A1\_P1\_Thea Rolskov Sloth

01/07/2018

# Assignment 1, Part 1: Language development in Autism Spectrum Disorder (ASD) - Brushing up your code skills

In this first part of the assignment we will brush up your programming skills, and make you familiar with the data sets you will be analysing for the next parts of the assignment.

In this first part of the assignment you will: 1) Create a Github (or gitlab) account and link it to your RStudio 2) Use small nifty lines of code to transform several data sets into just one. The final data set will contain only the variables that are needed for the analysis in the next parts of the assignment 3) Become familiar with the tidyverse package (especially the sub-packages stringr and dplyr), which you will find handy for later assignments.

## 0. First an introduction on the data

# Language development in Autism Spectrum Disorder (ASD)

Background: Autism Spectrum Disorder is often related to language impairment. However, this phenomenon has not been empirically traced in detail: i) relying on actual naturalistic language production, ii) over extended periods of time. We therefore videotaped circa 30 kids with ASD and circa 30 comparison kids (matched by linguistic performance at visit 1) for ca. 30 minutes of naturalistic interactions with a parent. We repeated the data collection 6 times per kid, with 4 months between each visit. We transcribed the data and counted: i) the amount of words that each kid uses in each video. Same for the parent. ii) the amount of unique words that each kid uses in each video. Same for the parent. iii) the amount of morphemes per utterance (Mean Length of Utterance) displayed by each child in each video. Same for the parent.

## 1. Let's get started on GitHub

Follow the link to a Github tutorial: <https://support.rstudio.com/hc/en-us/articles/200532077-Version-Control-with-Git-and-SVN>

In the assignments you will be asked to upload your code on Github and the GitHub repositories will be part of the portfolio, therefore all students must make an account and link it to their RStudio (you'll thank us later for this!).

N.B. Create a GitHub repository for the Language Development in ASD set of assignments and link it to a project on your RStudio.

You may also use Gitlab instead of Github (Malte will explain in class)

## 2. Now let's take dirty dirty data sets and make them into a tidy one

If you're not in a project in Rstudio, make sure to set your working directory here. If you created an rstudio project, then your working directory (the directory with your data and code for these assignments) is the project directory. You can always see which working directory is used with getwd(). Note that this may be different in the console than in the Rmd cells (don't ask me why).

library(tidyverse)

library(plyr)

library(stringr)

setwd("~/Studygroup/Assignment1&2\_LanguageASD")

Load the three data sets, after downloading them from dropbox and saving them in your working directory: \* Demographic data for the participants: <https://www.dropbox.com/s/w15pou9wstgc8fe/demo_train.csv?dl=0> \* Length of utterance data: <https://www.dropbox.com/s/usyauqm37a76of6/LU_train.csv?dl=0> \* Word data: <https://www.dropbox.com/s/8ng1civpl2aux58/token_train.csv?dl=0>

demo<-read.csv("demo\_train.csv", stringsAsFactors = FALSE)  
lu<-read.csv("LU\_train.csv", stringsAsFactors = FALSE)  
token<-read.csv("token\_train.csv", stringsAsFactors = FALSE)

Explore the 3 datasets (e.g. visualize them, summarize them, etc.). You will see that the data is messy, since the psychologists collected the demographic data, a linguist analyzed the length of utterance in May 2014 and the same linguist analyzed the words several months later. In particular: - the same variables might have different names (e.g. identifier of the child) - the same variables might report the values in different ways (e.g. visit) Welcome to real world of messy data :-)

Before being able to combine the data sets we need to make sure the relevant variables have the same names and the same kind of values.

So:

2a. Find a way to transform variable names. Tip: Look into the package dplyr (part of tidyverse) Tip: Or google "how to rename variables in R". Tip: Or look through the chapter on data transformation in R for data science (<http://r4ds.had.co.nz>).

#change colnames  
colnames(demo)[colnames(demo)=="Child.ID"] <- "SUBJ"  
colnames(demo)[colnames(demo)=="Visit"] <- "VISIT"  
  
#change to lower case in order to try to make them equal - this was not nessecary, since we used str\_extract later  
lu <- data.frame(lapply(lu, function(v) {  
 if (is.character(v)) return(tolower(v))  
 else return(v)  
}))  
  
token <- data.frame(lapply(token, function(v) {  
 if (is.character(v)) return(tolower(v))  
 else return(v)  
}))  
  
demo <- data.frame(lapply(demo, function(v) {  
 if (is.character(v)) return(tolower(v))  
 else return(v)  
}))  
  
#change visit variable  
lu$VISIT <- revalue(lu$VISIT, c("visit1." = "1", "visit2." = "2", "visit3." = "3", "visit4." = "4", "visit5." = "5", "visit6." = "6"))  
token$VISIT <- revalue(token$VISIT, c("visit1." = "1", "visit2." = "2", "visit3." = "3", "visit4." = "4", "visit5." = "5", "visit6." = "6"))

2b. Find a way to homogeneize the way "visit" is reported. If you look into the original data sets, you will see that in the LU data and the Token data, Visits are called "visit 1"" instead of just "1"" (which is the case in the demographic data set). Tip: There is a package called stringr, which will be very handy for manipulating (text) strings also in furture assignments. We will return to this package later, but for now use the str\_extract () to extract only the number from the variable Visit in each data set.

#extract numbers  
lu$VISIT <- str\_extract(lu$VISIT,"[0-9]+")  
token$VISIT <- str\_extract(token$VISIT,"[0-9]+")  
  
#alternatives: [:digit:] -> find a number/extract whatever is in []

2c. We also need to make a small adjustment to the content of the Child.ID coloumn in the demographic data. Within this column, names that are not abbreviations do not end with "." (i.e. Adam), which is the case in the other two data sets (i.e. Adam.). If The content of the two variables isn't identical the data sets will not be merged sufficiently. We wish to remove the "." at the end of names in the LU data and the tokens data.

Tip: stringr is helpful again. Look up str\_replace\_all Tip: You can either have one line of code for each child name that is to be changed (easier, more typing) or specify the pattern that you want to match (more complicated: look up "regular expressions", but less typing)

Tip: You will have to do identical work for both data sets, so to save time on the copy/paste use the cmd+f/ctrl+f function. Add the data frame name (e.g. LU\_data) in the first box, and the data frame name (e.g. Tokens\_data) you wish to change it to in the other box, and press replace. Or create a function that takes the data set and does the transformation. Then call the function on both data sets.

#removing the . at the end of the names  
  
lu$SUBJ = as.character(lu$SUBJ)  
token$SUBJ = as.character(token$SUBJ)  
  
#detect all the names with only 1 . and remove  
for (row in 1:nrow(lu)) {  
 if (str\_count(lu[row, "SUBJ"], "[.]") == 1) {  
 lu[row,"SUBJ"] <- str\_replace(lu[row, "SUBJ"],"[.]", "")  
 }  
}  
  
for (row in 1:nrow(token)) {  
 if (str\_count(token[row, "SUBJ"], "[.]") == 1) {  
 token[row,"SUBJ"] <- str\_replace(token[row, "SUBJ"],"[.]", "")  
 }  
}  
  
#alternative: ".$" -> looking for a dot in the end of the string

2d. Now that the nitty gritty details of the different data sets are fixed, we want to make a subset of each data set only containig the variables that we wish to use in the final data set. For this we use the tidyverse package dplyr, which contains the function select().

The variables we need are: Child.ID, Visit, Ethnicity, Diagnosis, Gender, Age, ADOS, MullenRaw, ExpressiveLangRaw, MOT\_MLU, MOT\_LUstd, CHI\_MLU, CHI\_LUstd, types\_MOT, types\_CHI, tokens\_MOT, tokens\_CHI.

* ADOS indicates the severity of the autistic symptoms (the higher the score, the worse the symptoms)
* MLU stands for mean length of utterance
* types stands for unique words (e.g. even if "doggie" is used 100 times it only counts for 1)
* tokens stands for overall amount of words (if "doggie" is used 100 times it counts for 100)
* MullenRaw indicates non verbal IQ
* ExpressiveLangRaw indicates verbal IQ

It would be smart to rename the last two into something you can remember (i.e. nonVerbalIQ, verbalIQ)

#renaming verbal IQ  
colnames(demo)[colnames(demo)=="MullenRaw"] <- "nonVerbalIQ"  
colnames(demo)[colnames(demo)=="ExpressiveLangRaw"] <- "VerbalIQ"  
  
#creating subsets  
selected\_demo <- select(demo, SUBJ, VISIT, Ethnicity, Diagnosis, Gender, Age, ADOS, nonVerbalIQ, VerbalIQ)  
selected\_lu <- select(lu, SUBJ, VISIT, MOT\_MLU, MOT\_LUstd, CHI\_MLU, CHI\_LUstd)  
selected\_token <- select(token, SUBJ, VISIT, types\_MOT, types\_CHI, tokens\_MOT, tokens\_CHI)

2e. Finally we are ready to merge all the data sets into just one. Tip: Look up "joins" in R for data science, or google "How to merge datasets in R" Tip: Joining / merging only works for two data frames at the time.

#merging  
merged\_data1 <- join(selected\_demo, selected\_lu, by= c("SUBJ", "VISIT"))  
fulldata <- join(merged\_data1, selected\_token, by= c("SUBJ", "VISIT"))  
  
#inner, left, right, full joins are options

Are we done? If you look at the data set now, you'll se a lot of NA's in the variables ADOS, nonVerbalIQ (MullenRaw) and verbalIQ (ExpressiveLangRaw). These measures were not taken at all visits. Additionally, we only want these measures for the first visit (Riccardo will explain why in class). So let's make sure that we select only these variables as collected during the first visit for each child and repeat these values throughout all other visits.

* A way to get around this in R, is to make a subset of the data containing only the values we are interested in, i.e. the values at visit 1. We only want to keep the relevant variables in this data set, i.e the ones which values are to be repeated. Either the the subset() function or select() and filter() can be used here. Solve this issue with both functions to familiarize yourself with these useful functions. N.B. save the subset of the dataset as a new dataset, do not overwrite the old one.
* In order to merge these new variables to the final data set, they'll need new names. E.g change the ADOS variable to ADOS1.
* Once you've changed the names, the subset can be merged to the final data set, and the score at the 1st visit in each variable will be spread to all 6.
* Lastly, there are too many unneccesary variables in the data set by now. Use the select() to choose only the variables you want in the data (e.g. remove the old ADOS, verbal and nonVerbal IQ variables, so you will not get confused later on) and define the order of the variables. Hint: You want the child identifier, followed by diagnosis, followed by demographic, cognitive and clinical features, followed by indexes of linguistic performance (utterances lenghts, types and tokens of words).

#using subset()  
subset1 <- subset(fulldata, VISIT == "1", select = c("SUBJ", "VISIT", "ADOS", "nonVerbalIQ", "VerbalIQ"))  
  
#using filter()  
subset2 <- filter(fulldata, VISIT == "1")

## Warning: package 'bindrcpp' was built under R version 3.4.4

subset2 <- select(subset2, SUBJ, VISIT, ADOS, nonVerbalIQ, VerbalIQ)  
  
#renaming ADOS  
colnames(subset1)[colnames(subset1)=="ADOS"] <- "ADOS1"  
colnames(subset1)[colnames(subset1)=="nonVerbalIQ"] <- "nonVerbalIQ1"  
colnames(subset1)[colnames(subset1)=="VerbalIQ"] <- "VerbalIQ1"  
  
#merge  
fulldata <- join(fulldata, subset1, by= "SUBJ")  
  
#selecting relevant data  
relevant\_data <- subset(fulldata, select = -c(ADOS, nonVerbalIQ, VerbalIQ))  
  
#reordering  
ordered\_data <- relevant\_data[c("SUBJ", "VISIT", "Diagnosis", "Ethnicity", "Gender", "Age", "ADOS1", "nonVerbalIQ1", "VerbalIQ1", "MOT\_MLU", "MOT\_LUstd", "CHI\_MLU", "CHI\_LUstd", "types\_MOT", "types\_CHI", "tokens\_MOT", "tokens\_CHI")]

Now, we are almost ready to start analysing the data. However, here are some additional finishing touches:

* in some experiments your participants must be anonymous. Therefore we wish to turn the CHILD.ID into numbers.

Tip: You will probably need to turn it into a factor first, then a number Tip: google "R how to convert character to integer" or look up the as.??? functions

* In order to make it easier to work with this nice, clean dataset in the future, it is practical to make sure the variables have sensible values. E.g. right now gender is marked 1 and 2, but in two weeks you will not be able to remember, which gender were connected to which number, so change the values from 1 and 2 to F and M in the gender variable. For the same reason, you should also change the values of Diagnosis from A and B to ASD (autism spectrum disorder) and TD (typically developing). Tip: Try taking a look at ifelse(), or google "how to rename levels in R".

#anonymous data  
ordered\_data$SUBJ = as.factor(ordered\_data$SUBJ)  
ordered\_data$SUBJ <- as.numeric(ordered\_data$SUBJ)  
  
#change gender  
ordered\_data <- mutate(ordered\_data, Gender = ifelse (Gender %in% c("1"), "M", "F"))  
#  
ordered\_data <- mutate(ordered\_data, Diagnosis = ifelse (Diagnosis %in% c("a"), "ASD", "TD"))

Write the data set into a csv file.

write.csv(ordered\_data, file = "Portfolio1THEA.csv")

1. Now that we have a nice clean data set to use for the analysis next week, we shall play a bit around with it. The following exercises are not relevant for the analysis, but are here so you can get familiar with the functions within the tidyverse package.

Here's the link to a very helpful book, which explains each function: <http://r4ds.had.co.nz/index.html>

USING FILTER List all kids who: 1. have a mean length of utterance (across all visits) of more than 2.7 morphemes. 2. have a mean length of utterance of less than 1.5 morphemes at the first visit 3. have not completed all trials. Tip: Use pipes %>% to solve this

detach(package:plyr)  
  
#1.  
kids1 <- ordered\_data %>%  
 filter(CHI\_MLU > 2.7)  
unique(kids1)

#2.   
kids2 <- ordered\_data %>%  
 filter(CHI\_MLU < 1.5, VISIT == 1)  
unique(kids2)

USING ARRANGE

1. Sort kids to find the kid who produced the most words on the 6th visit
2. Sort kids to find the kid who produced the least amount of words on the 1st visit.

#1. showing the ranking  
kids4 <- ordered\_data %>%  
 filter(VISIT == 6) %>%  
 arrange(desc(tokens\_CHI))  
unique(kids4)

#descenting from highest to lowest  
  
#2. showing the ranking  
kids5 <- ordered\_data %>%  
 filter(VISIT == 1) %>%  
 arrange(tokens\_CHI)  
unique(kids5)

USING SELECT

1. Make a subset of the data including only kids with ASD, mlu and word tokens
2. What happens if you include the name of a variable multiple times in a select() call?

#1.  
kids6 <- subset(ordered\_data, Diagnosis == "ASD", select = c("CHI\_MLU", "tokens\_CHI", "SUBJ"))   
  
kids6 <- subset(ordered\_data, Diagnosis == "ASD", select = c(Diagnosis, CHI\_MLU, tokens\_CHI))  
kids6

#2.  
ordered\_data %>%  
 select(c("VISIT", "VISIT"))

#it only gets included once

USING MUTATE, SUMMARISE and PIPES 1. Add a column to the data set that represents the mean number of words spoken during all visits. 2. Use the summarise function and pipes to add an column in the data set containing the mean amount of words produced by each trial across all visits. HINT: group by Child.ID 3. The solution to task above enables us to assess the average amount of words produced by each child. Why don't we just use these average values to describe the language production of the children? What is the advantage of keeping all the data?

#1.   
ordered\_data <- ordered\_data %>%  
 mutate(mean\_spokenwords = mean(tokens\_CHI, na.rm = T))  
  
#2.   
ordered\_data <- ordered\_data %>%  
 group\_by(SUBJ) %>%  
 mutate(mean\_CHIwords = mean(tokens\_CHI, na.rm = T))  
  
#the advantage would be to remain the progress of the children's language