STK4900 Oblig1

Daniel Heinesen, daniehei

March 12, 2019

Exercise 1

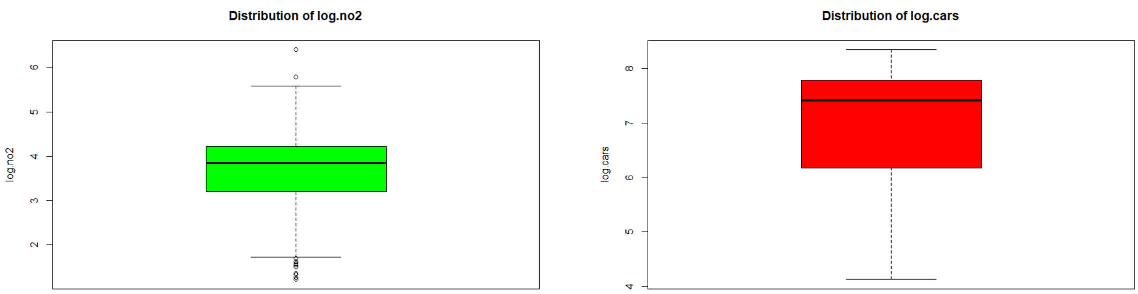
a)

log.no2	log.cars
Min. :1.224	Min. :4.127
1st Qu.:3.214	1st Qu.:6.176
Median :3.848	Median :7.425
Mean :3.698	Mean :6.973
3rd Qu.:4.217	3rd Qu.:7.793
Max. :6.395	Max. :8.349

Table 1: Summary of the main features of *log.no2* and *log.cars*

We are looking at the variables *log.no2* and *log.cars*, which represents the logarithm of the measured concentration of NO_2 and the logarithmic number of cars per hour. Table 1 shows a numerical summary of *log.no2* and *log.cars*. Looking at *log.no2* we see that the mean is close to the median, and that they both are more or less equidistant from *min* and *max*, and from the first and third quadrant. This indicates that the distribution is almost symmetric, with a slight skew. Looking at *log.car* we see that this is no longer the case. The mean and median is more different and lay closer to *max* and the third quadrant . This means that this distribution is very skewed.

check that it is called quadrant or quarter



(a) Box plot of the distribution of *log.no2*

(b) Box plot of the distribution of *log.cars*

Figure 1: Box plots showing the distributions for *log.no2* and *log.cars*.

Our ideas about the distribution we got from the numerical summaries, tab. 1, is confirmed by the box plots of the distribution found in fig. 1. Here we can see that *log.no2* is more or less symmetric, while *log.cars* is heavily skewed towards higher values.

We can look at the relationship between *log.no2* and *log.cars*. If we look at fig. 2 we see that there looks to be a dependence. We are going to try to fit a linear model to the data, but looking at the plot we see that there seems to be a higher clustering for higher values, which we expect from fig. 1. This means that there seems to be a larger spread for lower values than for the higher, making some of the assumptions used for a linear regression false. We will look at that later.

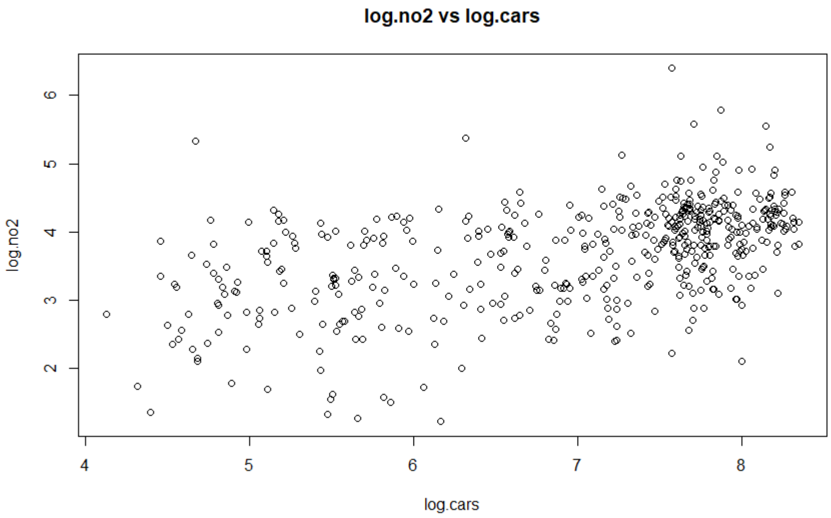


Figure 2: Scatter plot showing the relation between *log.no2* and *log.cars*.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.2331	0.1875	6.57	0.0000
log.cars	0.3535	0.0266	13.30	0.0000
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				

Table 2: Summary of the linear model for $log.no2$ and $log.cars$.

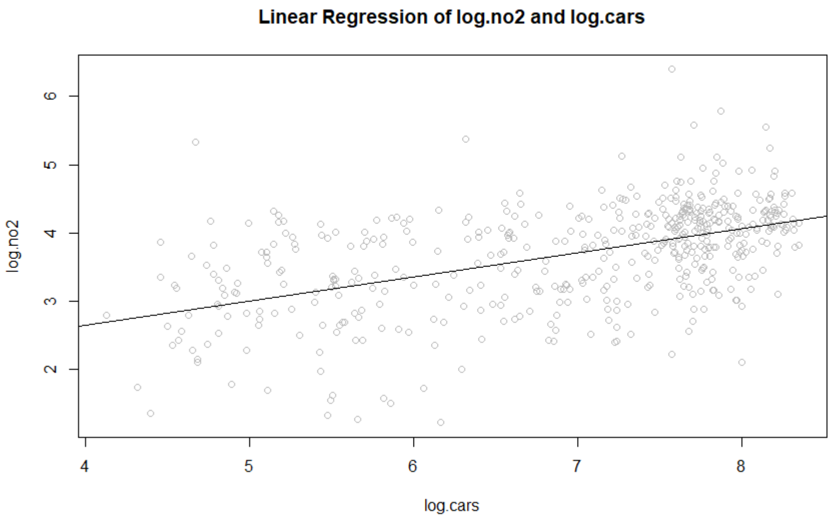


Figure 3: The linear model superimposed on the scatter plot from fig. 2

b)

We are no going to fit a linear model to our $log.no2$ and $log.cars$.

We start by looking at fig. 3. We can see that the line representing the linear model seems to be a good fit, but the residuals are quite large. If we look at the numerical summary in tab. 2 we see that we get the model

$$log.no2 = 1.2331 + 0.3535 \cdot log.cars + \epsilon. \tag{1}$$

This means that given no cars on the road we expect that $log.no2 = 1.2331$, and with each additional (logarithmic) car per hour, $log.no2$ increase by 0.3535. Looking at the p-value of $p \approx 0$ we can conclude that the linear fit is significant. But if we look at the $R^2 = 0.2622$ we see that this value is rather low. This can indicate that the data doesn't have a linear relationship – which our p-value indicates that our data have – or that $log.no2$ isn't described by $log.cars$ alone.

c)

As mentioned above, the there seems to be a bigger spread in the data in fig. 3 for lower values than for higher ones. This is one of the things we need to look at now.

For a linear model

$$y_i = \alpha + \beta_i x_i + \epsilon_i, \tag{2}$$

where ϵ_i is the residuals, we need to have some assumptions.

Linearity

For a linear regression one of the most important assumptions is that the model is linear in the predictor. We are going to check this with a component-plus-residual plot, CPR plot. This is where we plot the partial residuals $\beta_i x_i + r_i$ against the predictor x_i. Since we have a model with only one predictor, this will look like our scatter plot. But R also does a smoothing of the data so that we can look at how it compares to a straight line.

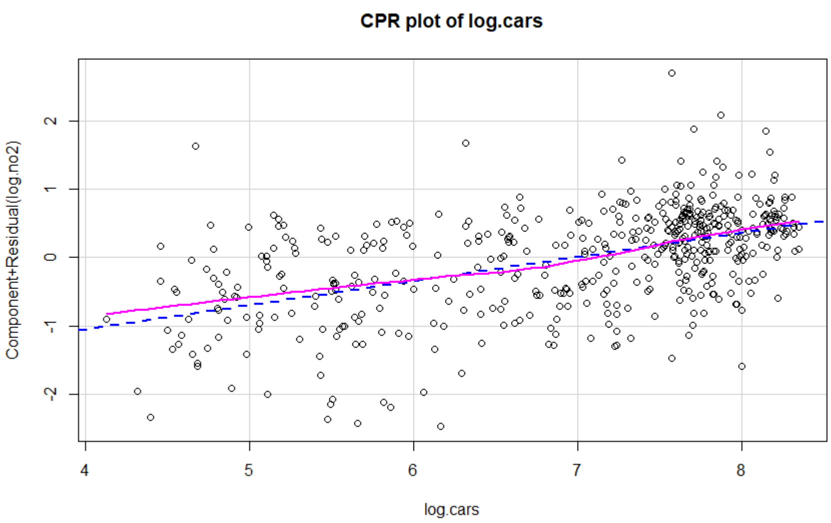


Figure 4: Component-plus-residual plot of $log.cars$.

From fig. 4 we see that the data is somewhat linear. There is some deviation from linearity for the lower values, but the deviation is not so great that we say that the model is non-linear.

Homoscedasticity

For our model we want our residuals to be normally distributed with the same distribution independent of the value of $log.cars$. To check this we are going to plot the standardized residuals – from a R function – against the fitted values. We expect that the residuals there are random around a mean for all the fitted values, meaning that we can fit a straight, horizontal line to the data.

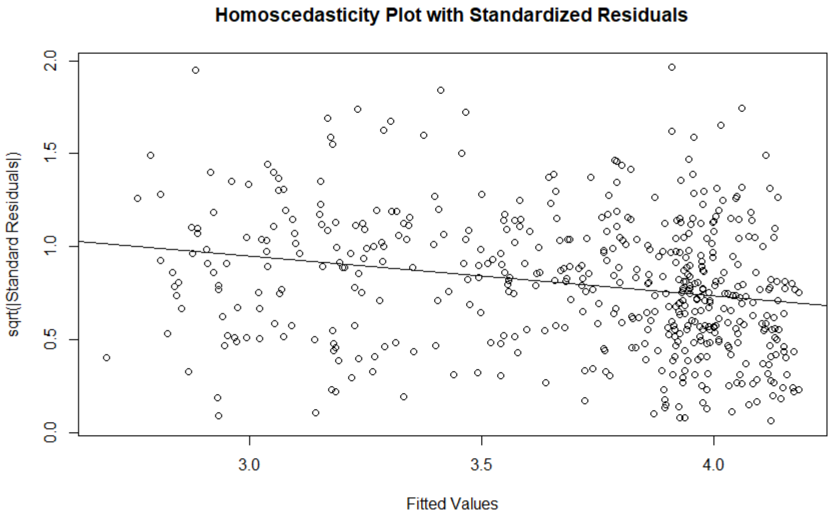


Figure 5: Figure showing $\sqrt{|\text{Standardized Residuals}|}$ against the fitted values.

If we look at fig. 5 we see that the fitted line is not horizontal, but instead decreasing. This indicates that the residuals as smaller for larger values of $log.cars$. This seems to verify what we expected from the start: That there was a larger spread for the smaller values of $log.cars$ than for the larger.

Normality

The last assumption we have is that the residuals are normal.

Looking at fig. 6 we see that the distribution looks more or less normal, but the median is somewhat of center, which indicates skewness. Fig. 7 shows that this is the case. The distribution is far from normal, and seems to be skew towards the negative values. From both these figures we can see that neither the median nor the mean is 0, which is something we would like it to be.

The last nail in the coffin is fig. 8. If the residuals were normally distributed, we would expect a the points to follow the straight line. The S-shape indicates a light tail, and the bend we see in the middle indicates a left skew – which corresponds well with our other plots –. This means that our assumptions that the residuals are normally distributed is false.

So we have seen that while our model is close to linear, the residuals are neither homoscedastic nor normal. So to use a linear model for our data is difficult to justify.

Exercise 2

Add titles to figs

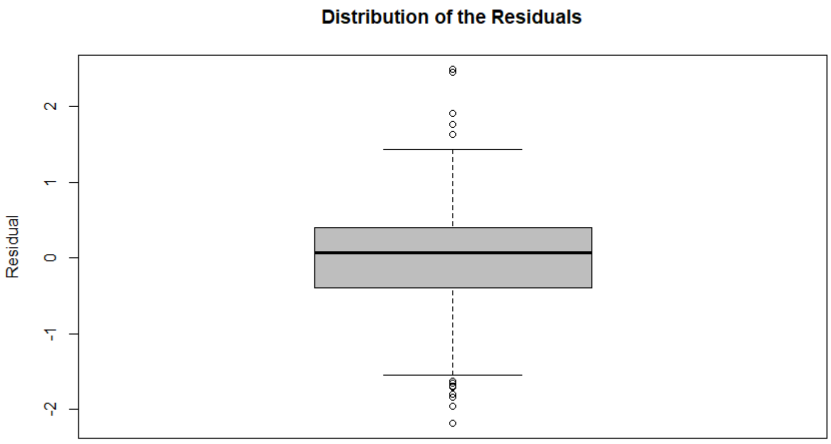


Figure 6: Figure a box plot of the residuals.

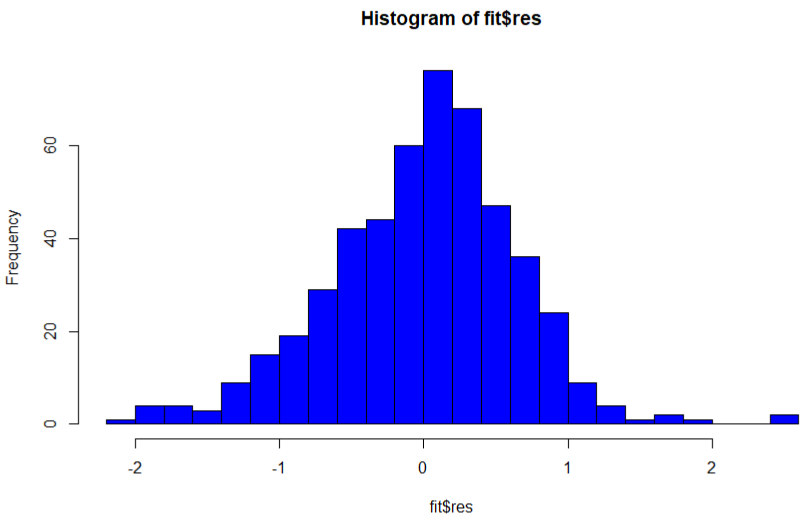


Figure 7: Figure the histogram of the residuals.

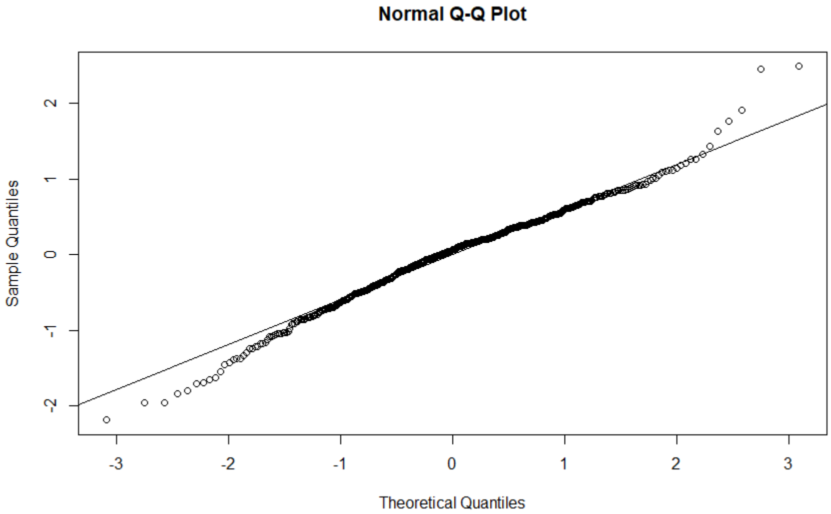


Figure 8: Figure a QQ-plot of the residuals.

a)

The data we are looking at is the systolic blood pressure for 36 people divided into three groups determined by their age. These groups are $30-45$ years, $46-59$ years and $60-75$ years.

Age Group	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
All	36	138.806	25.749	104	117.5	156.2	214
$30-45$ years	12	122.167	15.338	104	112	129	160
$46-59$ years	12	139.083	22.625	108	121.5	157.8	174
$60-75$ years	12	155.167	27.719	110	138	164	214

Table 3: A table showing different measures for the distribution of the systolic blood pressure for the three age groups, and the entire group as a whole. The increase in all the measures for the higher age groups seems to indicate a correlation between age and blood pressure.

If we look at table 3 we see the numerical summary of the blood pressure of the three age groups, and the group as a whole. From the mean column, we see that mean blood pressure seems to increase with increasing age group. This is also the case for all the other measures see in the table. This seems to indicate that there is some relationship between systolic blood pressure and age.

From fig. 9 we see that there is a clear increase in blood pressure with age. Both the median, max and min increases with age. This is exactly what we saw in tab. 3 – a difference is that in the table the mean was given, while in the box plot the median is given –. What we can see clearer from the plot is that the distributions also become wider as age increase. The interquartile width of the middle age group is quite wide, and overlaps with the other groups. This can make this group a bit more difficult to get a significant result from.

This may be just repetition of the caption...

All in all, both the numerical summary, tab. 3, and the box plot, fig. 9, seems to show the same: A increase in systolic blood pressure as age increases.

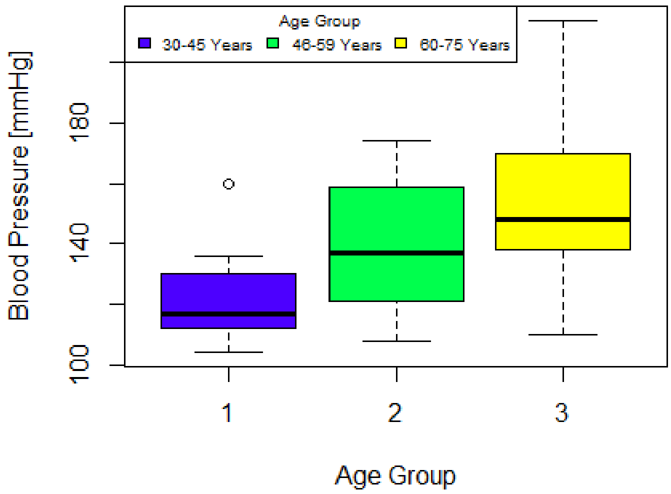


Figure 9: Boxplots showing the distribution of the blood pressure of the three age groups.

b)

We now want to look close at whether blood pressure varies between the different age groups. Above we used qualitative reasoning to to argue that this is the case. Now we want a quantified measure of this. To do this we will use a one-way ANOVA.

For the ANOVA we have the following hypotheses:

- $H_0 : \mu_1 = \mu_2 = \mu_2$
- $H_a : \mu_1 \neq \mu_2 \neq \mu_2$

In words: We have a null hypothesis that the mean blood pressure of all the groups are the same, and an alternative hypothesis that this is not that case. For the ANOVA to be valid, we need to assume that:

- The observations are independent of one another
- The observations in one group are a random sample with a normal distribution $N(\mu_k, \sigma^2)$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Age Group	2	6535.39	3267.69	6.47	0.0043
Residuals	33	16670.25	505.16		

Table 4: A table showing the result of an ANOVA test on the blood pressure dataset. We see that we have a high F-value, leading to a small, significant p-value.

From tab. 4 we see that we get a small p-value of $p = 0.0043$. This is below the $p = 0.05$ mark which is standard to use as the limit of significance. This means that we can throw out that null hypothesis, and conclude that blood pressure indeed varies across the age groups [1]

[1] It is normally bad practice to interpret the p-value in an article, but since this is an oblig, it is done here.

c)

We are now going to try to use a categorical regression on the data set. We are going to use treatment-contrast with group 1, the youngest, as the reference group. This means that our model will look like

$$y_i = \mu_1 + (\mu_2 - \mu_1) \cdot x_{1,i} + (\mu_2 - \mu_1) \cdot x_{2,i} + \epsilon_i, \tag{3}$$

where ϵ_i is a normally distributed error term, μ_j is the mean of the different age groups, and

$$x_{j-1,i} = \begin{cases} 1 \text{ for } i \text{ in group } j \\ 0 \text{ else} \end{cases} . \tag{4}$$

This means that if patient is in group 2, then $x_{1,i} = 1$ and $x_{2,i} = 0$, and vis versa if the patient is in group 3. If the patient is in group 1, both will be 0.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	122.1667	6.4882	18.83	0.0000
Age Group 2	16.9167	9.1757	1.84	0.0742
Age Group 3	33.0000	9.1757	3.60	0.0010
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				

Table 5: The summary of the categorical regression of the blood pressure with respect to the age groups.

In tab. 5 we see the summary of the regression model with the age groups as a categorical predictor variable. The intercept in the table is the mean of the youngest age group $\mu_1 = 122.167$ – our reference group –. The slopes of the two variables can be interpreted as how much the mean of blood pressure increases if you are in that age group, compared to the reference group. E.g. if you are in age group 2, we will expect your blood pressure to be 33.0 mmHg higher than the mean of the youngest group, i.e. it is expected to be 155.167 mmHg.

We can look at the p-values of tab. 5 we see that the p-value for age group 3 is $p = 0.001$. This means that the increase of 33 mmHg is significant. But for group 2 the p-value is just 0.0742, which means that this increase is not significant, so we can not say that we have an 17 mmHg increase compared with the reference. This might have to do with what we see in the box plots 9, where the distribution of the second age group is very wide. The reason might also be due to a non linear dependence on age.

We also see that we have a low $R^2 = 0.2816$, which means that the blood pressure is not fully described by the age groups.

But all in all we see that the we have a good regression model for group 1 and 3, with the change in group 2 not being significant. So there is some significant variation between the age groups, as we saw with the ANOVA.

Look over and rewrte...