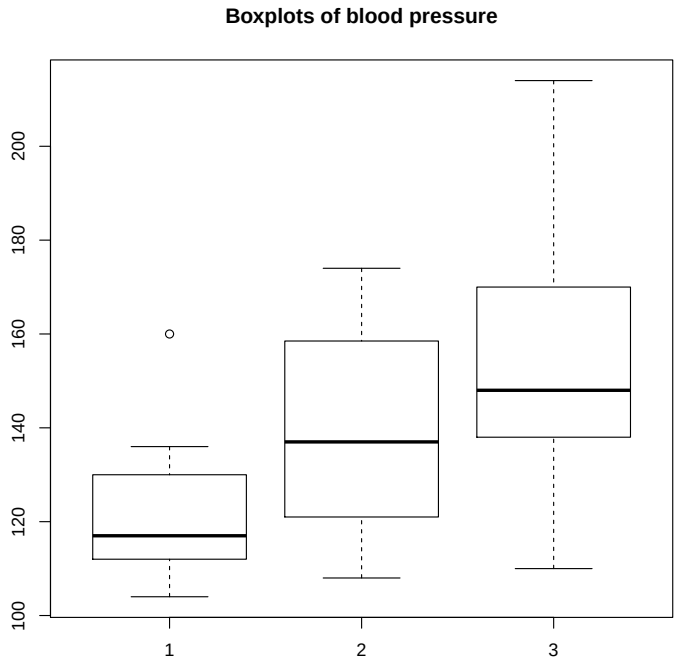


FIGURE 5. Boxplot of the blood pressures with respect to each age group

TABLE 8. One-way ANOVA table

	Df	Sum Sq	Mean Sq	F-value	Pr(>F)
age_group	2	6535.4	3267.7	6.4686	0.004263
Residuals	33	16670.2	505.2		

pressure with a p-value of $P = 0.004263$, and can thus reject the null-hypothesis. We also see from the F-value, where F = 6.4686, that there is a larger variance between the mean of each group compared to the variance of observations within each group.

c) By looking at the summary of the linear fit to our data, we get the values as shown in table 9. This shows that with age group 1 as our intercept, the blood pressure will be on average 122.167, but will increase by 16.917 if we move on to age group 2 from the intercept, and by 33.00 if we look at age group 3. By looking at the p-values as given in the table, with the age group 30-45 years as our reference group, we see that the age group 46-59 years has in comparison a p-value of P = 0.07423 $>$ 0.05, meaning that there does not exist a significant difference in the mean between this age group and the reference group. However, when looking at age group 60-75 years we see that the p-value is P = 0.00104 $<$ 0.05, and there must thus be a significant difference in the mean between this age group and the reference group. This also allows us to reject the null hypothesis for this specific case.
Comparing this to our discussion from task b, we see that this corresponds

TABLE 9. Summary of linear fit to the data

Coefficients:	Estimate	Std. Error	t-value	Pr(>|t|)
30-45 years	122.167	6.488	18.829	<2e-16
46-59 years	16.917	9.176	1.844	0.07423
60-75 years	33.000	9.176	3.596	0.00104

to the F-value, which states that the variance between the mean of each group is higher than the variance of observations within each group.

Throughout this assignment I have worked alongside with Maren Rasmussen.

REFERENCES

[1] Vittinghoff E., Shiboski S. C., Glidden D.V., McCulloch C.E., 2012, 'Regression Methods in Biostatistics', 2nd edn., Springer

Appendix A: Problem 1

```
#Problem 1:
library(latex2exp)
library(car)

setwd('C:\\Users\\Elisabeth\\Documents\\Courses\\STK4900\\Compulsory_assignment1')
#set working directory

no2data <- read.table("no2.txt",sep="\t",header=TRUE)
#sep="\t" => the values in the file are separated by tabs
#header = TRUE => the first row in the file contains the names of each column
no2data

summary(no2data)

log.cars = no2data$log.cars
log.no2 = no2data$log.no2
temp = no2data$temp
wind.speed = no2data$wind.speed
hour.of.day = no2data$hour.of.day

#a)
#open pdf-file
pdf("cars_NO2.pdf")
plot(log.cars, log.no2, xlab="log(cars␣per␣hour)", ylab = TeX("Concentration␣of␣NO$_2
     main =  TeX("Number␣of␣cars␣vs␣amount␣of␣$NO_2$"), col = "blue", pch=1)


#b)
fit.no2.cars = lm(log.no2~log.cars)
abline(fit.no2.cars)

#close pdf-file
dev.off()

legend = "Fit"
legend("topleft", inset = .03, legend, col = c("black"), lty=1, cex=1)

summary(fit.no2.cars)
#checking the correlation between the independent variable log(cars) and log(NO2)
cor.test(log.cars, log.no2)

#c) Checking various plot and residuals:
no2.cars.res = fit.no2.cars$residuals
no2.cars.fitval = fit.no2.cars$fitted.values
summary(no2.cars.res)

#check of constant variance (homoscedasticity)
plot(no2.cars.fitval, no2.cars.res, pch=1)
pdf("no2_cars_const_variance.pdf")
plot(fit.no2.cars, 1)
dev.off()
```

```
#check of normality
pdf("no2_cars_hist.pdf")
hist(no2.cars.res, breaks = 20, main = TeX("Histogram␣of␣residuals"),xlab = TeX("Resc
dev.off()

pdf("no2_cars_boxplot.pdf")
boxplot(no2.cars.res, main="Boxplot␣of␣residuals", border="blue")
dev.off()

pdf("no2_cars_qq.pdf")
qqnorm(no2.cars.res, col = "red"); qqline(no2.cars.res)
legend = c("Distribution", "Reference")
legend("topleft", inset = .03, legend, col = c("red", "black"), pch = c(1, NA), lty=c
dev.off()

#Check of linearity
crPlots(fit.no2.cars, terms=~log.cars)


#d) Best model fit
fit.no2.temp = lm(log.no2~temp)
no2.temp.res = fit.no2.temp$residuals
no2.temp.fitval = fit.no2.temp$fitted.values

cor.test(log(hour.of.day), log.no2)

fit.no2.windspeed = lm(log.no2~log(wind.speed))

fit.no2.hod = lm(log.no2~hour.of.day)

crPlots(fit.no2.temp, terms=~temp)
crPlots(fit.no2.windspeed, terms=~log(wind.speed))

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

fit.2 = lm(log.no2~log.cars + log(wind.speed) + hour.of.day + temp)
summary(fit.2)

fit.3 = lm(log.no2~log.cars + log(wind.speed) + log(hour.of.day) + temp)
summary(fit.3)

fit.4 = lm(log.no2~log.cars + wind.speed + log(hour.of.day) + temp)
summary(fit.4)

anova(fit.1)
anova(fit.2)

best_fit_values = fit.2$fitted.values
best_fit_res = fit.2$residuals
```

```
#check of constant variance (homoscedasticity)
plot(best_fit_values, best_fit_res, pch=1)
plot(fit.2, 1)

#check of normality
hist(best_fit_res, breaks = 20, main = TeX("Histogram␣of␣$NO_2$␣-␣cars␣residuals"),xl
boxplot(best_fit_res, main="Boxplot␣of␣residuals")
qqnorm(b, est_fit_res, col = "red"); qqline(best_fit_res)

#Check of linearity
crPlots(fit.2, terms=~log(wind.speed))
fit.sq2 = lm(log.no2~log.cars + log(wind.speed) + I(log(wind.speed)^2) + I(log(wind.s
crPlots(fit.sq2, terms=~log.cars + I(log(wind.speed)^2) + hour.of.day + temp)
summary(fit.sq2)

best_fit_values_sq = fit.sq2$fitted.values
best_fit_res_sq = fit.sq2$residuals
plot(best_fit_values_sq, best_fit_res_sq, pch=1)
plot(fit.sq2, 1)

#Attempt at better adjusted R^2:
fit.sq3 = lm(log.no2~log.cars + I(log(wind.speed)^2) +
               I(log(wind.speed)^3) + I(log(wind.speed)^4) + I(log(wind.speed)^5) + I
               I(hour.of.day^3) + temp)
pdf("best_fit_model.pdf")
crPlots(fit.sq3, terms=~log.cars + I(log(wind.speed)^2) + hour.of.day + temp)
dev.off()
summary(fit.sq3)
plot(fit.sq3)

anova(fit.sq3)
```

Appendix B: Problem 2

```
#Problem 2:
library(latex2exp)
library(car)

setwd('C:\\Users\\Elisabeth\\Documents\\Courses\\STK4900\\Compulsory_assignment1')
#set working directory

datablood <- read.table("blood.txt",sep=",",header=TRUE)
#sep="\t" => the values in the file are separated by tabs
#header = TRUE => the first row in the file contains the names of each column
datablood

bloodtr = datablood$blodtr; ages = datablood$alder

#Defining covariates as factors, i.e. categorical covariates:
age_group = factor(ages)
age_group
#a)
summary(bloodtr[age_group==1])
summary(bloodtr[age_group==2])
summary(bloodtr[age_group==3])

pdf("boxplot_a.pdf")
boxplot(bloodtr~alder,data=datablood, main="Boxplots␣of␣blood␣pressure")
dev.off()


#b)
#Assumptions of one-way ANOVA
#1. Each sample is taken from a normally distributed population
#2. Each sampe has been drawn independently from the others
#3. The variance of the data if the different groups are the same
#4. The variable, on which the measurements are based on, can be subdivided into sepa
#different age groups
#5. Three or more groups are to be compared. (Two-way ANOVA compares data where each
#samples)

#one-way ANOVA:
fit.datablood = lm(bloodtr~age_group, data=datablood)
anova(fit.datablood)


#c)
fit = lm(bloodtr~age_group, data=datablood)
summary(fit)
```