Report Template coursework assignment A - 2018

CS4125 Seminar Research Methodology for Data Science

Nathan Buskulic (4947916), Mitchell Deen(4396340),Þórunn Arna ÓmardÓttir (4917499)

4/3/2019

Contents

1 Part 1 - Design and set-up of true experiment

1.1 The motivation for the planned research.

The coffee is today the most consumed drink in the world and it is told to increase your performance and concentration. We want to challenge this idea and verify scientifically if this is a valid idea. We want to test how caffeine consumption affects the result of an IQ test. We are most interested and seeing what the affect is on TU Delft students like ourselves. So the participants will be recruited from the TU Delft student body.

1.2 The theory underlying the research.

There is a large body of literature available on the effects of caffeine on the performance in cognitive tasks. Literature generally supports the idea that coffee improves this performance, see e.g. (Jarvis, 1993; Nehlig, 2010; Rogers et al., 2008). In a brief survey of the relevant literature we did not find any studies specifically addressing students. We would like to investigate this part of the population in more detail.

Jarvis, M. J. (1993). Does caffeine intake enhance absolute levels of cognitive performance?. Psychopharmacology, 110(1-2), 45-52. Rogers, P. J., Smith, J. E., Heatherley, S. V., & Pleydell-Pearce, C. W. (2008). Time for tea: mood, blood pressure and cognitive performance effects of caffeine and theanine administered alone and together. Psychopharmacology, 195(4), 569. Nehlig, A. (2010). Is caffeine a cognitive enhancer?. Journal of Alzheimer's Disease, 20(s1), S85-S94.

1.3 Research questions

The research question that will be examined in the experiment (or alternatively the hypothesis that will be tested in the experiment)is: How does caffeine level in coffee increases IQ test score ?

1.4 The related conceptual model

The conceptual model can be found below. In the model the caffiene consumption is a independent variable, the IQ test score is a dependent variable, the sleepiness feeling of a person is a mediation vaiable and the coffee consumption habits is a moderating variable.

1.5 Experimental Design

The experiment is a four groups, post test only, randomized controlled trail. Participants will be separated randomly in 4 different groups that will get different amount of caffeine, after a certain amount of time they will then perform a test and get a score on this test.

1.6 Experimental procedure

The participants will be separated into four groups randomly. One group will do the IQ test without any prior coffee consumption while the three other groups will do the test half an hour after coffe consumption where the coffee will have different levels of caffeine according to the groups (20, 100 200 mg). This will allow us to measure the general impact of drinking coffee on an IQ test but it will also allow us to test the difference between each caffeine level.

1.7 Measures

The Coffe consumption will be measured in ml. The perfomance in an IQ test will be a simple integer number on the scale from 0-200 where the mean is around 100. Sleepingness will be given by the participants on the

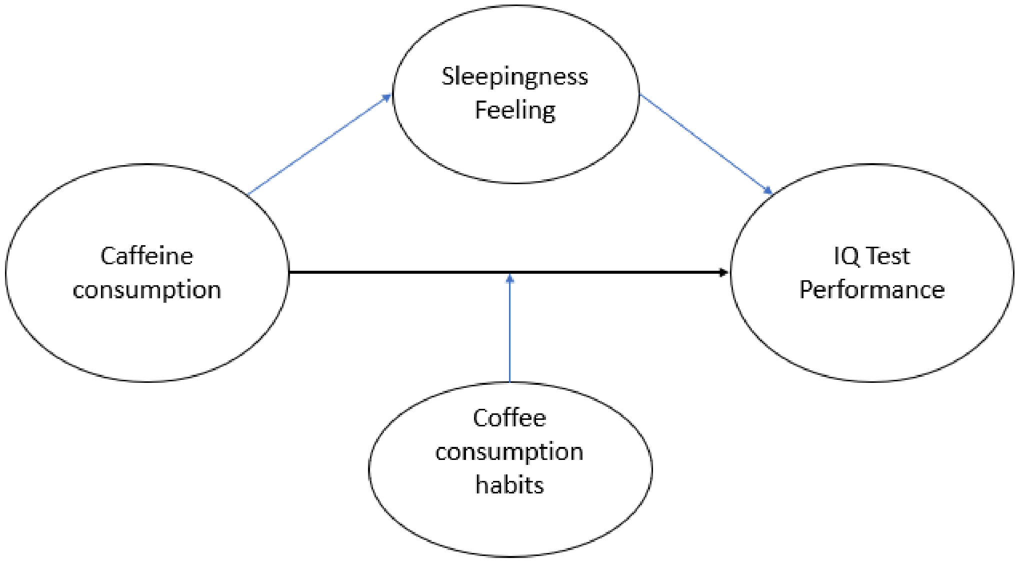


Figure 1: Caffeine consumption Conceptual model

scale from 0-10 where 10 means the highest level of sleepingness. The amount of caffeine will be measured in mg. Prior coffeedrinking habits will be given by participants. They will be asked how much coffee they typically drink on a normal day. The amount of coffe will be transformed to a scale from 0-10 where 0 is no caffiene consumption and 10 is the highest level of caffiene consumption.

1.8 Participants

Since we are just going to make this experiment on the effects of coffee consumption on students at TU Delft we need to find participants from that group of people. Emails will be sent out to the student body explaining the theory of the experiments and willing volunteers asked to fill in a form. We will try to contact an external company of some sort to get some credit or coupons that we can give to participants as a reward for helping out.

1.9 Suggested statistical analyses

We will use a one way Analysis of Variance (ANOVA) test between groups. Indeed, since the IQ test is designed to follow a gaussian distribution, we just want to compare the mean of each group.

2 Part 2 - Generalized linear models

2.1 Question 1 Twitter sentiment analysis (Between groups - single factor)

2.1.1 Collecting tweets, and data preparation

We collected Tweets for the three celebrities Beyonce, Madonna and Mickael Jackson. The code can be found in the markdown file.

2.1.2 Conceptual model

Here below we can see the conceptual model for our research question: Is there a difference in the sentiment of the tweets related to the different celebrities?

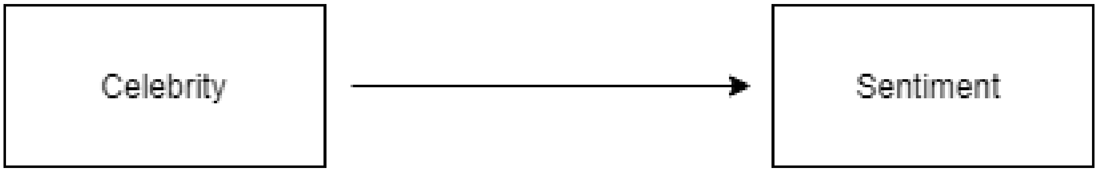


Figure 2: Celebrity sentiment Conceptual model

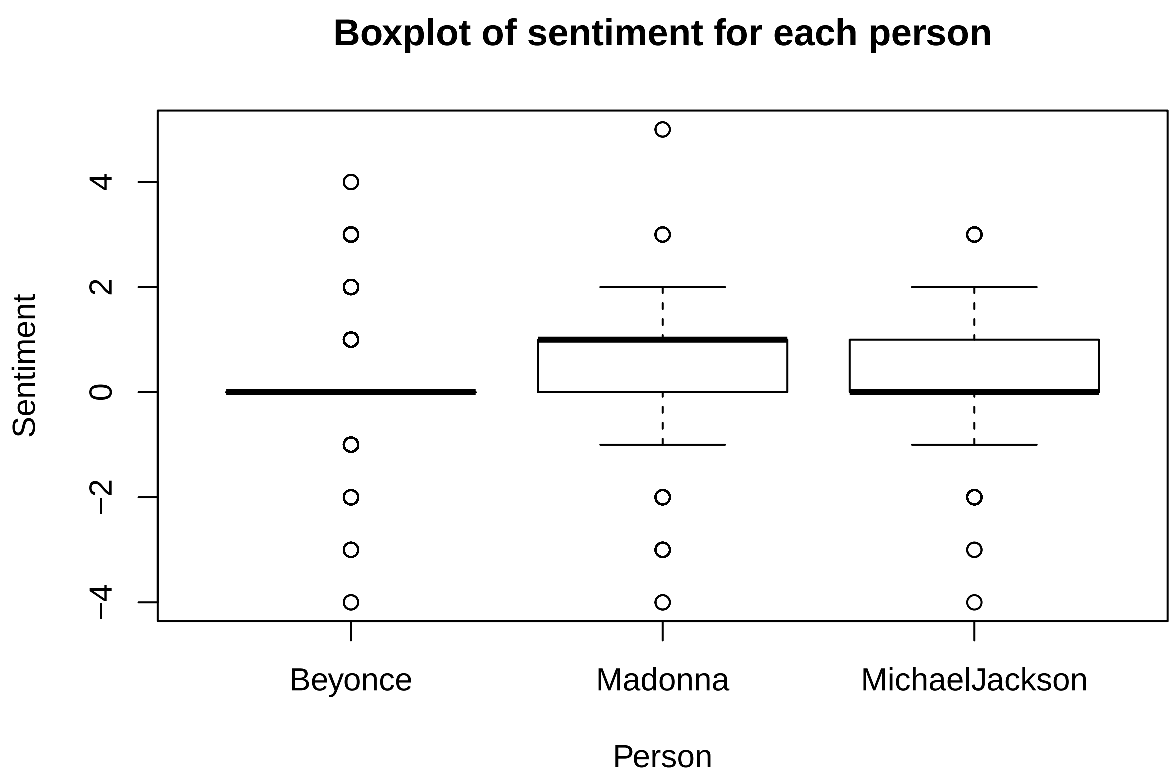
We can see that the sentiment of tweets related to different celebrity is directly connected to the celebrity itself. Therefor the conceptual model is very simple consisting of two variables, "Celebrity" and "Sentiment".

2.1.3 Homogeneity of variance analysis

Here we will analyze the homogeneity of variance of sentiments of the tweets of the different celebrities.

Lets start by looking at how the boxplot looks for each person and the relevant sentiment score. We can already see from looking at the boxplot that the variance does not seem to be the same for all celibirities. Madonna seems to have the broadest spectrum and the median line hits a bit different places depending on the celebrities.

```
#this was not here in the intermediate report.
#include your code and output in the document
boxplot(score ~ Person, data=semFrame, main="Boxplot of sentiment for each person",
        xlab="Person", ylab="Sentiment")
```



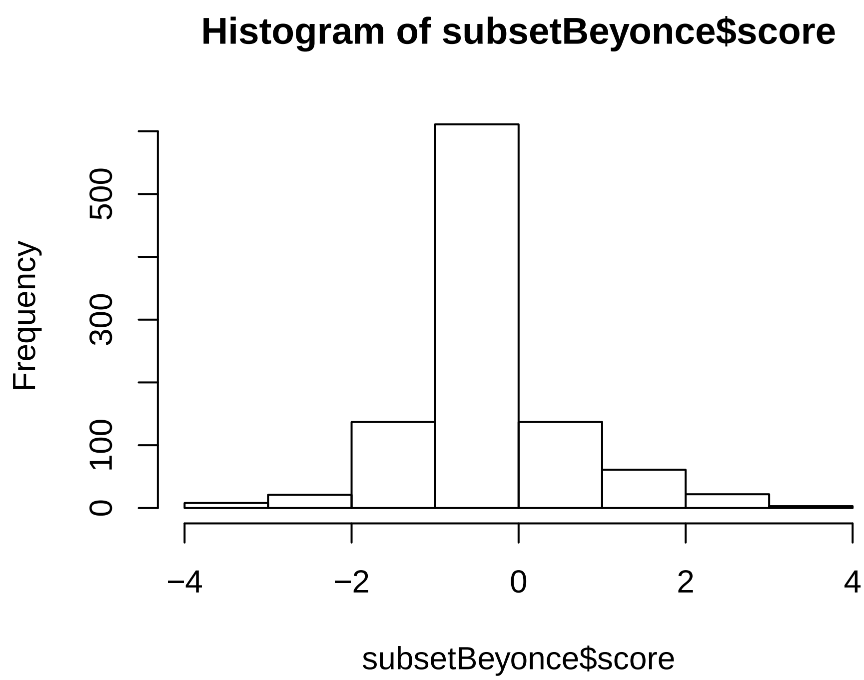

```
levene = leveneTest( semFrame$score,semFrame$Person, center = median)
```

The Levene test results in a very low p-value (P=0.000029). Therefore the hypothesis of equal variances is rejected and it is concluded that there is a difference between the variances in the population. Therefore the variance is not considered to be homogeneous.

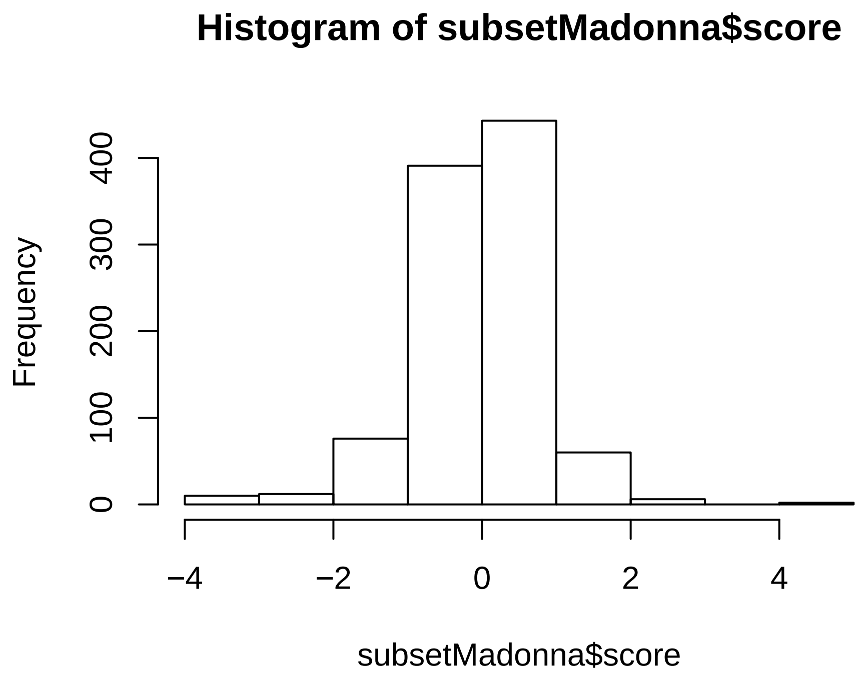
2.1.4 Visual inspection

Looking at the figures here below we see that the sentiment scores for all the celebrities follow a very similar distribution that looks a lot like a normal distribution. But by inspecting the histograms we can see that the distribution is not entirely the same. Therefore we will do a furter inspection to see how the distributions differ from each other.

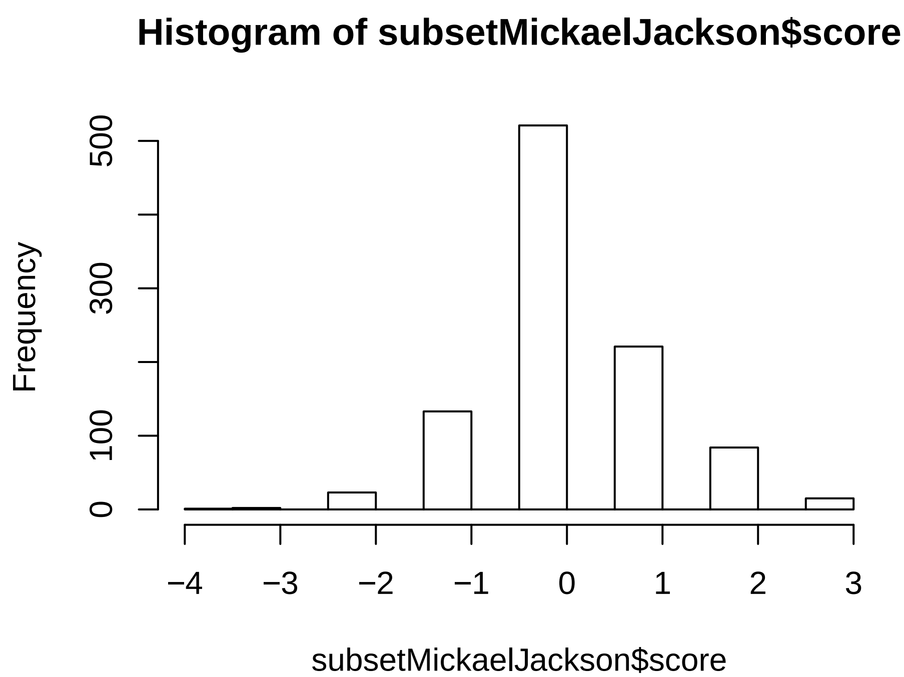
```
#include your code and output in the document

hist(subsetBeyonce$score)
```




```
hist(subsetMadonna$score)
```




```
hist(subsetMickaelJackson$score)
```



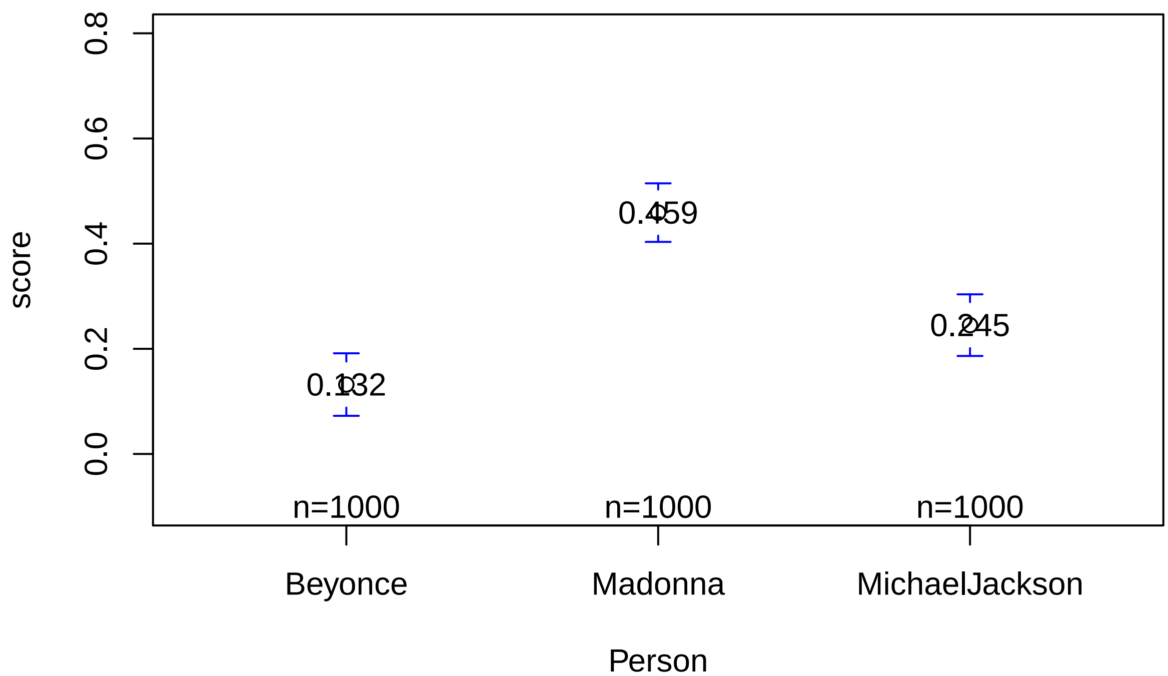

```
meanBeyonce<-mean(subsetBeyonce$score)
meanMadonna<-mean(subsetMadonna$score)
meanMickaelJackson<-mean(subsetMickaelJackson$score)
```

```r
stdBeyonce<-sd(subsetBeyonce$score)
stdMadonna<-sd(subsetMadonna$score)
stdMickaelJackson<-sd(subsetMickaelJackson$score)
```

2.1.5 Mean sentiments

Here below we plot the means of each class using plotmeans from the package gplots. We can see that the mean for Beyonce is 0.132, for Madonna is 0.459 and for Mickael Jackson it is 0.245. Where a lower value means that the sentiment analysis is more negative. Just by looking at the means we see that they are considerably far from each other. As well we have te standard deviations 0.958, 0.897, 0.944 in the same respective order as before.

```r
plotmeans(score ~ Person, data = semFrame,mean.labels = TRUE,connect = FALSE,ylim = c(-0.1,0.8))
```



2.1.6 Linear model

```r
#include your code and output in the document
 model0<- lm(score ~ 1, data = semFrame) #model without predictor
 model1<- lm(score ~ Person, data = semFrame) #model with predictor
 AnovaResults <-anova(model0,model1)
```

Here we have tied two different linear models to fit put outcome. Model0 is a simple model where the sentiment score stands alone, but model1 has the celebrity in there as well. The calculated f-value, $F(2,2997)$ is 31.640 and the p- value is very small and well below .00001. Since the p-value is so small it indicates that the sentiment of tweets is significantly different depending on what celebrity is mentioned in the tweet.

2.1.7 Post Hoc analysis

Now a post-hoc analysis is performed to examine which of the tweets differ from othere celebrity tweets

```
#include your code and output in the document
BonferroniResults <- pairwise.t.test(semFrame$score, semFrame$Person, paired = FALSE, p.adjust.method =
BonferroniP <- BonferroniResults$p.value
BonferroniP
```

```
##                    Beyonce      Madonna
## Madonna        1.975478e-14           NA
## MichaelJackson 2.050576e-02 9.449806e-07
```

We chose to use the Bonferroni correction to conduct this post-hoc analysis. There the p-values are multiplied bu the number of comparisons.

According to results here above all of the celebrity pair comparisons have a low p-value that indicates again that the sentiment is dependent on what celebrity it is in relation to.

2.1.8 Report section for a scientific publication

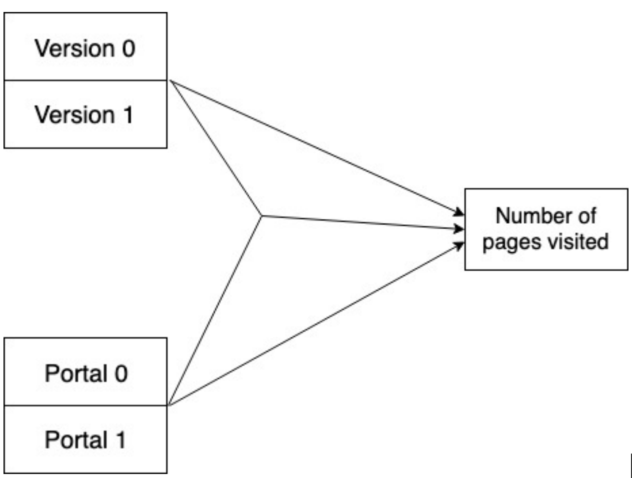
We analysed sentiment scores of tweets for thee dfferent celebrities. We had scores for Madonna (M=0.459, SD=0.897), Beyonce (M=0.132, SD=0.958) and Micheal Jackson (M=0.245, SD=0.944). The means and standard deviations are different between celebrities so they where inspected further.

A linear model was fitted on the number of the sentiment score, comparing the difference when taking the relative celbrity in account and not. We first conducted an Anova test and obtained the results (F(2,2997) = 31.65, p <<.0001) which states a significant difference in the sentiment, depending on which celebrity it is for. Then A Post Hoc analysis by the means of Bonferroni was conducted. There we again got p -values(<<.0001 ,0.02 ,<<.0001) that show us that the difference of scores is signifant depeding on celebrity.

2.2 Question 2 - Website visits (between groups - Two factors)

2.2.1 Conceptual model

The model can be found in the figure below.



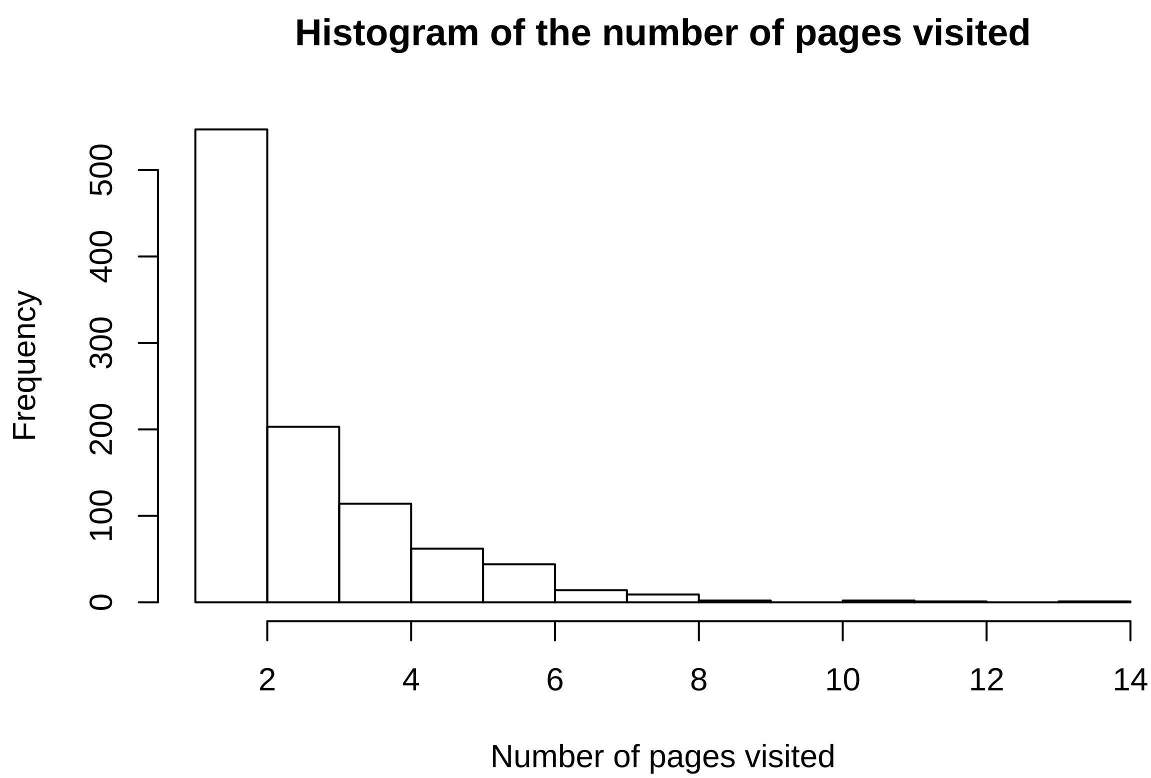
[]

2.2.2 Visual inspection

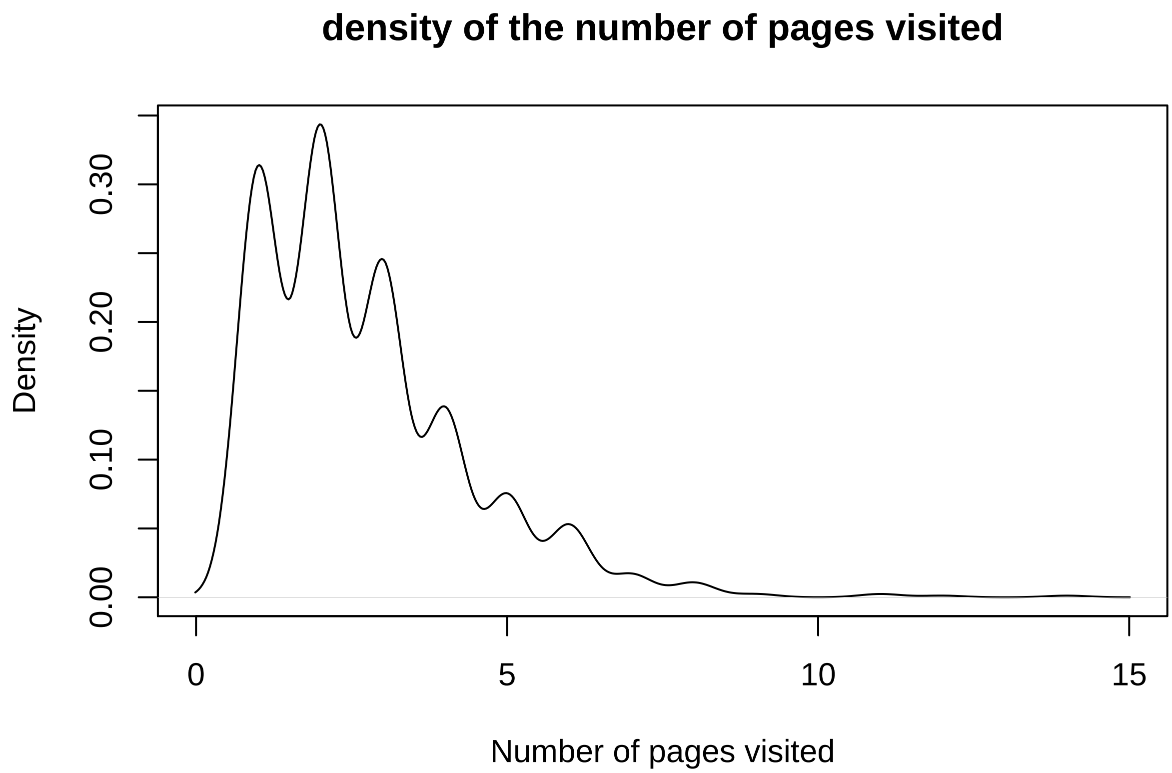
We will first examine the data and look at the distributions.

```
myData <- read.csv("webvisita.csv",header=TRUE)
# We transform into factors what need to be.
```

```r
myData$user <- factor(myData$user)
myData$version <- factor(myData$version, levels=c(0:1), labels=c("old","new"))
myData$portal <- factor(myData$portal, levels=c(0:1),labels=c("consummer","company"))

hist(myData$pages, xlab="Number of pages visited", main = "Histogram of the number of pages visited")
```




```r
plot(density(myData$pages), xlab="Number of pages visited", main = "density of the number of pages visi
```



It appears that the data do not look like normally distributed, We will provide more analysis to understand these distributions.

2.2.3 Normality check

We can see that the data does not seems to come from a normal distribution, thus we will do a normality test.

```
shapiro.test(myData$pages)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  myData$pages
## W = 0.8436, p-value < 2.2e-16
```

The really small p-value indcates here that there is a high probability that this data do not come from a normal distribution. From what we can visually see it seems that the data comes from a Poisson distribution, which is an important information for the model analysis.

2.2.4 Model analysis

Since the data are not normally distributed, we cannot use a simple linear model which assume the normality of the data. Thus we will fit generalized linear model with a poisson distribution assumption.

```
# We create all the different models
model0 <- glm(pages ~ 1, data=myData, family="poisson")
model1 <- glm(pages ~ version, data=myData, family="poisson")
model2 <- glm(pages ~ portal, data=myData, family="poisson")
```

```
model3 <- glm(pages ~ version + portal, data=myData, family="poisson")
model4 <- glm(pages ~ version + portal + version:portal, data = myData, family="poisson")
```

Since we are using generalized models, we cannot use an F-test and we decided to use a Chi-Square test instead in our analysis.

```
pander(anova(model0,model1,test="Chisq"),caption = "Version as main effect on the number of pages visit
```

Table 1: Version as main effect on the number of pages visited

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
998	938.6	NA	NA	NA
997	931.2	1	7.36	0.006671

```
pander(anova(model0,model2,test="Chisq"),caption = "Portal as main effect on the number of pages visite
```

Table 2: Portal as main effect on the number of pages visited

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
998	938.6	NA	NA	NA
997	878.2	1	60.39	7.793e-15

```
pander(anova(model3,model4,test="Chisq"),caption = "Interaction effect on top of the two main effect")
```

Table 3: Interaction effect on top of the two main effect

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
996	872.4	NA	NA	NA
995	843.5	1	28.98	7.318e-08

```
pander(anova(model4, test="Chisq"),caption = "Effect of version, portal and interaction effect on the n
```

Table 4: Effect of version, portal and interaction effect on the number of pages visited

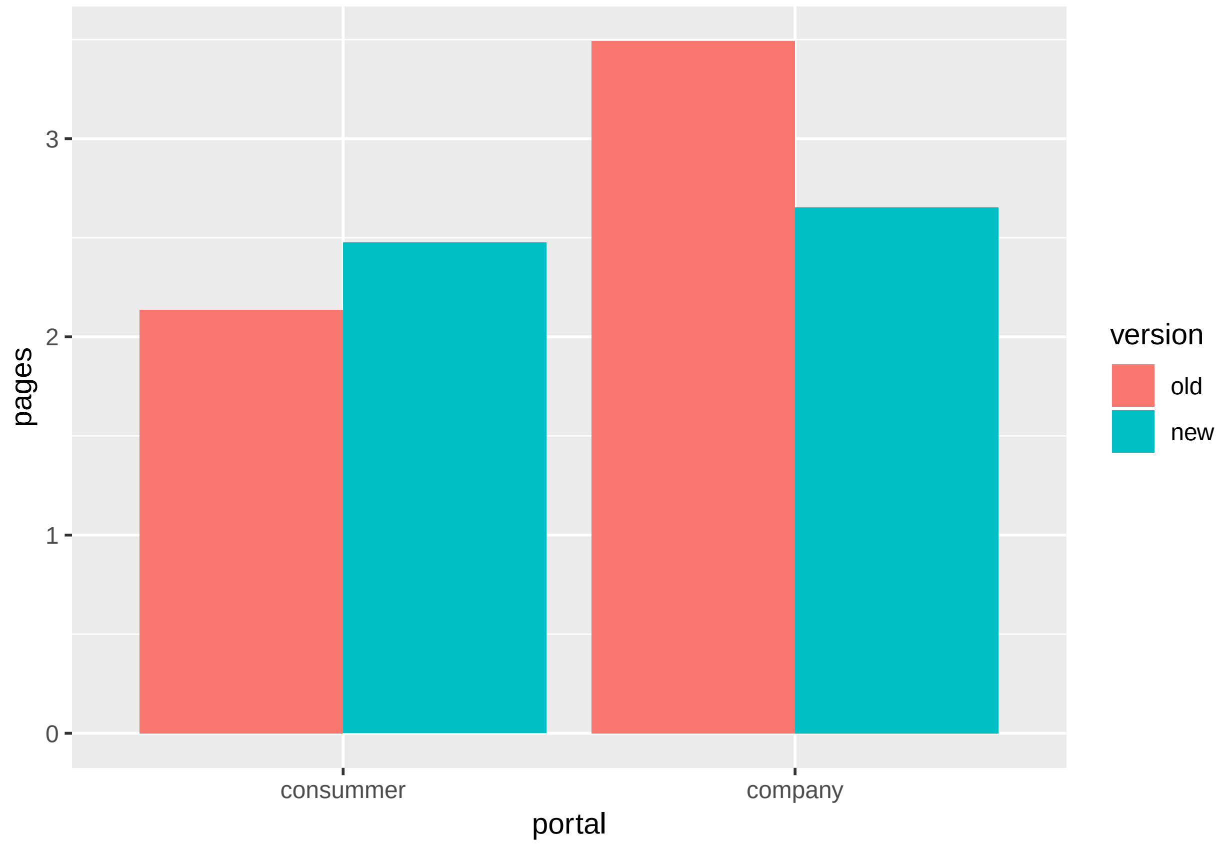
	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
NULL	NA	NA	998	938.6	NA
version	1	7.36	997	931.2	0.006671
portal	1	58.78	996	872.4	1.764e-14
version:portal	1	28.98	995	843.5	7.318e-08

We can observe a significant main effect of the version ($p < 0.01$) and the portal ($p < 0.01$). We also see a significant two-way interaction effect ($p < 0.01$), we will thus perform a simple effect analysis to better understand this interaction effect.

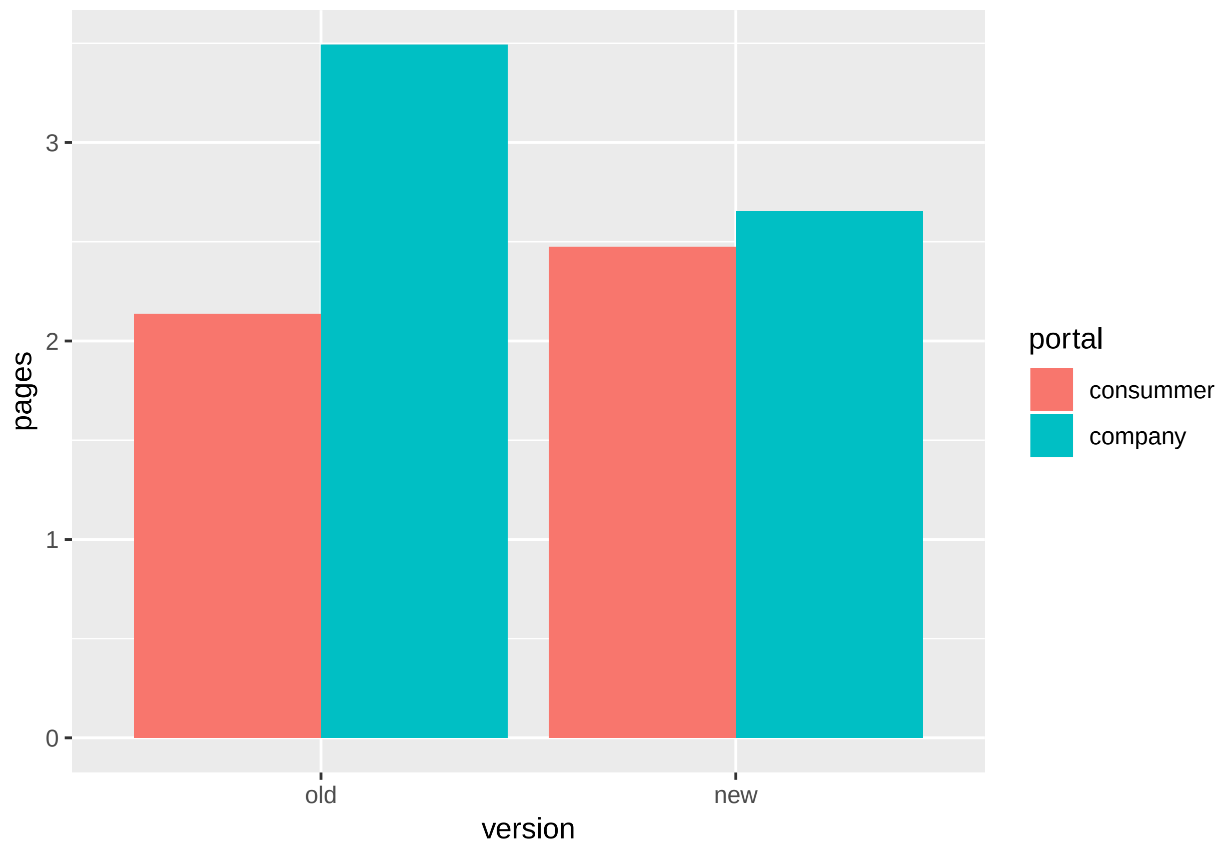
2.2.5 Simple effect analysis

Since we found a significant two-way interaction effect we will conduct a simple effect analysis to explore this interaction effect. We will first plot the effect of the means according to the version and the portal in two different figures to know which effect we should explore in more details.

```
bar <- ggplot(myData, aes(portal, pages, fill = version))
bar + stat_summary(fun.y = mean, geom = "bar", position="dodge")
```




```
bar <- ggplot(myData, aes(version, pages, fill = portal))
bar + stat_summary(fun.y = mean, geom = "bar", position="dodge")
```



From the two figures, we decided that the most interesting effect to explore was the one that compare the two versions and try to see if the change of version change the fact that there is a significant difference in the number of pages visited according to the portal.

```
myData$simple <- interaction(myData$version, myData$portal) #merge two factors

contrastOldVersion <-c(1,0,-1,0) #Only the old version data
contrastNewVersion <-c(0,1,0,-1) #Only the new version data

SimpleEff <- cbind(contrastOldVersion,contrastNewVersion)
contrasts(myData$simple) <- SimpleEff #now we link the two contrasts with the factor simple
pander(simpleEffectModel <-glm(pages ~ simple , data = myData, na.action = na.exclude, family = "poisso
```

Table 5: Simple effect analysis

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.9729	0.01969	49.4	0
simplecontrastOldVersion	-0.246	0.02698	-9.119	7.551e-20
simplecontrastNewVersion	-0.0347	0.0287	-1.209	0.2267
simple	-0.06349	0.03939	-1.612	0.107

The results of the simple effect analysis shows that while there was a significant difference ($P < 0.001$) between the two portals in the old version, we cannot find that significant difference anymore in the new version.

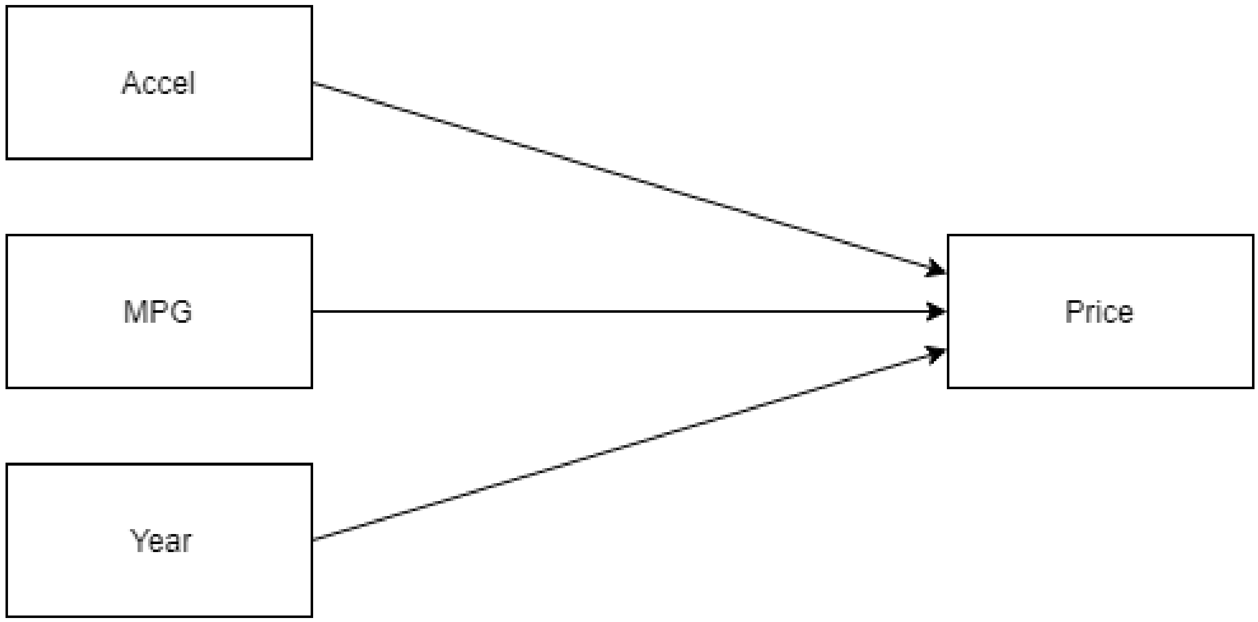


Figure 3: Conceptual model of the four considered variables

2.2.6 Report section for a scientific publication

A generalized linear model was fitted on the number of pages visited of a website, taking the version of the website (an old version and a new one) and the portal that was used to access the website (a portal for consumers and one for companies) as independent variables, and including a two-way interaction between these variables. The analysis found a significant main effect for the version (χ^2 (1,997) = 7.36, p. < 0.01)and for the portal (χ^2 (1,997) = 60.39, p. < 0.001). It also found a significant two-way interaction effect (χ^2 (1,995) = 28.98, p. < 0.001) between these two variables. The two-way interaction was further examined by a Simple Effect analysis on the effect of the version on the differences in portals. It found a significant difference for the portal given the old version (z = -9.1, p. < 0.001) but no difference could be found when using the new version (z = -1.2, p. = 0.23).

2.3 Question 3 - Linear regression analysis

2.3.1 Conceptual model

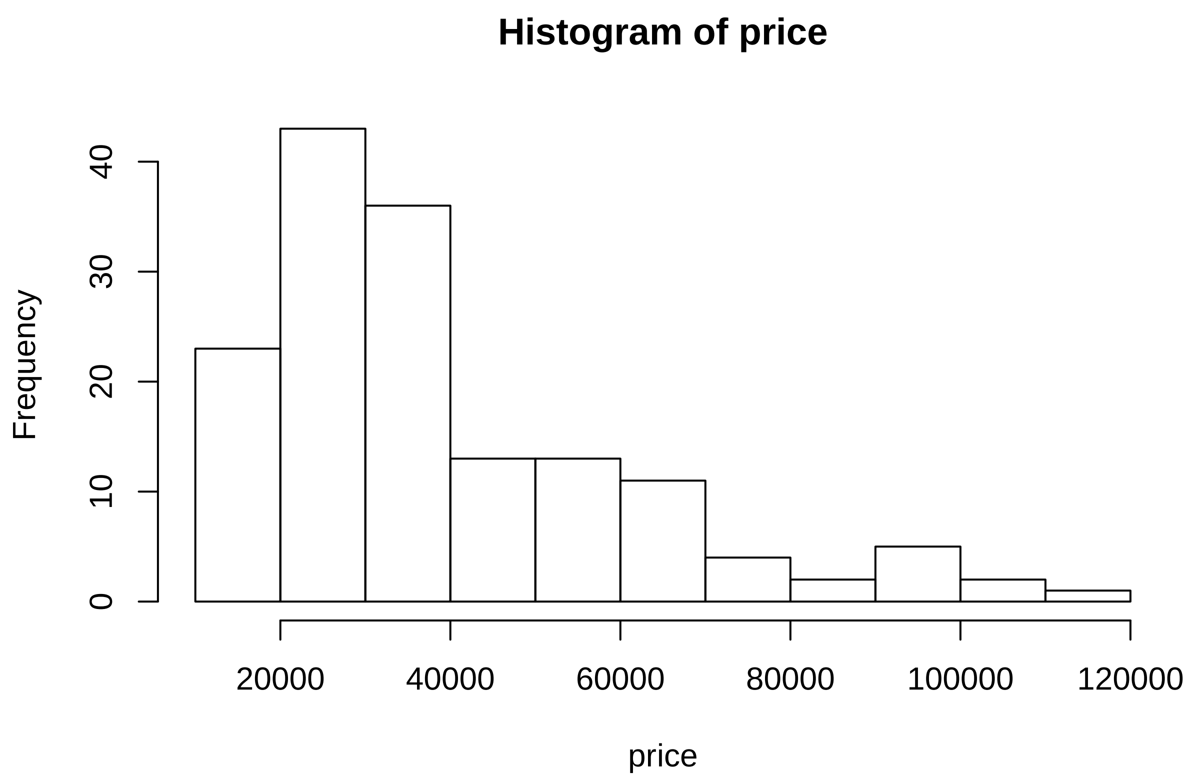
For this assignment we retrieved a data set from http://www.stat.ufl.edu/~winner/datasets.html. The dataset contains facts about 153 hybrid cars, including their price, year built, acceleration data and fuel consumption; those are the four quantitative variables that will be the subject of the linear model in this question. We would like to predict the price of the car (response variable), using data on acceleration rate of the car, the fuel consumption and the year that it was built. The conceptual model for this research looks like this:

2.3.2 Visual inspection

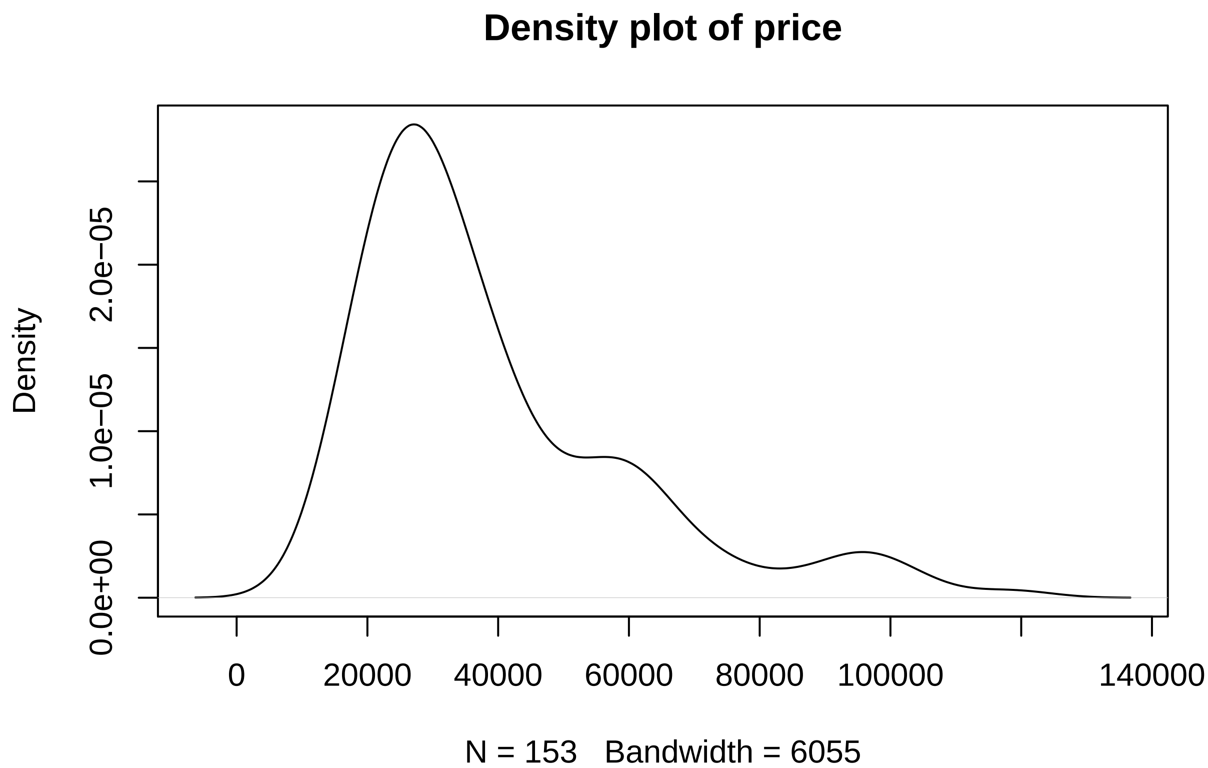
The distribution of the independent variable is displayed here.

```
# Reading in the necessary packages
library(readr)
d <- read_csv("hybrid_reg.csv")
mpg <- d$mpg
year <- d$year
accel <- d$accelrate
price <- d$msrp
```

```r
# Histogram of the distribution of the price variable
hist(price)
```

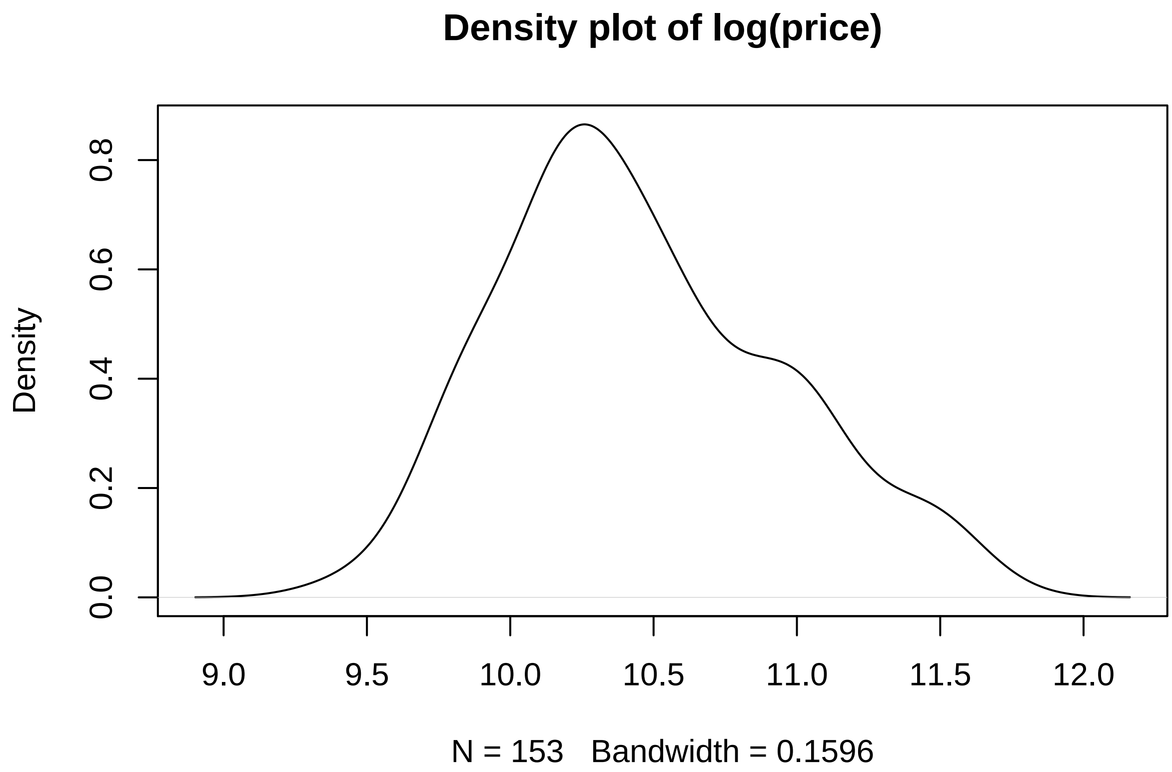



```r
# Density plot of the price variable
plot(density(price),main="Density plot of price")
```



Visual inspection of the plots reveals that the distribution of price deviates from a normal distribution. Especially the right tail of the density distribution has more mass than it should have. Since the distribution is right skewed, a logarithmic transformation is effective in increasing the normality; the result can be seen in the figure below.

```r
# Density plot of the transformed price variable
plot(density(log(price)),main="Density plot of log(price)")
```




```r
shapiro.test(price)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  price
## W = 0.85261, p-value = 4.345e-11
```

```r
shapiro.test(log(price))
```

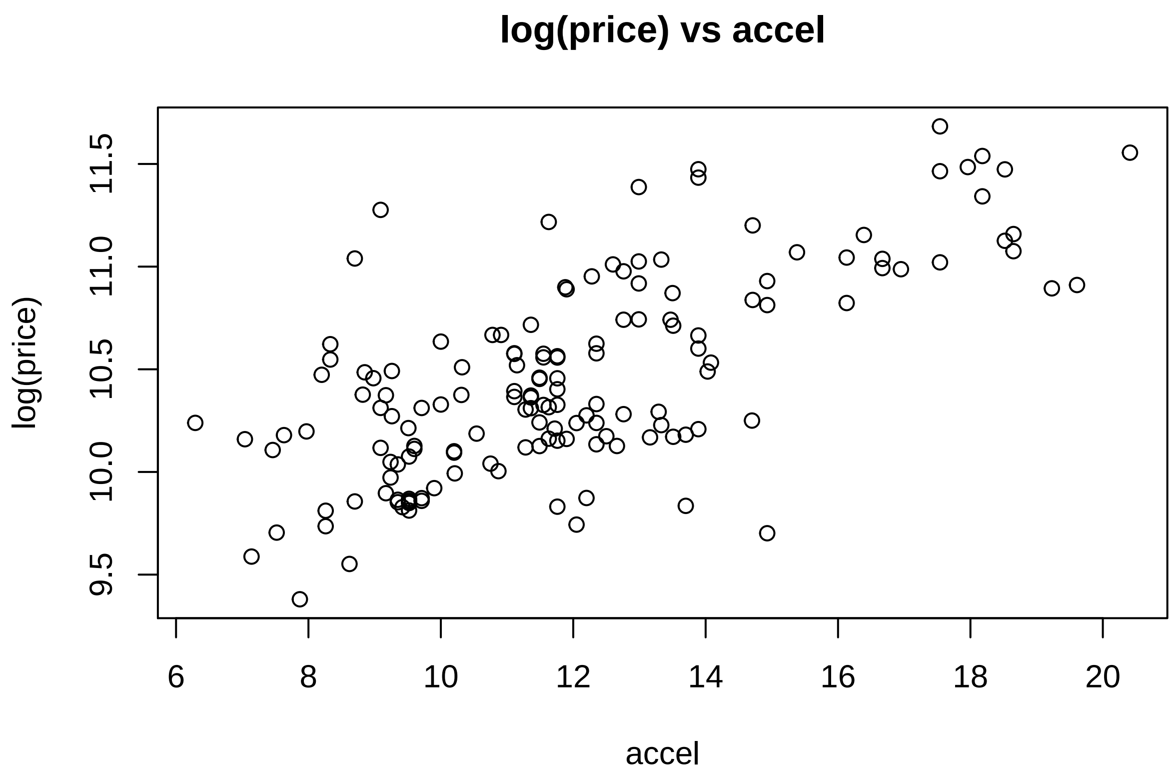
```
##
##  Shapiro-Wilk normality test
##
## data:  log(price)
## W = 0.97322, p-value = 0.004424
```

As we will see later on, the logarithmic transformation is also necessary to justify the choice of linear regression, as without it the assumptions for linear regression do not hold.

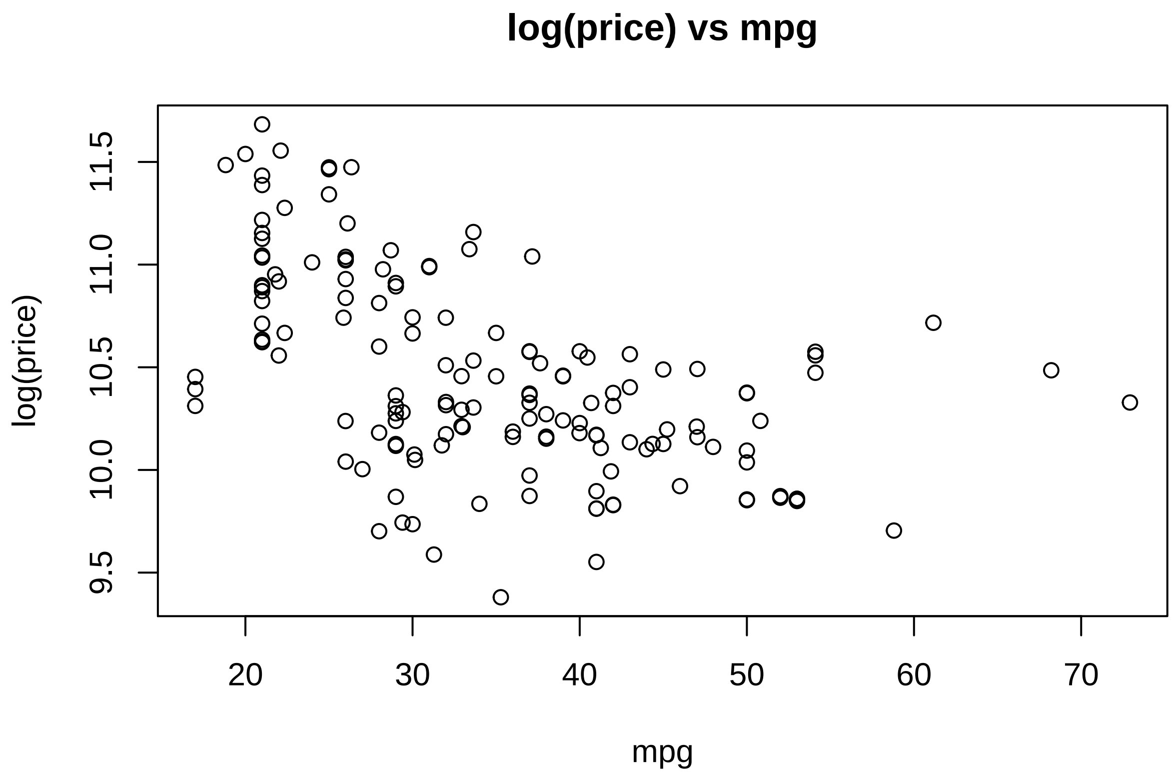
2.3.3 Scatter plot

Using scatter plots we can visually examine the relationship between two variables. The following figures show the scatter plots of the response variable price paired with each of the predictors.

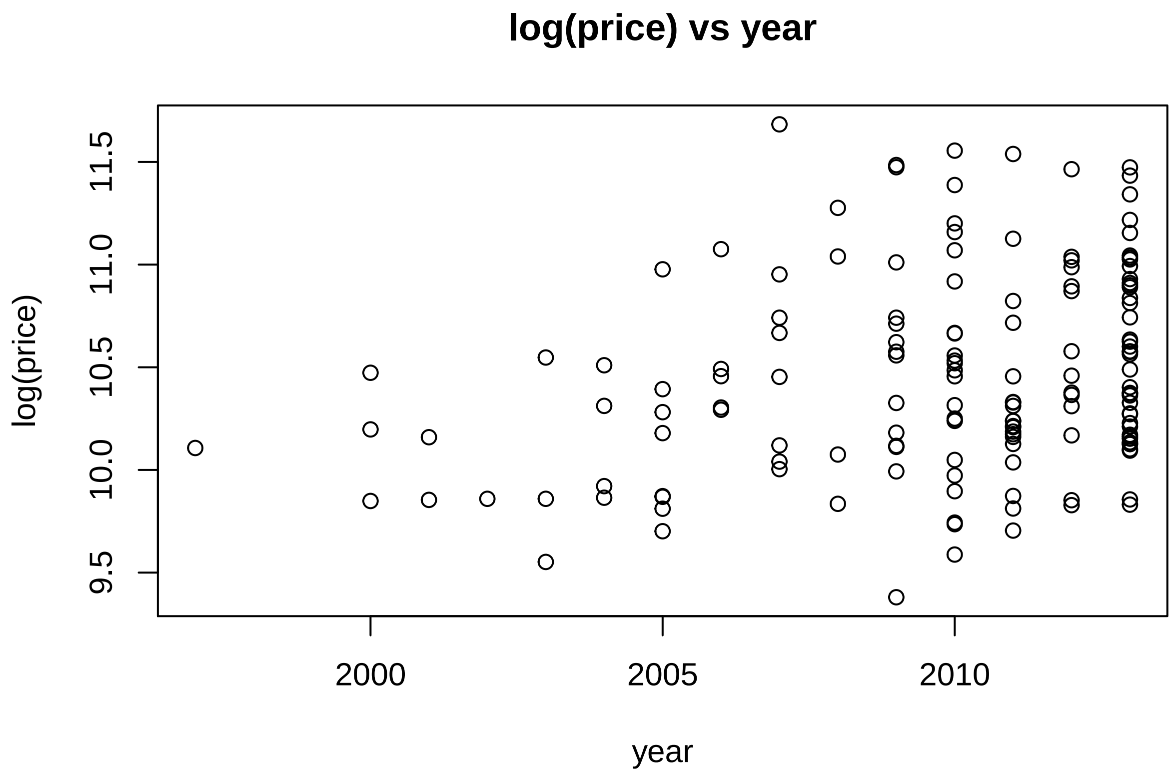
```r
plot(log(price) ~ accel,main="log(price) vs accel")
```




```r
plot(log(price) ~ mpg,main="log(price) vs mpg")
```




```
plot(log(price) ~ year,main="log(price) vs year")
```



We can make a few observations based on these scatter plots. Considering the acceleration variable, a linear pattern can be distinguished. In general, higher accel values also correspond to higher prices. The inverse seems to hold for the mpg variable; low mpg values often have high values attached to them. The third scatter plot does not give the indication of a relationship between year and price. We will now dive deeper into these variables to find out whether our visual inspection matches with the results of linear regression.

2.3.4 Linear regression

From the anova table we can see that the accel and mpg variables are able to significantly improve the model. The year variable is not able to explain any additional significant variance in the price variable. Therefore we exclude the year variable from further analysis. The R^2 value of 0.52 indicates that we can explain about half of the variance in the price using the two remaining independent variables.

```r
library(pander)
model0 <- lm(log(price) ~ 1)
model1 <- lm(log(price) ~ accel)
model2 <- lm(log(price) ~ accel + mpg)
model3 <- lm(log(price) ~ accel + mpg + year)

pander(anova(model0,model1,model2,model3),
       caption = "Model comparison to predict the price of a car")
```

Table 6: Model comparison to predict the price of a car

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
152	35.77	NA	NA	NA	NA
151	18.83	1	16.94	146.9	5.776e-24

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
150	17.19	1	1.639	14.22	0.0002343
149	17.18	1	0.01013	0.08785	0.7673

```
pander(summary(model2),
       caption = "Summary of model 2")
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	9.728	0.2023	48.08	4.898e-93
accel	0.09282	0.01083	8.573	1.15e-14
mpg	-0.01097	0.002901	-3.782	0.0002241

Table 8: Summary of model 2

Observations	Residual Std. Error	R^2	Adjusted R^2
153	0.3385	0.5194	0.513

```
library(QuantPsyc)
pander(confint(model2),
       caption = "#95% confidence interval of the estimates")
```

Table 9: #95% confidence interval of the estimates

	2.5 %	97.5 %
(Intercept)	9.328	10.13
accel	0.07142	0.1142
mpg	-0.0167	-0.00524

```
pander(lm.beta(model2),
       caption = "standardised regression coefficients") # standardised regression coefficients
```

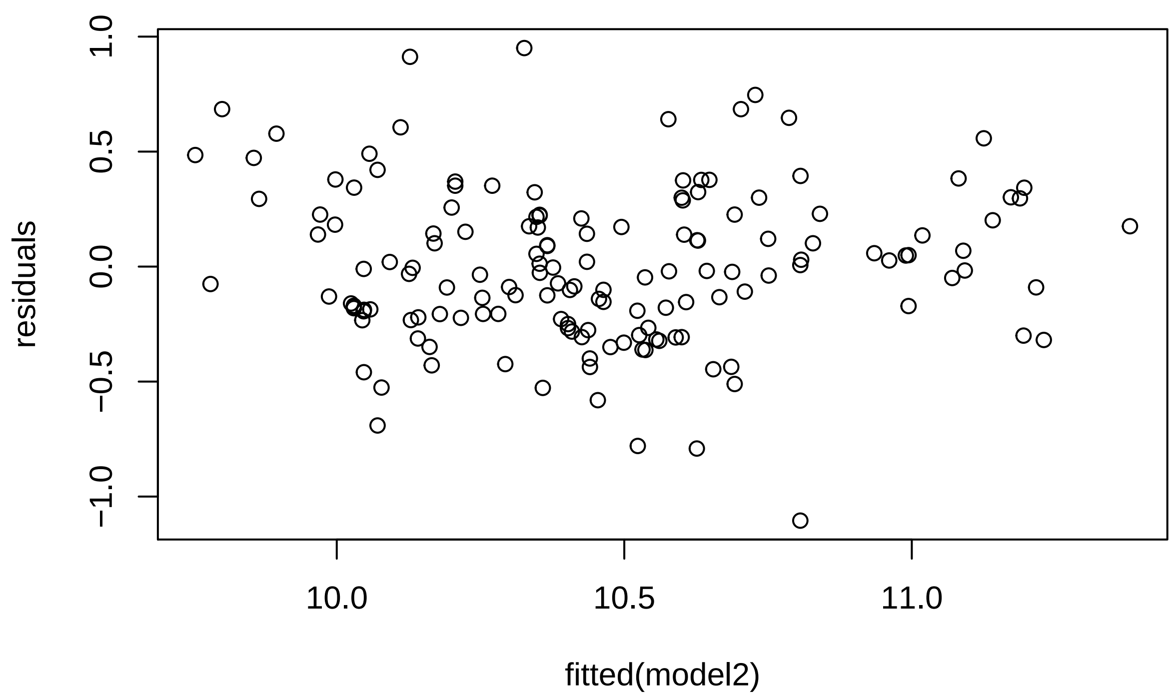
accel	mpg
0.5626	-0.2482

Finally we check the confidence intervals of our estimates to have a better idea of our model's certainty. The standardized regression coefficients serve as a means to compare the magnitude of the effects of the individual coefficients, as their values represent the increase in price associated with an increase in one standard deviation of the predictor. We can see that the effect of the acceleration on the price appears to be more substantial.

2.3.5 Examine assumption

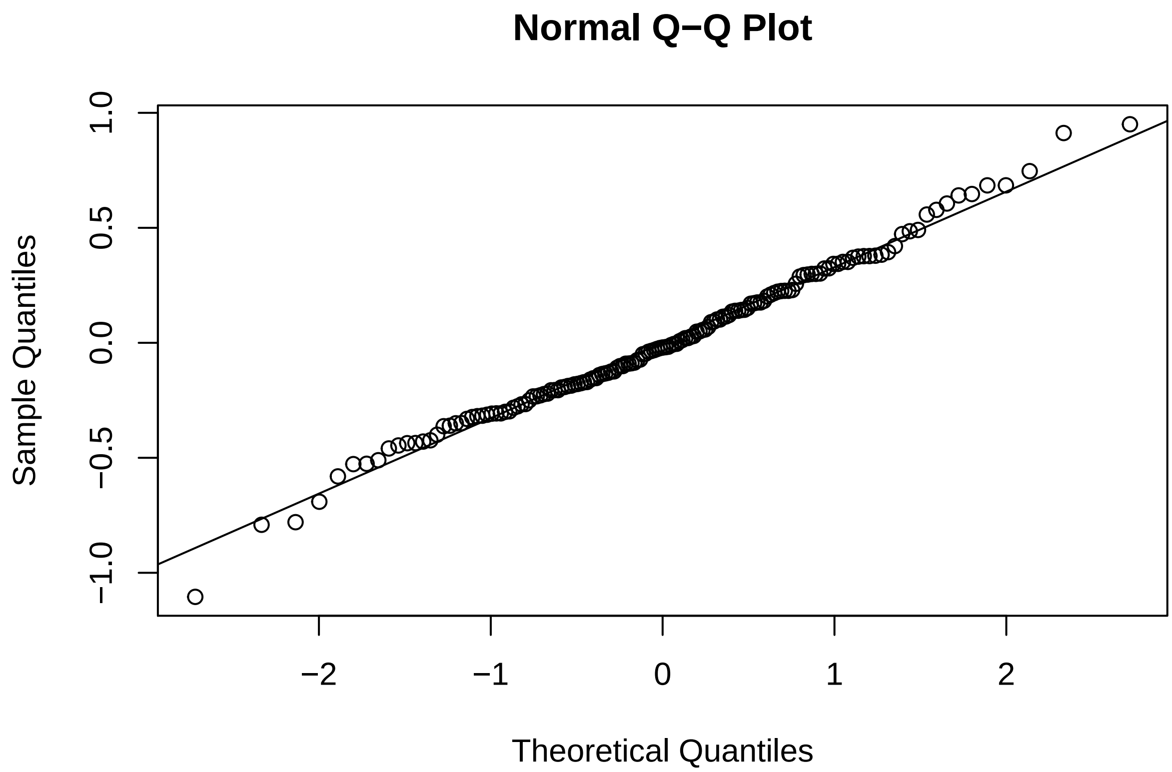
The residuals vs fitted plot is a useful tool in examining the linearity and equal variances assumptions. We are looking for residuals that are distributed normally with expected value 0 and equal variance. Rough inspection of the plot does not reveal major violations of these assumptions.

```r
residuals = resid(model2)
plot(residuals ~ fitted(model2))
```



For normality we can check the qq-plot. All observations seems to be quite close to the theoretical normal distribution, hence from visual inspection we see no real violation of the normality assumption.

```r
library(car)
qqnorm(residuals)
qqline(residuals)
```



Multicollinearity is not an issue in our model. Tolerance values smaller than 0.2 indicate a problem; we are nowhere near this range.

```
vif(model2)
```

```
##   accel     mpg
## 1.34428 1.34428
```

```
1/vif(model2) # Tolerance
```

```
##     accel       mpg
## 0.7438928 0.7438928
```

Checking our model for independent errors revealed a violation; there is an autocorrelation significantly different from 0 between our errors.

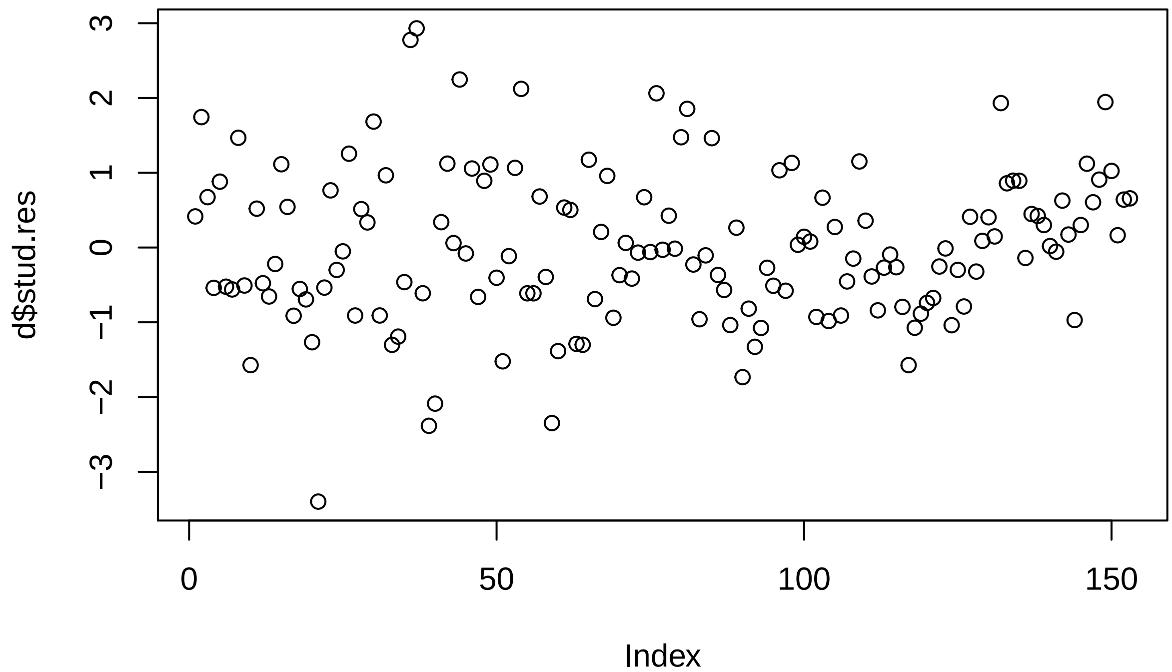
```
show(durbinWatsonTest(model2))
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1       0.2303107      1.535374       0
##  Alternative hypothesis: rho != 0
```

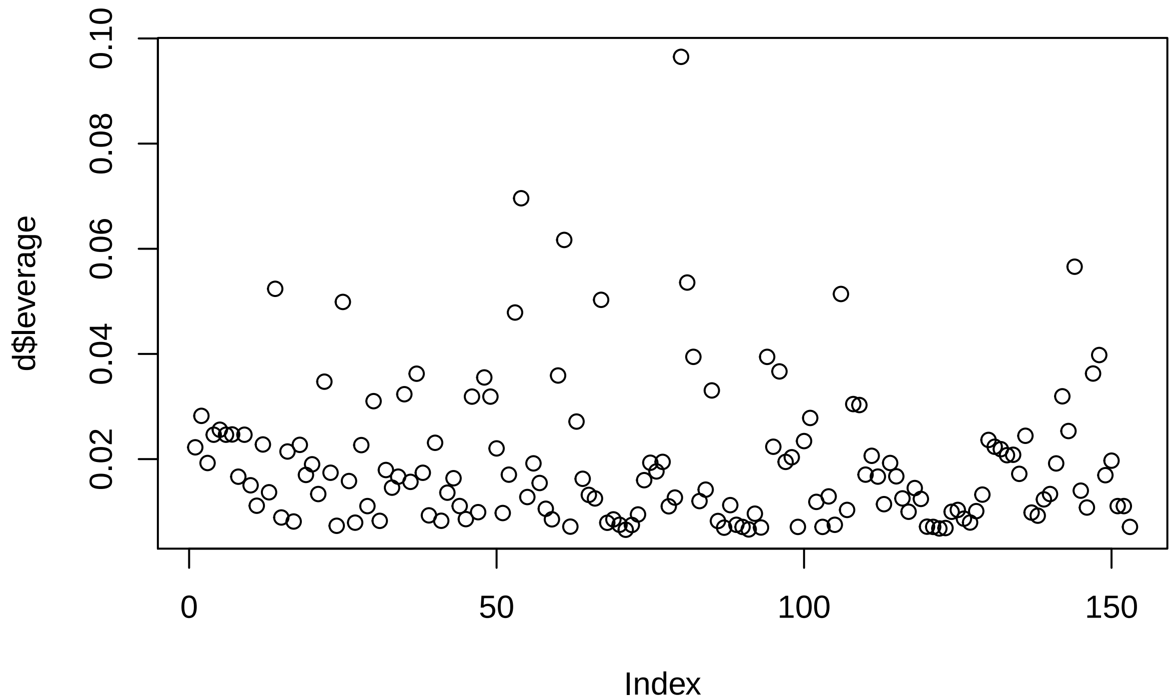
2.3.6 Impact analysis of individual cases

To conclude our analysis we inspect the behavior of individual cases in our model, as some of them might have a disproportionally large influence on the result of the regression. The three methods used to inspect outliers and influential points did not reveal any points that need special attention.

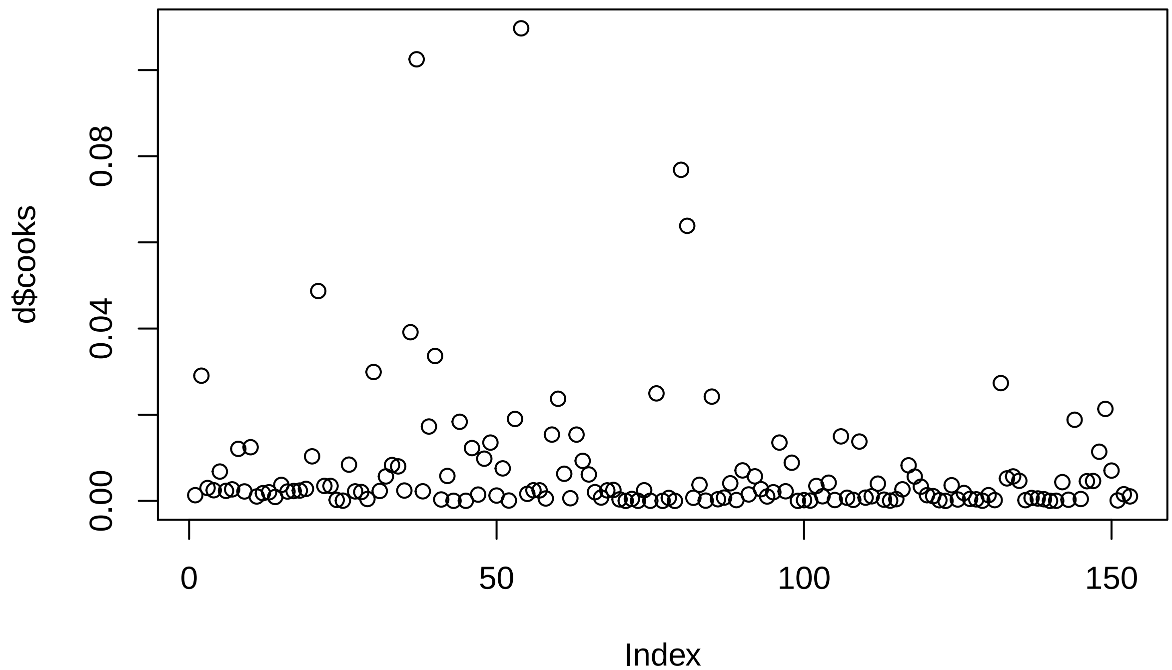
```
d$stud.res<-rstudent(model2)
plot(d$stud.res)
```




```
d$leverage<-hatvalues(model2)
plot(d$leverage)
```




```
d$cooks<-cooks.distance(model2)
plot(d$cooks)
```



2.3.7 Report section for a scientific publication

In this section we briefly present the results of fitting a linear regression model in order to predict the price of cars using data on their acceleration rate, their fuel efficiency in miles per gallon, and their build year.

First of all we can conclude that a logarithmic transformation was necessary to increase normality of the distribution of the price variable. While the distribution still significantly differs from a normal one (W = 0.973, p = 0.004), it is an improvement over the original distribution (W = 0.853, p = 4.3e-11). Furthermore, the transformation was necessary to justify the choice of performing linear regression, as without it the assumptions for linear regression do not hold, especially the linearity assumption.

Inspecting the scatter plots, it becomes clear that a linear relationschip between the natural logarithm of price and the year the car was built is absent. The scatter plots of the other two variables show some indication that a relationship might exist.

Fitting the model revealed that the year variable is indeed not able to explain any additional variance in the price variable on top of the accel and mpg variables (F = 0.088, df = 1, p = 0.77). Based on this result we decided to exclude the independent variable year from the model. Hence we end up with the following model:

log(price) = 9.73 + 0.093 x accel - 0.011 x mpg

This model quantifies the effect of both predictors on the price (R^2 = 0.52). In order to justify its validity, we checked the assumptions of the linear regression. We found that the distribution of the residuals is normal with expected value 0 and (roughly) constant variance. Testing for independence showed a violation of the assumption (D-W = 1.54, p = 0.008). Violating the independence assumption is quite problematic, but for the sake of the exercise we will continue the analysis. Additionally, no multicollinearity could be found in our model.

Analysis of influential and leverage points revealed no severe outlying cases that undermine the linear

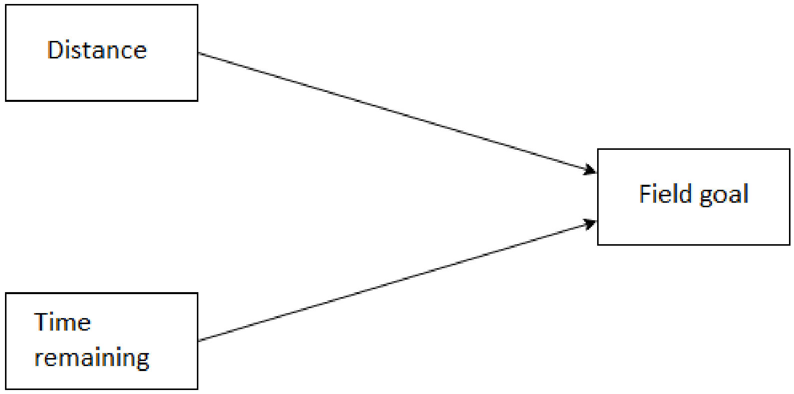


Figure 4: Conceptual model of the four considered variables

regression model.

The interpretation of the coefficients is slightly tricky, since we are dealing with a transformed dependent variable. Instead of additive, the model becomes multiplicative, and each coefficients has to be interpreted as an exponent (i.e. the intercept becomes e^9.73 = 16,815). The standardized coefficients tell us that the effect of one higher standard deviation in the acceleration rate has about twice the effect on the price of one higher standard deviation in the fuel efficiency.

To conclude, we were able to formulate a linear model that is to some extent able to predict the price of a car based on its acceleration rate and its fuel efficiency (to be precise, around 50% of the variance in price can be explained by our model). Since not all assumptions of linear regression were met, interpretation of the results requires caution.

2.4 Question 4 - Logistic regression analysis

2.5 Data set

For this task we will be using a data set with details of all field goal attempts in the NFL season of 2008. The data set contains information in the form of 23 variables for 1039 attempts. We will focus on three of these variables, by investigating the effect of the distance to the goal and the time remaining in the game, on whether or not the attempt was succesful. This setting is graphically depicted in the following conceptual model:

2.5.1 Logistic regression

A general linear model was fitted on the data. The predictor values have been rescaled to represent deviations from their mean; this was done to allow for easier interpretation of the estimated coefficients.

```
library(readr)
library(pander)
library(gmodels)
```

```
## Warning: package 'gmodels' was built under R version 3.5.3
```

```
nfl2008_fga <- read_csv("nfl2008_fga.csv")
```

```
## Parsed with column specification:
## cols(
##   .default = col_double(),
##   AwayTeam = col_character(),
##   HomeTeam = col_character(),
```

```
##   kickteam = col_character(),
##   def = col_character(),
##   name = col_character()
## )
```

```
## See spec(...) for full column specifications.
```

```
d <- nfl2008_fga
goal <- factor(d$GOOD)

d$distanceC <- d$distance - mean(d$distance)
d$timeleftC <- d$timerem - mean(d$timerem)

distance <- d$distanceC
timeleft <- d$timeleftC

model0 <- glm(goal ~ 1, family = binomial())
model1 <- glm(goal ~ distance, family = binomial())
model2 <- glm(goal ~ distance + timeleft, family = binomial())

pander(anova(model0,model1,model2,test="Chisq"),caption="Model comparison for predicting field successfu
```

Table 11: Model comparison for predicting field successful field goals We can observe that model1 fits significantly better than the intercept only model, so the distance from the goal can explain to some extent whether a field goal attempt will be successful. This is, however, not the case for the time remaining in the game. This gives the indication that the time at which a field goal attempt occurs in the game is not important for whether it is successful. We will now look at a summary of the model with the significant predictor distance. Since direct interpretation of the estimates is not straightforward, we will return to their meaning when we consider the odd ratio.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1038	817.7	NA	NA	NA
1037	686.9	1	130.8	2.744e-30
1036	686.9	1	0.03772	0.846

```
pander(summary(model1),
       caption = "Summary results of the final model")
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.332	0.1298	17.96	3.82e-72
distance	-0.1208	0.01229	-9.836	7.886e-23

(Dispersion parameter for binomial family taken to be 1)

Null deviance:	817.7 on 1038 degrees of freedom
Residual deviance:	686.9 on 1037 degrees of freedom

To assess the effectiveness of the model we can first of all look at (pseudo) r^2 measures:

```
logisticPseudoR2s <- function(LogModel)
  #taken from Andy Field et al. book on R, p.334
{
  dev <- LogModel$deviance
  nullDev <- LogModel$null.deviance
  modelN <- length(LogModel$fitted.values)
  R.l <- 1 - dev / nullDev
  R.cs <- 1 - exp(-(nullDev - dev) / modelN)
  R.n <- R.cs / (1 - (exp(-(nullDev / modelN))))
  cat("Pseudo R^2 for logistic regression\n")
  cat("Hosmer and Lemshow R^2:  ", round(R.l, 3), "\n")
  cat("Cox and Snell R^2:       ", round(R.cs, 3), "\n")
  cat("Nagelkerke R^2:          ", round(R.n, 3), "\n")
}
logisticPseudoR2s(model1)
```

```
## Pseudo R^2 for logistic regression
## Hosmer and Lemshow R^2:   0.16
## Cox and Snell R^2:        0.118
## Nagelkerke R^2:           0.217
```

The pseudo measures indicate that a fairly low proportion of the variance can be explained by the predictor in the model (10-20%). This is not surprising, since intuitively it would not make sense for this single variable to have great explanatory power. Continuing our analysis, we will further inspect the estimates through calculation of the odds ratio.

```
o <- exp(model1$coefficients)
show(o)
```

```
## (Intercept)    distance
##  10.3000026   0.8861796
```

These values can be interpreted as follows. For an average distance from the goal, it is roughly ten times more likely to score than to miss. Furthermore, for each added yard further away than average, the odds to score become less likely by a factor of 0.8. These odds have the following conidence intervals on the 95% confidence level.

```
pander(exp(confint(model1)),
       caption = "95% confidence interval of odds ratio")
```

Waiting for profiling to be done...

Table 14: 95% confidence interval of odds ratio

	2.5 %	97.5 %
(Intercept)	8.073	13.44
distance	0.8644	0.9071

2.5.2 Crosstable predicted and observed responses

```
d$goalpred[fitted(model1) <=0.5] <- 0
d$goalpred[fitted(model1) > 0.5] <- 1
goalpred <- factor(d$goalpred)
CrossTable(d$goalpred, d$GOOD, prop.r=FALSE, prop.c = FALSE,
```

```
           prop.t = FALSE,
           prop.chisq=FALSE, format = "SPSS",
           fisher = FALSE, chisq = TRUE,
           expected = FALSE, sresid = FALSE)
```

```
##
##    Cell Contents
## |-------------------------|
## |                   Count |
## |-------------------------|
##
## Total Observations in Table:  1039
##
##              | d$GOOD
##   d$goalpred |        0  |        1  | Row Total |
## -------------|-----------|-----------|-----------|
##            0 |        7  |        6  |       13  |
## -------------|-----------|-----------|-----------|
##            1 |      132  |      894  |     1026  |
## -------------|-----------|-----------|-----------|
## Column Total |      139  |      900  |     1039  |
## -------------|-----------|-----------|-----------|
##
##
## Statistics for All Table Factors
##
##
## Pearson's Chi-squared test
## ------------------------------------------------------------
## Chi^2 =  18.60402     d.f. =  1     p =  1.60881e-05
##
## Pearson's Chi-squared test with Yates' continuity correction
## ------------------------------------------------------------
## Chi^2 =  15.23574     d.f. =  1     p =  9.489051e-05
##
##
##        Minimum expected frequency: 1.739172
## Cells with Expected Frequency < 5: 1 of 4 (25%)
```

When we compare the predicted success of field goal attempts with our prediction, we observe that our model predicts success in 1026 of the 1039 cases; hence we correctly predict 894 of the 900 successful attempts in the data set. However, we only get 7 out of the 139 unsuccessful attempts correct. From the Chi-square test we can conclude that the results differ significantly from each other.

2.5.3 Report section for a scientific publication

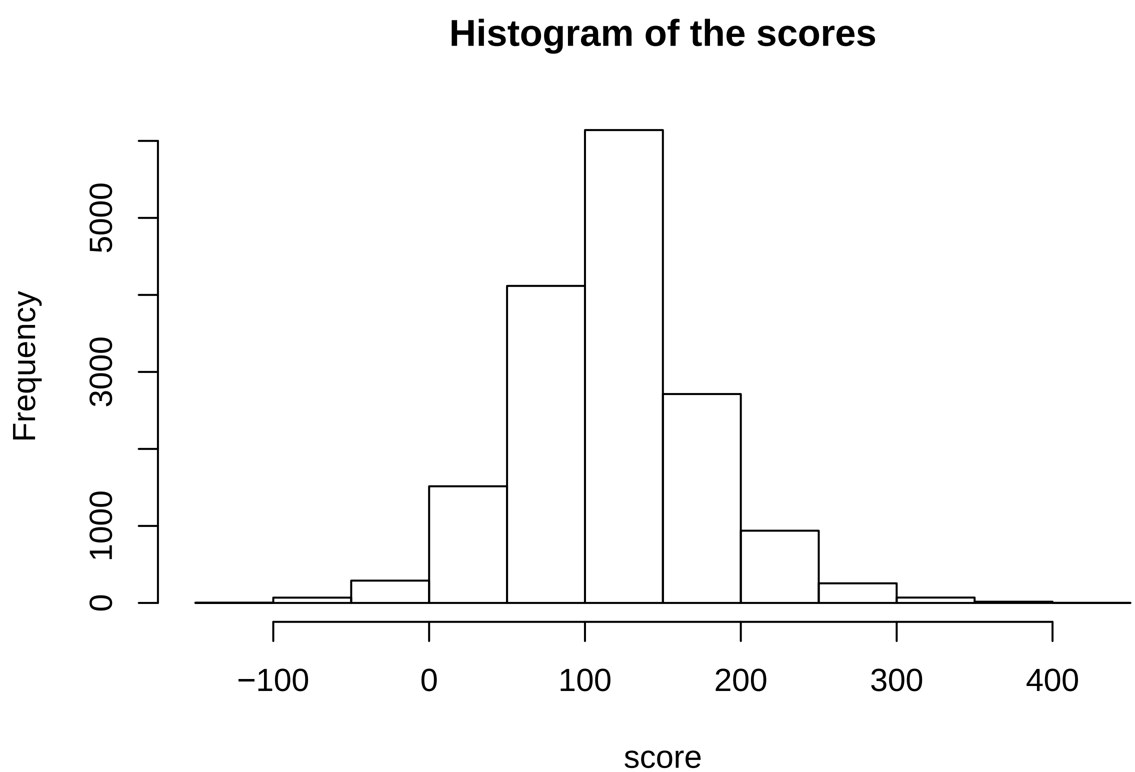
In this exercise we have examined the effect of two predictors on a dichotomous variable: the success of a field goal attempt. The two predictors considered in the analysis are the distance to the goal and the time remaining in the match. We have found that adding the remaining time to the model did not improve the fit ($Chi^2 = 0.038$, df = 1, p = 0.846), compared to the model with only the distance as a predictor ($Chi^2 = 130.8$, df = 1, $p < 0.0001$). The main tangible result of the model can be summarized by the odds ratios. The average odds of scoring a goal (8.07 - 13.44) decrease by a factor of (0.86 - 0.91) for each yard added to the average distance, on a 95% confidence level. The results of the model differed significantly ($Chi^2 = 18.6$, df = 1, $p<0.0001$) from the true data, as our model was only able to explain between 10 and 20% of the variance in the data. Nonetheless, our results confirm the intuition that field goal attempts are generally less successful from a larger distance.

3 Part 3 - Multilevel model

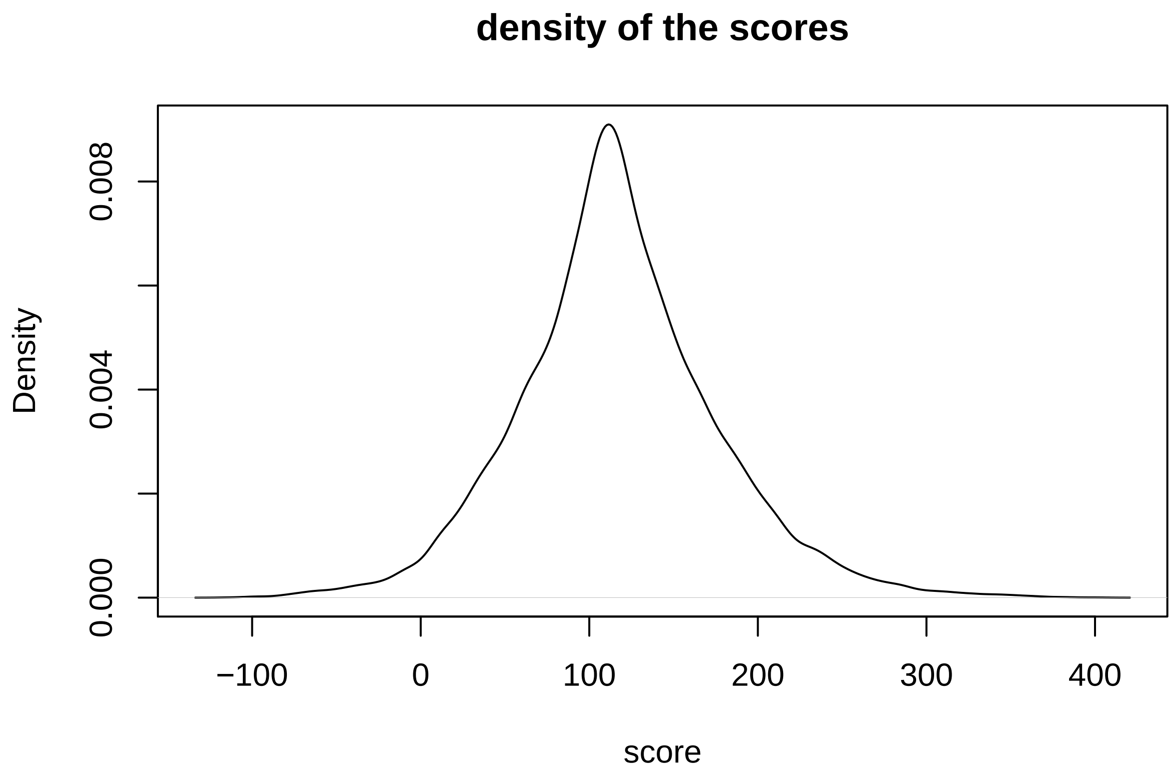
3.1 Visual inspection

First, we inspect the distribution of the scores.

```
set1 <- read.csv("set1.csv", header = TRUE)
set1$Subject <- factor(set1$Subject)
# Should we consider the session as factor ?
#set1$session <- factor(set1$session)

hist(set1$score, xlab="score", main="Histogram of the scores")
```

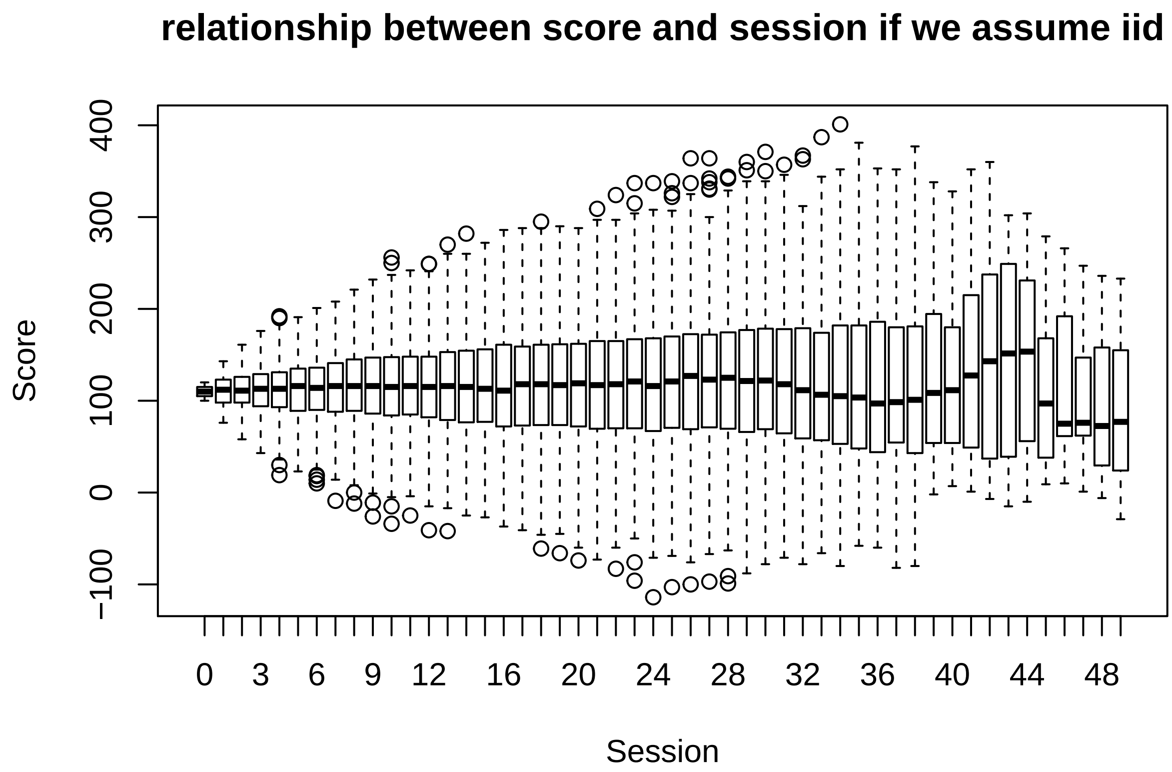



```
plot(density(set1$score), xlab="score", main="density of the scores")
```



If we just look at the score distribution, they seems to be normally distributed and centered around 150. We will now look at how the session affect the scores with an assumption of iid of the variables (which is false).

```
#Plot of the relationship between session and score
# Assuming iid of the variables which is not true, shouldn't do that.
boxplot(score ~ session, data=set1, xlab="Session", ylab="Score", main="relationship between score and s
```



It seems that even without looking at each person individually the session seems to have an impact on the scores. We will further analyse this effect by taking into account the evolution of each person now.

3.2 Multilevel analysis

Conduct multilevel analysis and calculate 95% confidence intervals, determine:

We will conduct a multilevel analysis of the dataset where each subject get a random intercept. We first want to know if the session has an impact on people scores.

```
randomIntercept <- lme(score ~ 1, data = set1, random = ~1|Subject, method="ML")
addSession <- update(randomIntercept, .~. + session)
pander(anova(randomIntercept,addSession), caption="comparisons of models when session is added as a fixe
```

Table 15: comparisons of models when session is added as a fixed factor (continued below)

	call	Model	df	AIC
randomIntercept	lme.formula(fixed = score ~ 1, data = set1, random = ~1 | Subject, method = “ML”)	1	3	162711
addSession	lme.formula(fixed = score ~ session, data = set1, random = ~1 | Subject, method = “ML”)	2	4	162545

	BIC	logLik	Test	L.Ratio	p-value
randomIntercept	162734	-81352		NA	NA
addSession	162576	-81269	1 vs 2	167.7	2.317e-38

The addition of the session significantly improves the model (p. < 0.001), we will now verify the 95% confidence bound.

```
intervals(addSession)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                   lower        est.       upper
## (Intercept) 106.8665678 111.0675622 115.2685566
## session       0.3126229   0.3682005   0.4237781
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: Subject
##                     lower    est.    upper
## sd((Intercept)) 43.67471 46.5146 49.53914
##
##  Within-group standard error:
##    lower     est.    upper
## 34.68269 35.06933 35.46028
```

We see that for the fixed effect, the session deviates from 0 in the 95% interval which means that there is a significant impact of the session on people's scores.

We will now look if there is a significant variance between the participants in their score.

```
intervals(randomIntercept)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                lower     est.    upper
## (Intercept) 112.7019 116.8139 120.9259
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
##   Level: Subject
##                     lower     est.    upper
## sd((Intercept)) 43.68633 46.52747 49.55338
##
##  Within-group standard error:
##    lower     est.    upper
## 34.86891 35.25763 35.65067
```

We can see that in the Random effect, the standard deviation of the intercept does not include 0 in the 95% interval, thus there is a significant variance between the participants in their score.

3.3 Report section for a scientific publication

We fitted a linear Mixed-Effects model with a random intercept for each subject and we then build a new model that used the session as independant variables to predict the scores. It appears that there is a significant main effect of the session over participants score ($\chi^2(1) = 167.7$, p. < 0.001) when compared to the baseline at a 95% confidence interval (estimate(session) = [0.31;0.42]). We can also show that there is a significant variance between participants in their score (sd(intercept) = [43.69;49.55]) at a 95% confidence interval. From these results we can draw the conclusion that the session in which the participants is has an impact on his score which can be interpreted as improvement over each exercice session since the estimates are positive. We can also conclude that each participant is different and that indeed, we cannot consider the observations to be independent.