Experimental Methods 3 - Portfolio 1 & 2

Study Group “Dumbledore’s Army”: Astrid, Sebastian, Lisa, Fredrik & Lena

Portfolio 1

#1
#We got set up in GitHub

#2a
#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. We check the variable names:

- names(token_train)
- names(demo_train)
- names(LU_train)

#Then fit them to a common format

a) demo_train <- plyr::rename(demo_train, c("Child.ID" = "SUBJ" , "Visit"="VISIT"))

#2b
#homogenizing "visit"
#Remove letters "visit1." -> "1."

- LU_train$VISIT <- stringr::str_extract(LU_train$VISIT, "\\-*\\d+\\.*\\d*")
- token_train$VISIT <- stringr::str_extract(token_train$VISIT, "\\-*\\d+\\.*\\d*")

#Remove the dot after the number "1." -> "1"

- LU_train$VISIT <- as.numeric(LU_train$VISIT)
- token_train$VISIT <- as.numeric(token_train$VISIT)

#2c
#We also need to make a small adjustment to the content of the Child.ID column in the demographic data.
#Remove all punctuation

- token_train$SUBJ <- str_replace_all(token_train$SUBJ, c("Adam." = "Adam", "Albert." = "Albert", "Alfie." = "Alfie", "Allison." = "Allison","Annie." = "Annie", "Charles." = "Charles", "Dirk." = "Dirk", "Eduardo." = "Eduardo", "Frankie." = "Frankie", "Jack." = "Jack", "Jason." = "Jason", "Jerry." = "Jerry", "Johan." = "Johan", "John." = "John", "Judson." = "Judson", "Kara." = "Kara", "Kevin." = "Kevin", "Lester." = "Lester", "Luis." = "Luis", "Marius." = "Marius", "Milo." = "Milo", "Omar." = "Omar", "Rory." = "Rory", "Ryder." = "Ryder", "Tim." = "Tim", "Tina." = "Tina", "Todd." = "Todd", "Vick." = "Vick", "Witt." = "Witt"))
- LU_train$SUBJ <- str_replace_all(LU_train$SUBJ, c("Adam." = "Adam", "Albert." = "Albert", "Alfie." = "Alfie", "Allison." = "Allison","Annie." = "Annie", "Charles." = "Charles", "Dirk." = "Dirk", "Eduardo." = "Eduardo", "Frankie." = "Frankie", "Jack." =

"Jack", "Jason." = "Jason", "Jerry." = "Jerry", "Johan." = "Johan", "John." = "John", "Judson." = "Judson", "Kara." = "Kara", "Kevin." = "Kevin", "Lester." = "Lester", "Luis." = "Luis", "Marius." = "Marius", "Milo." = "Milo", "Omar." = "Omar", "Rory." = "Rory", "Ryder." = "Ryder", "Tim." = "Tim", "Tina." = "Tina", "Todd." = "Todd", "Vick." = "Vick", "Witt." = "Witt"))

#2d
#We now make a subset of each data set only containing the variables that we wish to use in the final data set.
select only these for a subset
#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.

- demo_train <- select(demo_train, SUBJ, VISIT, Ethnicity, Diagnosis, Gender, Age, ADOS, MullenRaw, ExpressiveLangRaw)
- LU_train <- select(LU_train, SUBJ, VISIT, MOT_MLU, MOT_LUstd, CHI_MLU, CHI_LUstd)
- token_train <-select(token_train, SUBJ, VISIT, types_MOT, types_CHI, tokens_MOT, tokens_CHI)

#Rename to nonVerbalIQ and verbalIQ

- demo_train <- plyr::rename(demo_train, c("MullenRaw" = "nonVerbalIQ" , "ExpressiveLangRaw"="verbalIQ"))

#2e
#Now we merge all the data sets into just one.
#Demo_train data includes some children with no MLU or TOKEN data -> These are cut off in the merging

- mergy <- merge(demo_train,LU_train)
- merg <- merge(mergy,token_train)

#Lastly
ADOS, nonVerbalIQ (MullenRaw) and verbalIQ (ExpressiveLangRaw) needs help. They have NA values for all the entries that aren't the first!

#First we make a subset containing the values

- sub_merg <- subset(merg, VISIT==1)

#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

- sub_merg <- plyr::rename(sub_merg, c("nonVerbalIQ" = "nonVerbalIQ1" , "verbalIQ" = "verbalIQ1", "ADOS" = "ADOS1"))

#Now we merge and make sure to keep all the SUBJ from the original merg file

- mergmaster <- merge(merg,sub_merg,by="SUBJ")

#Now we kill the old ADOS, VerbIQ and NonVerbIQ

- names(mergmaster)

- mergmaster <- select(mergmaster,SUBJ,VISIT.x,Ethnicity.x,Diagnosis.x,Gender.x,Age.x,ADOS1,nonVerbalIQ1,verbalIQ1,MOT_MLU.x,MOT_LUstd.x,CHI_MLU.x,CHI_LUstd.x,types_MOT.x,types_CHI.x,tokens_MOT.x,tokens_CHI.x)

#Getting rid of the x´s in the column headers
- mergmaster <- rename(mergmaster, Visits = VISIT.x, Ethnicity = Ethnicity.x, Diagnosis = Diagnosis.x, Gender = Gender.x, Age = Age.x, MOT_MLU = MOT_MLU.x, MOT_LUstd = MOT_LUstd.x, CHI_MLU = CHI_MLU.x, CHI_LUstd = CHI_LUstd.x,types_MOT = types_MOT.x, types_CHI = types_CHI.x, tokens_MOT = tokens_MOT.x, tokens_CHI = tokens_CHI.x)

#Getting rid of the 1´s in the column headers
- mergmaster <- rename(mergmaster, ADOS = ADOS1, nonVerbalIQ = nonVerbalIQ1, verbalIQ = verbalIQ1)

#Very last
#We will now Annonymize the kids
- mergmaster$SUBJ <- as.factor(mergmaster$SUBJ)
- mergmaster$SUBJ <-as.integer(mergmaster$SUBJ)

#For some reason the subj starts at 2 instead of 1 - so I subtract 1 from the whole coloumn
- mergmaster$SUBJ <-mergmaster$SUBJ-1

#change the values from 1 and 2 to F and M in the gender variable.
- mergmaster$Gender <- (ifelse(mergmaster$Gender == 1, "F", "M"))

#change the values of Diagnosis from A and B to ASD (autism spectrum disorder) and TD (typically developing).
- mergmaster$Diagnosis <- (ifelse(mergmaster$Diagnosis == "A", "ASD", "TD"))

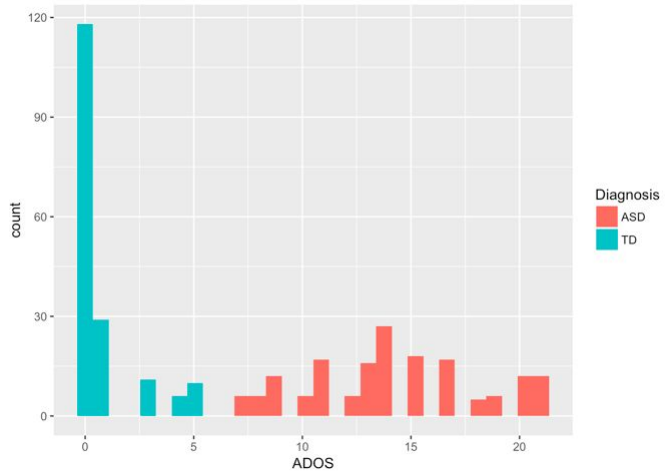
#Then we write the .CSV
- write.csv(mergmaster, file = "First_Assignment_3_Semester.CSV")

Portfolio 2

Exercise 1) Describe the participant samples in the dataset (e.g. by diagnosis, age, etc.). Do you think the two groups are well balanced? If not, what do you think was the reason?

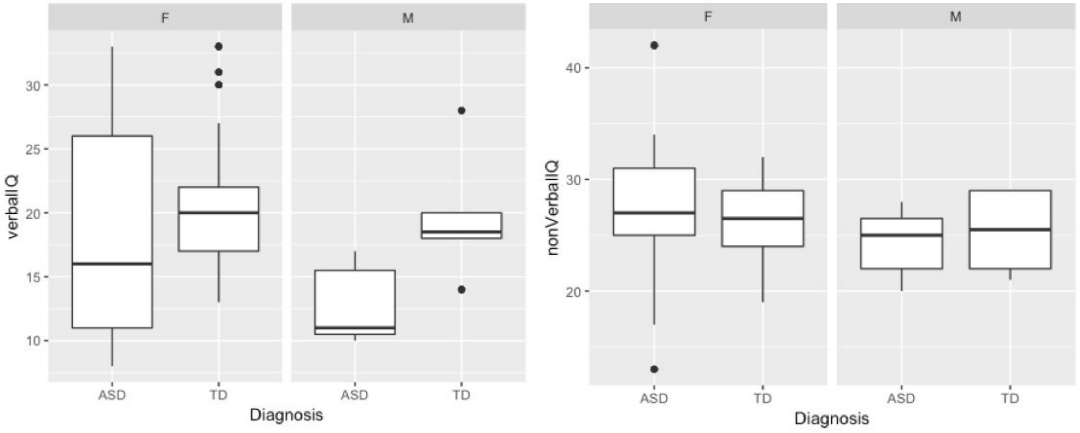
After visualizing the data and playing round with some some t-tests, we came to the following conclusions:
In these categories we found significant differences between the diagnosed autistic children group and the control: Total word count in the children, verbal IQ; unique words used for both mothers and children; mean length of utterance.

Age-wise, the ASD-group is significantly older than the normal group (the children are selected based on similar language-skills at visit one; thus this goes on to show, that the children with autism were lacking behind in their development from the beginning).
The autists differed clearly from the normal children on the autism scale.
Also there's a great variation regarding the ASD-score of ASD-group participants, since the Autism is a spectrum.



Interestingly, categories that did not differ between the groups includes non-verbal IQ; amount of words spoken by mothers. Gender is equally distributed in the two groups; and also relatively similar ethnicity-wise.

Here is one visualization that shows how Verbal IQ differed between the diagnosed autistic children group and the control, but nonVerbal IQ did not.

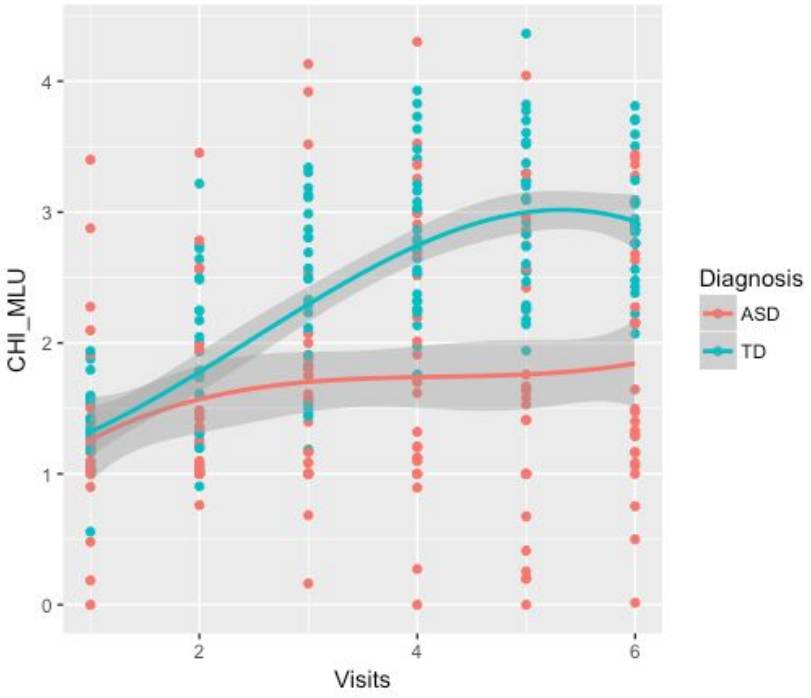


Exercise 2) Describe linguistic development in TD and ASD children in terms of Mean Length of Utterance (MLU)?

Diagnosis did not affect the MLU of the child as much as we thought it would over time. Not different from the null_model, which only includes Visits as a predictor, $\chi 2(1)=2.28$, $p=0.13$ (insignificant)

After playing around with growth models, we compared the model_1, the cubic and the quadratic and it turns out that the cubic model (p= 3.737e-09 and incredibly lower BIC as well as AIC scores) is our best choice:

```
## Data: Data
## Models:
## object: CHI_MLU ~ Diagnosis + Visits + (1 + Visits | SUBJ)
## ..1: CHI_MLU ~ Visits + I(Visits^2) + Diagnosis + (1 + Visits | SUBJ)
## ..2: CHI_MLU ~ Visits + I(Visits^2) + I(Visits^3) + Diagnosis + (1 +
## ..2:     Visits + I(Visits^2) + I(Visits^3) | SUBJ)
##        Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## object  7 590.09 616.89 -288.05   576.09
## ..1     8 561.67 592.30 -272.84   545.67 30.419      1  3.481e-08 ***
## ..2    16 522.30 583.56 -245.15   490.30 55.372      8  3.737e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Exercise 3) Describe how parental use of language changes over time in terms of MLU. What do you think is going on?

Diagnosis and Visits seems both to be sufficient predictors of Mother_MLU and compared to the null-model with anova() the model including both predictors was significantly better (p= 2.986e-05) at predicting Mother_MLU also, lower BIC and AIC score! It seems that the mothers adapt to the child's vocabulary/ability to learn new words in some sense.

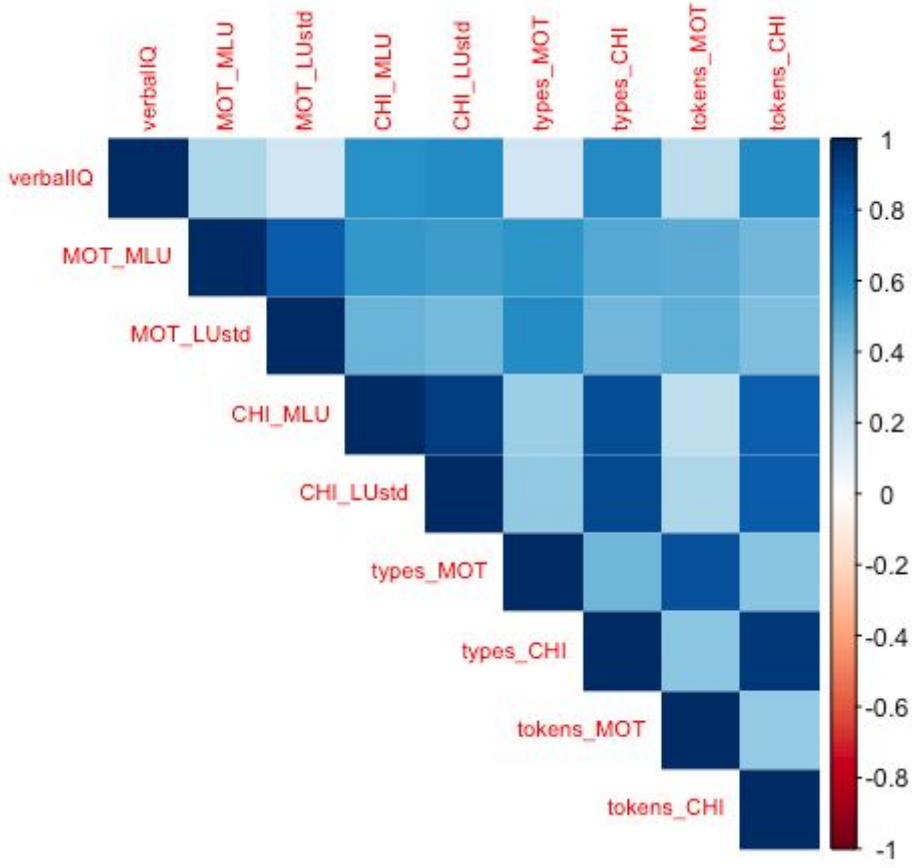
Exercise 4) Looking into "individual differences" (demographic, clinical or cognitive profiles)

We came to the conclusion that the most plausible predictors of language production (mean length of utterance) to be diagnosis, time, unique words, and verbal IQ; and keeping subject as random intercept, with time as random slope to accommodate that each child has a different starting point and progress differently.

We tested Data$types_CHI, Data$tokens_CHI for correlation and it turns out there is a very high correlation between these two. Supposedly, we can just include one. We picked types - we suspect the total amount of words spoken are more dependent on the mood of the child on the day, than the amount of unique words.

Also, we wanted to keep the model as simple as possible, so we only added verbal IQ, which seemed to us could be plausible predictor of CHI_MLU.

There were many variables to choose from, but we decided to pick those that to us seemed plausible, rather than p-hack.



Above is a correlational matrix that we conducted after creating our model. Looking into the CHI_MLU (child's mean length of utterance) row, we do indeed spot high correlation with the two other utterance-parameters: tokens_CHI (total words) and types_CHI (unique words).

Link to code:
Lena (Representative of the whole group for 'part 2'):

https://github.com/coolasacucumber/upgraded-lamp/blob/master/Sem3_AS1-Part2.Rmd

Lisa:
https://gitlab.com/superpowerLisa/almost_online/commit/ab28a768fe15f7d4503013aae4a78c32a110ef23#610dab264146f8c8b193337b19acefa3ddaf355f

https://gitlab.com/superpowerLisa/almost_online/

Sebastian:
https://github.com/sebsebar/Alouishes