Week THREE

This week is about logistic regression. Logistic regression is a common approach when we want to construct a regression model where the dependent variable is binary. Logistic regression model gives us a probability of the occurrence of the event (binary variable) with certain values of the explanatory variables. This probability is according to that model, and would be different in another model.

Here is the code for the data wrangling part:

```
#################
#Paula Bergman
#7.2.2017
#Data wrangling to create the alcohol consumption dataset
#Data downloaded from the website
#https://archive.ics.uci.edu/ml/machine-learning-databases/00356/

#Set the working directory
setwd("C:/Users/Paula/Documents/GitHub/IODS-project/data")

#Read the datasets into RStudio
mat<-read.table("student-mat.csv",header=T,sep=";")
por<-read.table("student-por.csv",header=T,sep=";")

#Check the dimensions and the structure of the datasets
dim(mat)
str(mat)
#We can see that the mat-data  consists of 395 observations and 33 variables,
#17 of which are of type factor and the rest integer.
dim(por)
str(por)
#We can see that the por-data consists of 649 observations and 33 variables,
#which are exactly the same than in the mat-data.

#Access the dplyr library
library(dplyr)

#Join the datasets by using the variables "school", "sex", "age", "address",
#"famsize", "Pstatus", "Medu", "Fedu", "Mjob", "Fjob", "reason", "nursery",
#"internet" as (student) identifiers. This way only the students present in
#both datasets will be included in the data
join_by <- c("school","sex","age","address","famsize","Pstatus","Medu","Fedu",
             "Mjob","Fjob","reason","nursery","internet")
mat_por <- inner_join(mat, por, by = join_by)
mat_por<-inner_join(mat,por,by=join_by,suffix=c(".mat",".por"))

dim(mat_por)
str(mat_por)
#Now the dataset has 382 observations and 53 variables. Those variables that I
#used for joining, are there only once but for the rest of the variables there are
#two of each, separated with the endings .mat and .por

#Create a new data frame with only the joined columns
alc <- select(mat_por, one_of(join_by))
```

```
#The columns in the datasets which were not used for joining the data
notjoined_columns <- colnames(mat)[!colnames(mat) %in% join_by]

#For every column name not used for joining...
for(column_name in notjoined_columns) {
  #select two columns from 'math_por' with the same original name
  two_columns <- select(mat_por, starts_with(column_name))
  #select the first column vector of those two columns
  first_column <- select(two_columns, 1)[[1]]

  #if that first column vector is numeric...
  if(is.numeric(first_column)) {
    #take a rounded average of each row of the two columns and
    #add the resulting vector to the alc data frame
    alc[column_name] <- round(rowMeans(two_columns))
  } else { # else if it's not numeric...
    #add the first column vector to the alc data frame
    alc[column_name] <- select(two_columns, 1)[[1]]
  }
}

#Create an alcohol consumption variable alc_use to the joined data by
#taking the average of alcohol consumption variables in the data.
#Those are Dalc = Daily alcohol consumption and Walc = Weekly alcohol consumption
alc$alc_use<-(alc$Dalc+alc$Walc)/2

#Create a variable high_use to the joined data to describe whether the student
#uses alcohol a high amount or not
alc$high_use<-ifelse(alc$alc_use>2,TRUE,FALSE)

#Glimpse at the joined and modified data to make sure everything is in order
glimpse(alc)

#It seems that everything is in the order: There are 382 observations and
#35 variables as it is supposed to be.

#Save the joined dataset into the data-folder in csv-form. Since the working
#directory is already set into that folder, it doesn't have to be referenced
#anymore.
write.csv(alc,"alc.csv",row.names=F)
```

First I read back into RStudio the dataset created in data wrangling part and checked the names of the variables in the dataset. I also called the possibly necessary packages from library.

```
setwd("C:/Users/Paula/Documents/GitHub/IODS-project/data")
alc<-read.table("alc.csv",header=T,sep=",")
names(alc)
```

```
##  [1] "school"     "sex"        "age"        "address"    "famsize"
##  [6] "Pstatus"    "Medu"       "Fedu"       "Mjob"       "Fjob"
## [11] "reason"     "nursery"    "internet"   "guardian"   "traveltime"
## [16] "studytime"  "failures"   "schoolsup"  "famsup"     "paid"
## [21] "activities" "higher"     "romantic"   "famrel"     "freetime"
```

```
## [26] "goout"      "Dalc"       "Walc"       "health"     "absences"
## [31] "G1"         "G2"         "G3"         "alc_use"    "high_use"
```

```
# Call packages needed this week, ggplot2, dplyr and GGally from the library
library(ggplot2)
library(dplyr)
library(GGally)
```

This is a dataset of students of maths course and Portuguese course. (Data Set Information: P. Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th FUture BUsiness TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.) It consists of 35 variables, describing students' basic information, family background, free time activities, school performance and alcohol usage. The last two variables were created in the data wrangling part and explanations can be found in that code, and the more detailed description of the rest of the variables can be found here.

The purpose of this weeks analysis is to study the relationship between high/low alcohol consumption and some other variables of my choice in the data. I chose

Pstatus, which tells whether the parents of the student are together (=T) or apart (=A) (dicotomous/binary variable)

age, which tells the student's age in years (between 15 and 22, numeric variable)

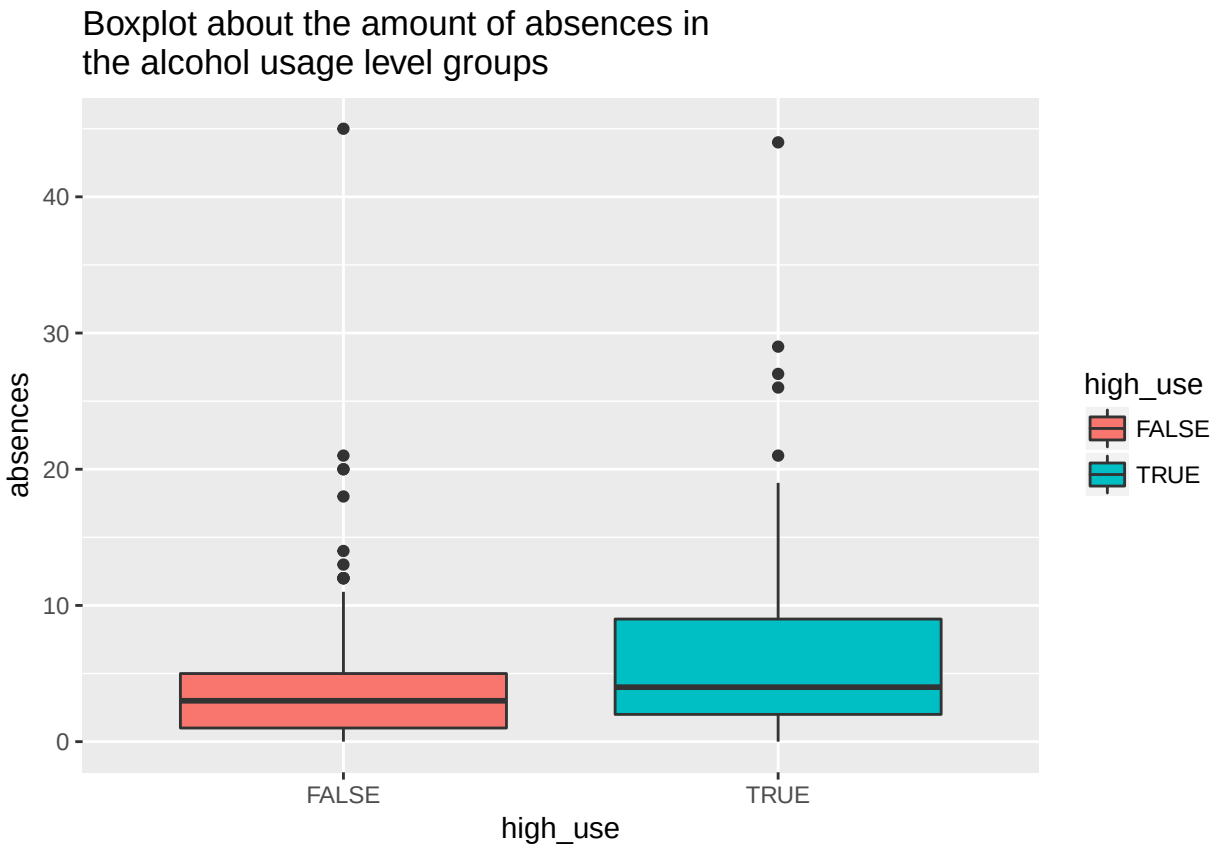
studytime, which tells how much time the student spends weekly studying (numeric variable: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)

absences, which tells how many absences the student has had from school (numeric variable, 0-93)

My hypothesis for the parents' status is that if the parents are apart, it is more likely for the student to use high amount of alcohol. I also suppose that the older the student is, more probable is high alcohol consumption. I also would guess that the less time student spends studying, the more they use alcohol and the more absences they have from school the more they use alcohol. I hypotesize that the last two variables indicate that the student has less interest in education in general and I think that could be one predictive factor for high alcohol usage.

I first decided to look at these variables graphically, starting from a similar graph I made last week.

```
# Plot
qplot(high_use, absences, data=alc, geom="boxplot", fill=high_use,main="Boxplot about the amount of abs
```



By looking at the boxes, especially the median value, it seems that the absences are more likely in the high use group than in the group of students who don't use high amounts of alcohol. We can also see that both groups have an outlier. This could be useful to remember later.

To understand better the relationship between Pstatus and high_use, I decided to use the traditional way and to crosstabulate them

```
attach(alc)
table(high_use,Pstatus)
```

```
##         Pstatus
## high_use   A   T
##    FALSE  26 242
##    TRUE   12 102
```

```
library(xtable)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")

data=alc
xtable(alc[c("high_use","Pstatus"),])
```

```
## % latex table generated in R 3.3.2 by xtable 1.8-2 package
## %
## \begin{tabular}{rllrllllrrrllllllrrrrrrrrrrrrl}
```

```
##   \hline
##  & school & sex & age & address & famsize & Pstatus & Medu & Fedu & Mjob & Fjob & reason & nursery &
##   \hline
## NA &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &
##   NA.1 &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &  &
##    \hline
## \end{tabular}
```

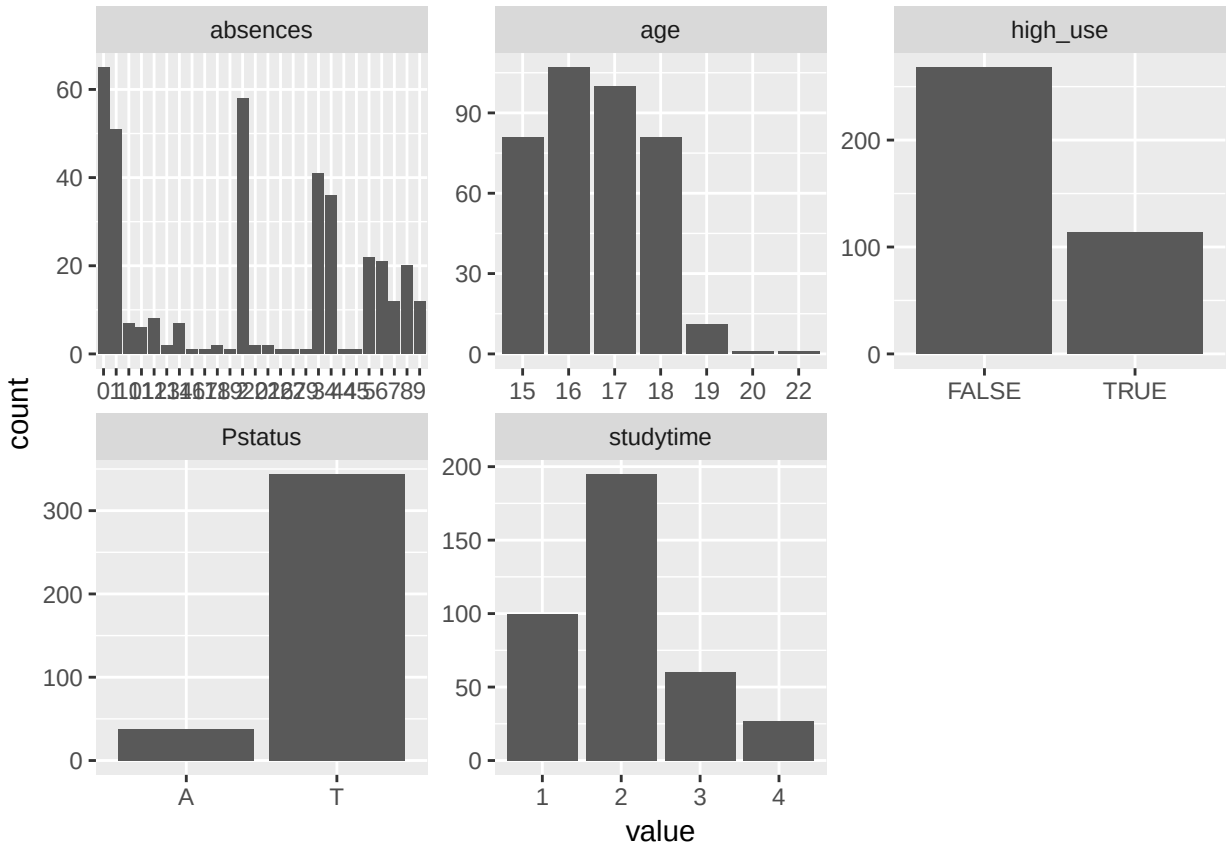
```
#Access the tidyverse library tidyr
library(tidyr)

#Take a subset of the alc-data by using only the explanatory variables
#and high_use
myvars<-c("high_use","Pstatus","age","studytime","absences")
alc1<-alc[myvars]

# use gather() to gather columns into key-value pairs and then glimpse() at the resulting data
gather(alc1) %>% glimpse
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```

```
## Observations: 1,910
## Variables: 2
## $ key   <chr> "high_use", "high_use", "high_use", "high_use", "high_us...
## $ value <chr> "FALSE", "FALSE", "TRUE", "FALSE", "FALSE", "FALSE", "FA...
```

```
# draw a bar plot of each variable
gather(alc1) %>% ggplot(aes(value)) + facet_wrap("key", scales = "free") + geom_bar()
```

```
## Warning: attributes are not identical across measure variables; they will
## be dropped
```


absences
age
high_use
Pstatus
studytime
count
value
FALSE
TRUE
A
T