STK4900 Mandatory Assignment 1

Olivia Beyer Bruvik

Due 3/12/2021

Problem 1)

```
## read in no2.txt data
pollution_data <- read.table('https://www.uio.no/studier/emner/matnat/math/STK4900/v21/obliger/no2.txt'

## view start of data
head(pollution_data)
```

```
##   log.no2 log.cars temp wind.speed hour.of.day
## 1 3.71844  7.69120  9.2        4.8          20
## 2 3.10009  7.69894  6.4        3.5          14
## 3 3.31419  4.81218 -3.7        0.9           4
## 4 4.38826  6.95177 -7.2        1.7          23
## 5 4.34640  7.51806 -1.3        2.6          11
## 6 4.16044  7.67183  2.6        1.6          19
```

```
## view end of data
tail(pollution_data)
```

```
##     log.no2 log.cars temp wind.speed hour.of.day
## 495 2.56495  4.58497  1.8        2.3           4
## 496 4.30946  7.68202  3.5        5.0          11
## 497 2.94444  6.52942  9.5        6.5          10
## 498 4.17439  7.75791  5.2        4.6          14
## 499 2.95491  5.78996  8.4        0.5           7
## 500 4.03247  8.16223  4.7        5.9          17
```

Problem 1a)

The logarithms of the concentration of NO2 and number of cars have ranges of 5.17131 and 4.22141, and interquartile ranges of 1.0030675 and 1.6173325, respectively.

The data of the number of cars is more spread out than the data of the NO2 concentrations: the mean log-transformed concentrations of NO2 is 3.6983679 with a standard deviation of 0.7505966, whereas the mean log-transformed number of cars per hour is 6.9733421 with a standard deviation of 1.0871664.

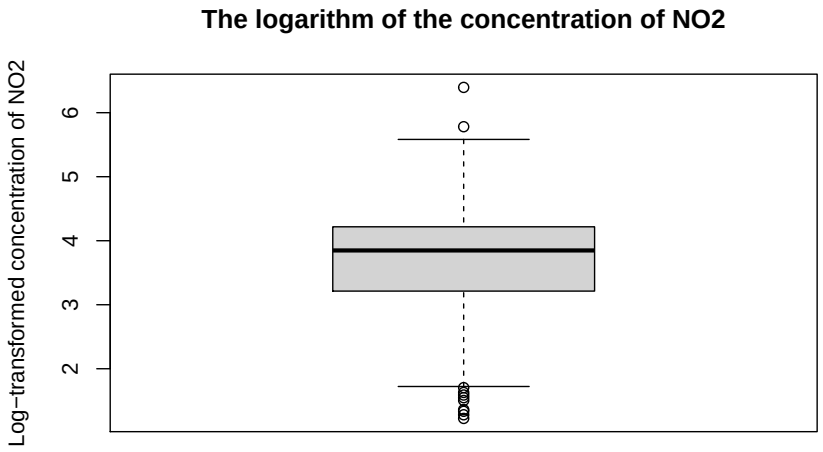
As can be seen in the scatterplot of car frequency vs. NO2 concentration, the log-transformed concentration of NO2 and the log-transformed number of cars per hour have a positive correlation, reflected in the Pearson's

correlation coefficient of 0.5120504. The majority of the data points have high values of both log-transformed concentration of NO2 and log-transformed number of cars per hour.

```
## Main features and visualization of the log.no2 variable
summary(pollution_data$log.no2)
```

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

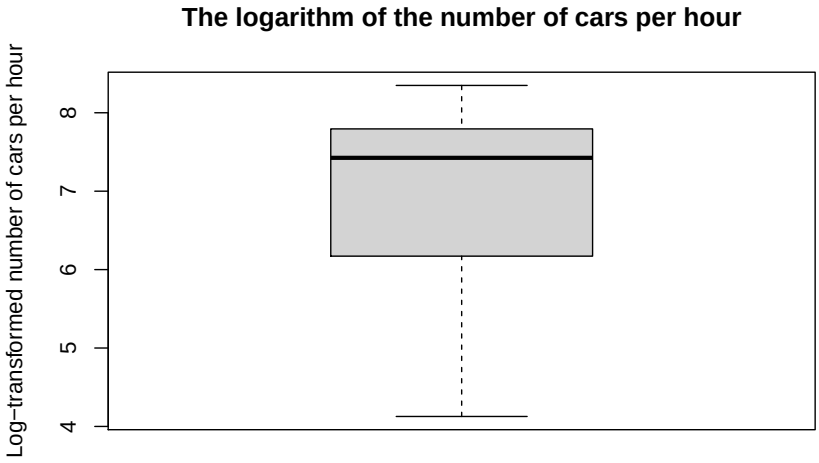
```
boxplot(pollution_data$log.no2,
        main = "The logarithm of the concentration of NO2",
        ylab = "Log-transformed concentration of NO2")
```



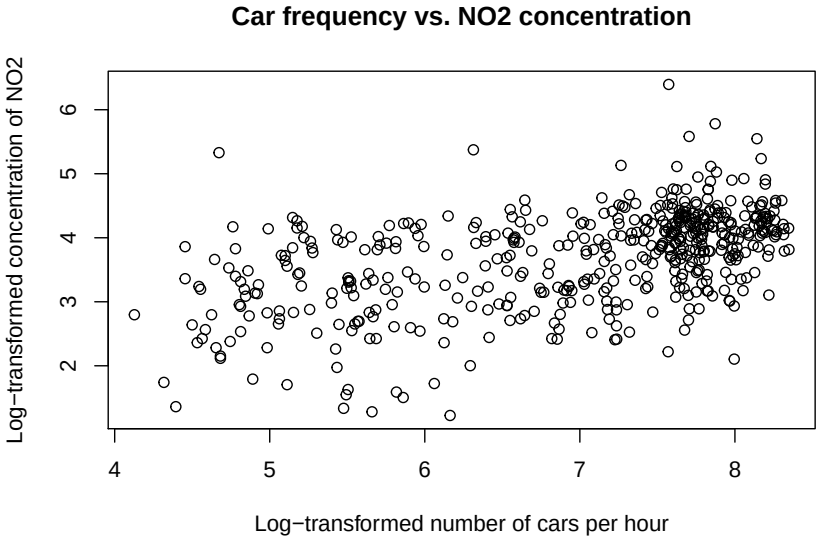

```
## Main features and visualization of the log.cars variable
summary(pollution_data$log.cars)
```

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

```
boxplot(pollution_data$log.cars,
        main = "The logarithm of the number of cars per hour",
        ylab = "Log-transformed number of cars per hour")
```




```
## Scatterplot of log.cars against log.no2
plot(x = pollution_data$log.cars,
     y = pollution_data$log.no2,
     xlab = "Log-transformed number of cars per hour",
     ylab = "Log-transformed concentration of NO2",
     main = "Car frequency vs. NO2 concentration")
```




```
## Pearson's correlation of log.cars and log.no2
cor(pollution_data$log.no2, pollution_data$log.cars)
```

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

Problem 1b)

I fitted a simple linear model with log.no2 as the outcome and log.cars as the explanatory variable. The simple linear model is summarized and shown on the scatterplot of log.cars against log.no2 below.

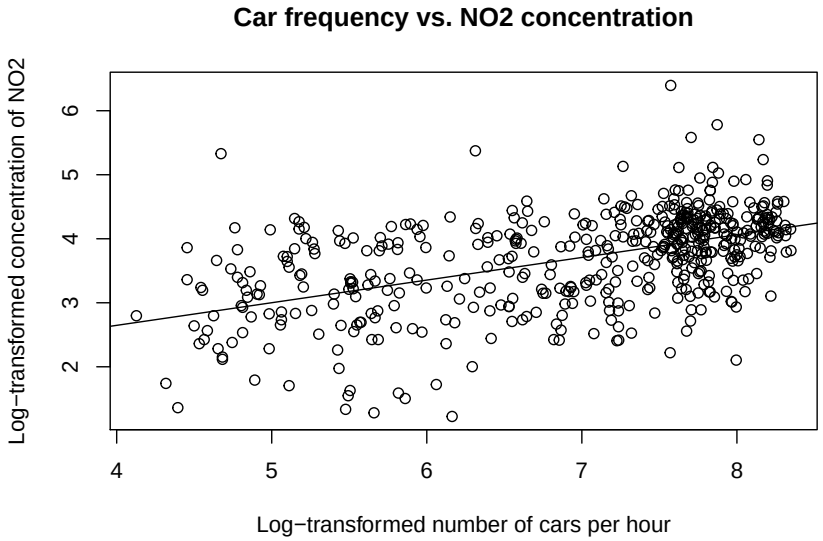
The estimated intercept and coefficient for log.cars are 1.233 and 0.354, respectively. The estimated intercept of 1.233 can be interpreted as the log-transformed concentration of NO2 when there is no traffic, measured by the log-transformed number of cars per hour. Conversely, the estimated log.cars coefficient of 0.354 can be interpreted as the average increase of log-transformed concentration of NO2 per unit increase in the log-transformed number of cars per hour, holding all else constant.

The coefficient of determination, or the multiple R-squared measure, is 0.262. This value reports that this model accounts for a proportion of 0.262 of the total variability in the log-transformed concentration of NO2, suggesting that the model has room for improvement.

```
## Simple linear model of log concentration of No2 explained by amount of traffic
fit.no2_traffic <- lm(log.no2 ~ log.cars, data = pollution_data)
summary(fit.no2_traffic)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars, data = pollution_data)
##
## 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
```

```
## Scatterplot of log.cars against log.no2 with fitted line
plot(x = pollution_data$log.cars,
     y = pollution_data$log.no2,
     xlab = "Log-transformed number of cars per hour",
     ylab = "Log-transformed concentration of NO2",
     main = "Car frequency vs. NO2 concentration",
     abline(fit.no2_traffic))
```



Problem 1c)

The various residual plots below suggest that the model assumptions are reasonable.

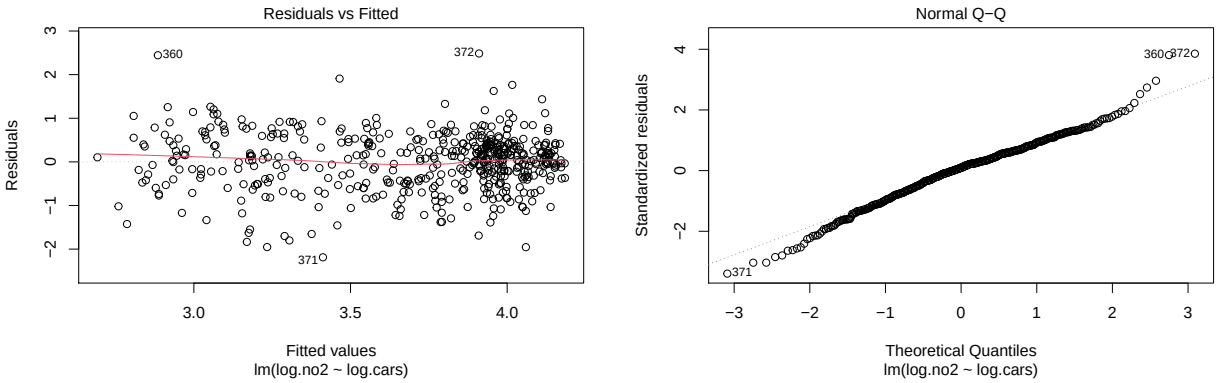
First, the Residuals vs. Fitted plot shows homoscedasticity, hence indicating that the true relationship is close to linear because the red line has very little curvature and is centered at Residuals = 0.

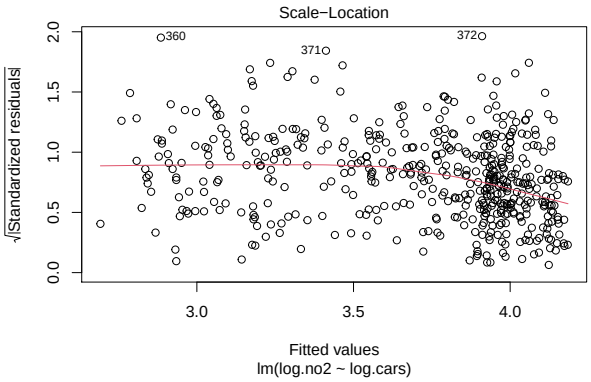
The scale-location plot confirms this with the relatively horizontal red fitted line; however the line does tilt down towards larger fitted values, possibly because of the higher frequency of data points at high values, as discussed above.

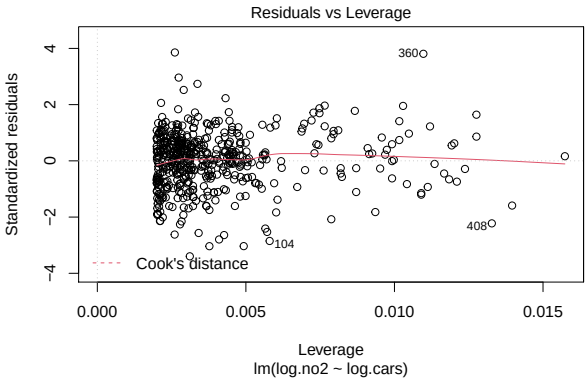
In the quantile-quantile plot, the standardized residuals follow the straight dashed line, indicating that the residuals are approximately normally distributed.

In the Residuals vs. Leverage plot, the datapoints labeled 104, 408 and 360 indicate datapoints with an absolute standardized residuals value above 2, suggesting that these datapoints may possibly be outliers and can be influential in the model.

```
## various residual plots to judge model assumptions
plot(fit.no2_traffic)
```







Problem 1d)

I fit various multiple regression models with log.no2 as the outcome, and the four other variables as explanatory. I found the 'best' model by looking at the coefficient of determination for each model. The model with the highest coefficient of determination (r^2 = 0.4833) that I could find included all four explanatory variables, but I log-transformed wind.speed.

The other models that I tested and their respective coefficients of determination can be found in the appendix.

```
## multiple regression with log-transformed wind.speed
fit.multiple = lm(log.no2 ~ log.cars + temp + log(wind.speed) + hour.of.day, data = pollution_data)
summary(fit.multiple)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + log(wind.speed) + hour.of.day,
##     data = pollution_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18068 -0.31840  0.03765  0.33429  1.81757
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.070868   0.170984   6.263 8.21e-10 ***
## log.cars         0.457188   0.027931  16.369  < 2e-16 ***
## temp            -0.026724   0.003837  -6.964 1.06e-11 ***
## log(wind.speed) -0.419388   0.036362 -11.534  < 2e-16 ***
## hour.of.day     -0.012297   0.004374  -2.811  0.00513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5417 on 495 degrees of freedom
## Multiple R-squared:  0.4833, Adjusted R-squared:  0.4791
## F-statistic: 115.7 on 4 and 495 DF,  p-value: < 2.2e-16
```

Problem 1e)

I fitted a multiple regression model with log.no2 as the outcome and log.cars, temp, hour.of.day and log(wind.speed) as the explanatory variables.

The estimated intercept is 1.071, indicating that the log-transformed concentration of NO2 is 1.071 when there is no traffic, a temperature of zero degrees and no wind at midnight.

The estimated coefficient for each explanatory variable can be interpreted as the average change of log-transformed concentration of NO2 per unit increase for the particular explanatory variable, holding all else constant.

The coefficients for temp, log(wind.speed) and hour.of.day are negative, suggesting that the log-transformed concentration decreases as these variables increase. Conversely, the positive log.cars coefficient demonstrates the positive relationship between log.cars and log.no2.

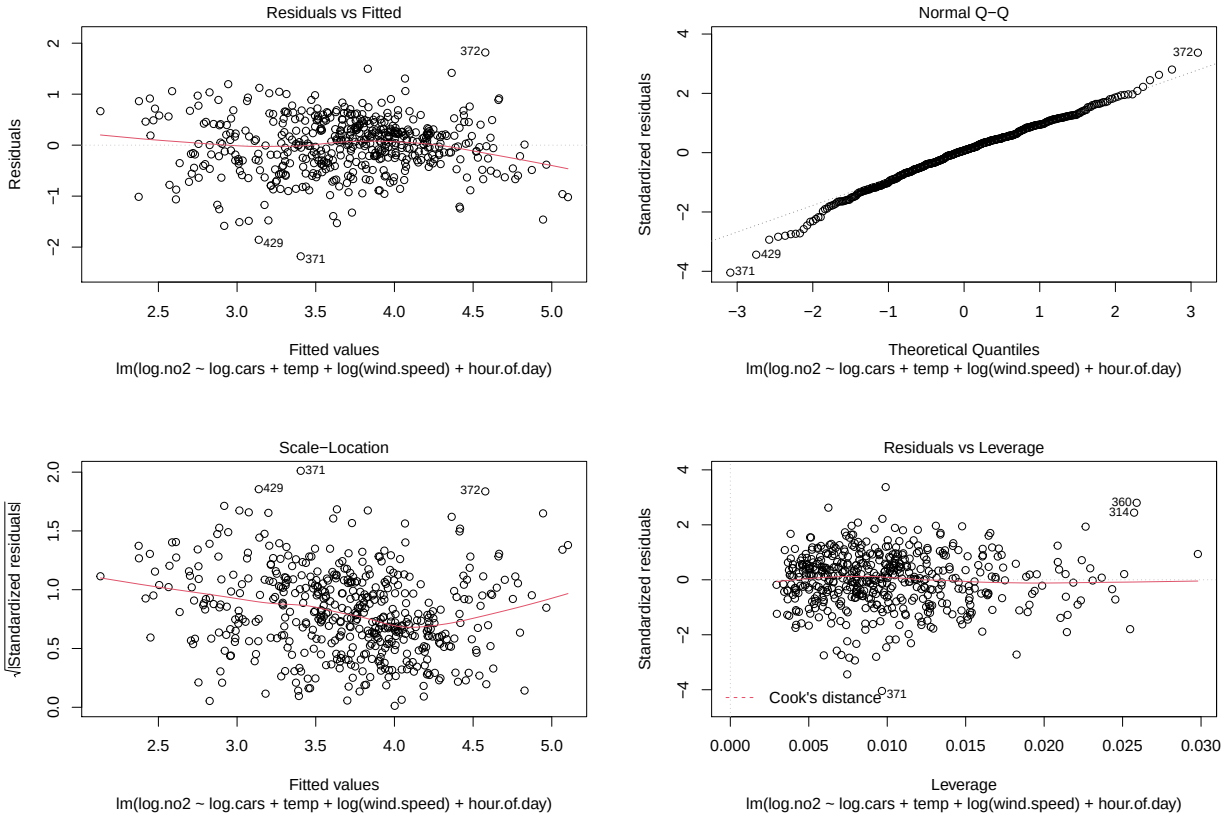
The coefficient of determination, or the multiple R-squared measure, is 0.262. This value reports that this model accounts for a proportion of 0.262 of the total variability in the log-transformed concentration of NO2, suggesting that the model has room for improvement.

The various residual plots below suggest that the model assumptions are reasonable, because the plots demonstrate homoscedasticity in the data and a general normal distribution of errors.

```
## multiple regression with log-transformed wind.speed
fit.multiple = lm(log.no2 ~ log.cars + temp + log(wind.speed) + hour.of.day, data = pollution_data)
summary(fit.multiple)
```

```
##
## Call:
## lm(formula = log.no2 ~ log.cars + temp + log(wind.speed) + hour.of.day,
##     data = pollution_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.18068 -0.31840  0.03765  0.33429  1.81757
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1.070868   0.170984   6.263 8.21e-10 ***
## log.cars         0.457188   0.027931  16.369  < 2e-16 ***
## temp            -0.026724   0.003837  -6.964 1.06e-11 ***
## log(wind.speed) -0.419388   0.036362 -11.534  < 2e-16 ***
## hour.of.day     -0.012297   0.004374  -2.811  0.00513 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5417 on 495 degrees of freedom
## Multiple R-squared:  0.4833, Adjusted R-squared:  0.4791
## F-statistic: 115.7 on 4 and 495 DF,  p-value: < 2.2e-16
```

```
## various plots to judge model assumptions
plot(fit.multiple)
```



Problem 2

```
## read in blood.txt data
bp_data <- read.table('https://www.uio.no/studier/emner/matnat/math/STK4900/v21/obliger/blood.txt', hea

## view start of data
head(bp_data)
```

```
##   Bloodpr age
## 1     128   1
## 2     104   1
## 3     132   1
## 4     112   1
## 5     136   1
## 6     124   1
```

```
## view end of data
tail(bp_data)
```

```
##    Bloodpr age
## 31     188   3
## 32     158   3
```

```
## 33     182   3
## 34     148   3
## 35     138   3
## 36     136   3
```

```r
# Define the age groups as factors (categorical):
bp_data$age <- factor(bp_data$age)
```

Problem 2a)

The blood pressure measurements have ranges of 56, 66 and 104 for age groups 1, 2 and 3, respectively.

As can be seen in the boxplot, the mean bloodpressure increases as the age group, and by extension ages, increases.

The bloodpressure data is also more spread out in higher age groups: the standard deviations for age group 1, 2 and 3 are 15.3376147, 22.6252407 and 27.7188263.

```r
## Main features and visualization of the data
# boxplot
boxplot(Bloodpr ~ age,
        data = bp_data,
        main = "Blood pressure for each age group",
        xlab = "Age group",
        ylab = "Blood pressure")
```

```r
# Summary all age groups
summary(bp_data)
```

```
##     Bloodpr      age
##  Min.   :104.0   1:12
##  1st Qu.:117.5   2:12
##  Median :136.0   3:12
##  Mean   :138.8
##  3rd Qu.:156.2
##  Max.   :214.0
```

```r
# Summary age group 1
summary(bp_data$Bloodpr[bp_data$age==1])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   104.0   112.0   117.0   122.2   129.0   160.0
```

```r
# Summary age group 2
summary(bp_data$Bloodpr[bp_data$age==2])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   108.0   121.5   137.0   139.1   157.8   174.0
```

```r
# Summary age group 3
summary(bp_data$Bloodpr[bp_data$age==3])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   110.0   138.0   148.0   155.2   164.0   214.0
```

Problem 2b)

I ran a one-way ANOVA test below, with blood pressure as the outcome and age group as the explanatory variable. The assumptions involved in this test primarily involves that the data is normally distributed and that the age groups represent random samples. Furthermore, observations are assumed to be independent of each other.

I am testing the null hypothesis that the mean blood pressure for each age group are all equal, and the alternative hypothesis that the mean blood pressure for each age group are not all equal.

```r
# One-way ANOVA test to see how blood pressure varies across age groups.
aov.bp <- aov(Bloodpr~age, data = bp_data)
anova(aov.bp)
```

```
## Analysis of Variance Table
##
## Response: Bloodpr
##           Df  Sum Sq Mean Sq F value   Pr(>F)
## age        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
```

Problem 2c)

Here, I formulated a regression model with age group as a categorical predictor variable. I used treatment contrast method and by default, the youngest age group was the reference.

The results suggest that the null hypothesis can be rejected because of the large F value and t value in the one-way anova test and regression model.

The blood pressure also evidently increases as the age group increases, evident in the coefficients.

```
# Regression with categorical predictor variables
lm.bp <- lm(Bloodpr~age, data = bp_data)
summary(lm.bp)
```

```
##
## Call:
## lm(formula = Bloodpr ~ age, data = bp_data)
##
## 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
```

Appendix

Problem 1d)

```
## coefficient of determination of other multiple regression models
summary(lm(log.no2 ~ log.cars + temp + wind.speed + hour.of.day, data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.465836
```

```
summary(lm(log.no2 ~ log.cars + log(temp + 0.000001) + wind.speed + hour.of.day, data = pollution_data))
```

```
## $r.squared
## [1] 0.43038
```

```r
summary(lm(log.no2 ~ log.cars + log(temp + 0.000001) + log(wind.speed) + hour.of.day, data = pollution_
```

```
## $r.squared
## [1] 0.4581861
```

```r
summary(lm(log.no2 ~ log.cars + I(temp^2) + log(wind.speed) + hour.of.day, data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.4339509
```

```r
summary(lm(log.no2 ~ log.cars + temp + I(wind.speed^2) + hour.of.day, data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.4223266
```

```r
summary(lm(log.no2 ~ log.cars + temp + log(wind.speed) + log(hour.of.day), data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.4807469
```

```r
summary(lm(log.no2 ~ log.cars + temp + I(wind.speed^2) + hour.of.day, data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.4223266
```

```r
pollution_data$temp_K <- pollution_data$temp - 273
summary(lm(log.no2 ~ log.cars + temp + I(wind.speed^2) + hour.of.day, data = pollution_data))[8]
```

```
## $r.squared
## [1] 0.4223266
```