Assignment 1 - Language Development in ASD - part 2

Signe Kløve Kjær and Thea Rolskov Sloth

July 7, 2017

Link to code:

<https://github.com/YpipY/EXPMETH3/blob/master/Assignment%201%20and%202%20(autism)/signe_thea_A1_P2_LangASD_trajectories_code_only.Rmd>

# 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.

This data is in the file you prepared in the previous class.

NB. A few children have been excluded from your datasets. We will be using them next week to evaluate how good your models are in assessing the linguistic development in new participants.

We then want to test the language trajectory of child and parent over time.

This RMarkdown file is structured in the following way:

1. The exercises: read them carefully. Under each exercise you will have to write your answers, once you have written and run the code. This is the part that you have to directly send to the teachers.
2. An (optional) guided template full of hints for writing the code to solve the exercises. Fill in the code and the paragraphs as required. Then report your results under the exercise part.
3. In exercise 4 you will be asked to create the best possible model of language development in TD and ASD children, picking and choosing whatever additional variables you want from the dataset. Next time, the models produced by the different groups will compete against each other to see who can produce the best model, so choose carefully!

You will have to have a github repository for the code and send the answers to Malte and Riccardo without code (but a link to your github/gitlab repository). This way we can check your code, but you are also forced to figure out how to report your analyses :-)

Remember to submit only your findings, and not just the code. To do this you can either - Write your answers in a separate document - Write your answers in the template, but tell rstudio not to print the code chunks when you knit it with the chunk option include=FALSE

## [1] 11

To re-iterate one more time: Hand in a document with your findings but without code (html or pdf or word) and a link to your github/gitlab with your Rmd file.

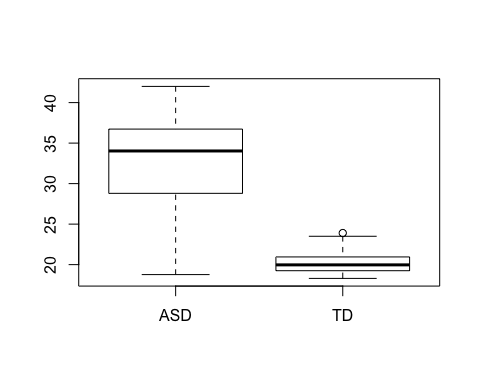
N.B. The following lines are a summary of the questions to be answered, the step-by-step instructions and tips are in the template.

## Exercise 1) Preliminary Data Exploration

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?

When analyzing the two participant samples according to diagnose, the boxplot indicates that the sampled children with autism are generally older than the sampled typically developed children. The average age of children wiht autism is 43.2 (SD = ). The average age of typically developed children is 30.6 (SD = ).

# Making three dataframs with only the first visit and for only the ASD and TD group  
visit1<- filter(data, VISIT==1)  
visit1ASD<- filter(data, VISIT==1 & Diagnosis=='ASD')  
visit1TD<- filter(data, VISIT==1 & Diagnosis=='TD')

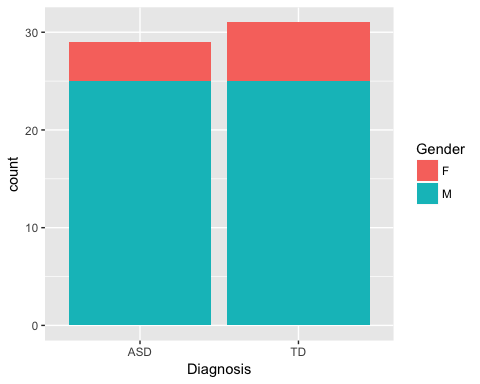


##   
## Welch Two Sample t-test  
##   
## data: visit1$Age by visit1$Diagnosis  
## t = 11.761, df = 31.823, p-value = 4.022e-13  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 10.43004 14.80080  
## sample estimates:  
## mean in group ASD mean in group TD   
## 32.99897 20.38355

## [1] 5.588385

## [1] 1.511186

The gender distributions between the two samples are well-matched. It is worth noticing that there is a larger proportion of males in both samples. The sampled children with autism is represented with 5 females and 26 males. For the sampled typically developed children, the distribution is 6 females and 30 males.

 Do you think the two groups are well balanced? If not, what do you think was the reason?

As the descriptive of the data claryfies, the two samples are matched according to "linguistic performance at visit 1". This is an explanation for why the two samples are not matched according to age. Children with autism compared to typically developed children reach the same level of linguistic performace at a later point in age.

On a linear model, this would not cause a problem for the predictive ability of the model. However when looking at quadratic and cubic models, where you might expect a learning curve to develop in a specific way in relation to age, it can cause a problem as we are looking at two different points on the curve. I.e., since they differ in age they are located at two different points of the curve, which will result in different slopes for the two samples.

As we want to predict the development in linguistic performance, we still consider it to be meanignful to match the samples according to MLU, rather than to match them according to age.

In the population there are a higher frequency of males with autism compared to females with autism. This difference in frequency is represented in the sample. Since this is well-balanced between the two samples, we do not consider this as a source of error.

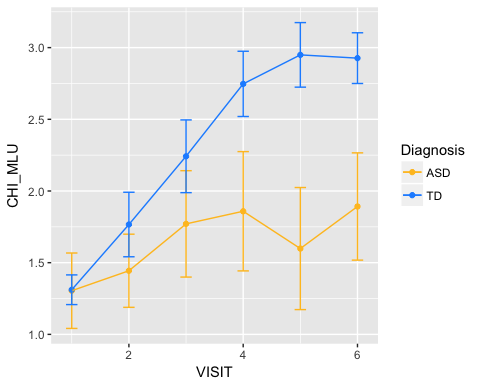
Generally we consider the two samples to be well-matched.

### Exercise 2) Children learning language: the effects of time and ASD

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

H1: The child's MLU (mean length of utterance) changes over time according to diagnosis.

H0: The child's MLU (mean length of utterance) does not change over time according to diagnosis.

 ```

null model is visit wihtout diagnosis

We modelled the data with linear, quadratic, and cubic mixed effects and concluded that the quadratic model fitted the data best.

Report R^2

Due to overfitting, the cubic model was not used.

We used a quadratic mixed effects model to predict the childs MLU from visit (indicator of time) according to diagnosis. As fixed effects we used visit and diagnosis with interaction in the model. As random effects we used random intercepts for subjects and by-subject random slopes for the effect of visit. By eyeballing residual plots we concluded that the assumptions of homoskedasticity and normality of residuals were met. We found that childrens means length of utterance significantly changes over time according to diagnosis: β (interaction) = 0.25 (SE = 0.04), t = 6.683, p < .05. P-values were obtained by likelyhood ratio test of the full model with time according to diagnosis as predictor against the null model without the effect of time according to diagnosis. The interaction effect between visit and diagnosis indicates that the diagnosis has an effect on how the childs mean length of utterance developes over time, this means that the effect of time is determined by the diagnosis of the child.  
Thus, we can reject the null-hypothesis, that the child's MLU does not change over time according to diagnosis. In the model 33.18 % of the variance is explained by the fixed effects and 85.91% of the variance can be explained by the fixed effects together with the random effects.

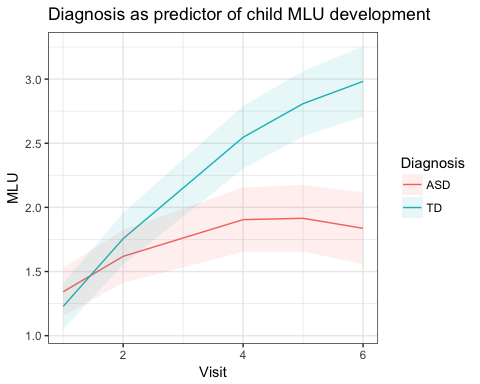
β (visit) = 0.41 (SE = 0.07), t = 5.51 β (diagnosis) = -0.35 (SE = 0.14), t = -2.46

Below is a plot of the model.

languageModel2 = lmer(CHI\_MLU ~ Diagnosis \* VISIT + I(VISIT^2) + (1+VISIT+ I(VISIT^2)|SUBJ), data=data, REML=FALSE)  
  
data$Diagnosis <- as.factor(data$Diagnosis)  
  
ee <- effect(c("Diagnosis","VISIT"),languageModel2)

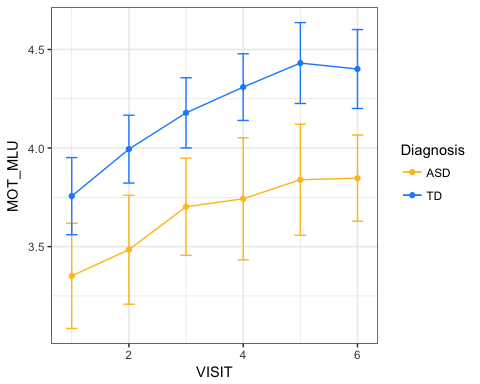
## NOTE: DiagnosisVISIT is not a high-order term in the model

theme\_set(theme\_bw())  
ggplot(as.data.frame(ee),  
 aes(VISIT,fit,colour=Diagnosis,fill=Diagnosis))+  
 geom\_line()+  
 ## colour=NA suppresses edges of the ribbon  
 geom\_ribbon(colour=NA,alpha=0.1,  
 aes(ymin=lower,ymax=upper))+  
 labs(title = "Diagnosis as predictor of child MLU development", x = "Visit", y = "MLU")



### Exercise 3) Child directed speech as a moving target

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



We used a quadratic mixed effects model to predict the parents MLU from visit (indicator of time) according to diagnosis. As fixed effects we used visit and diagnosis with interaction in the model. As random effects we used random intercepts for subjects and by-subject random slopes for the effect of visit.

By eyeballing residual plots we concluded that the assumptions of homoskedasticity and normality of residuals were met.

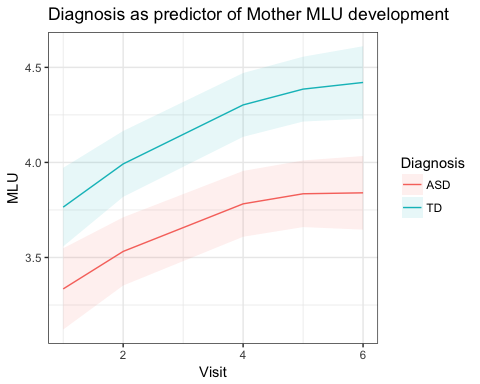
We found that parents means length of utterance significantly changes over time according to diagnosis: β(interaction) = 0.03 (SE = 0.03), t = 0.86, p < .05, which indicates that the way the mothers MLU is affected by time is determined by which diagnosis the child has. P-values were obtained by likelyhood ratio test of the full model with time according to diagnosis as predictor against the null model without the effect of time according to diagnosis. THereby we can reject the null hypothesis. The interaction effect however is rather small, which can also be eyeballed from the plot of the data. Where it seems like the slope of the parents MLU over time is similar.

In the model 23.83 % of the variance is explained by the fixed effects and 69.34% of the variance can be explained by the fixed effects together with the random effects.

languageModelMot = lmer(MOT\_MLU ~ Diagnosis \* VISIT + I(VISIT^2) + (1+VISIT+ I(VISIT^2)|SUBJ), data=data, REML=FALSE)  
  
ee <- effect(c("Diagnosis","VISIT"),languageModelMot)

## NOTE: DiagnosisVISIT is not a high-order term in the model

theme\_set(theme\_bw())  
ggplot(as.data.frame(ee),  
 aes(VISIT,fit,colour=Diagnosis,fill=Diagnosis))+  
 geom\_line()+  
 ## colour=NA suppresses edges of the ribbon  
 geom\_ribbon(colour=NA,alpha=0.1,  
 aes(ymin=lower,ymax=upper))+  
 labs(title = "Diagnosis as predictor of Mother MLU development", x = "Visit", y = "MLU")



The samples of the children are matched according to MLU. The ideal situation would be that the mothers' MLU would follow the same slope as the childrens', due to language theories that suggest, that the parents MLU should systematically follow the childs MLU, but always at a higher level in order to appropriately challenge the child. The plot displays how the sampled mothers have differet starting points. We can suspect that since the parents have been told that their child has autism, they have changed the way that they address their child. Thus, the diagnosis has possibly affected the mothers mean length of utterance when talking to the child, regardless of the childs actual linguistic competences.

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

The dataset contains some additional variables characterizing the kids’ cognitive and clinical profile: ADOS (autism severity), MSEL EL (Expressive Language, that is, verbal IQ, or linguistic skills at first visit as assessed by a psychologist using Mullen Scales of Early Learning), MSEL VR (Visual Reception, used as a proxy for non verbal IQ at first visit), Age, Gender, Ethnicity. Would it make sense to add any of them to your model of linguistic trajectories? Create the best possible model (the one that best explain the data, with MLU as outcome). Next time your model will be tested on new participants, and we will proclaim a winner. Describe your strategy to select the best models (how did you choose the variables to include?) and send the code to Riccardo and Celine.

When trying to create the best possible model, we considered if the variables were correlated and thereby would account for much of the same variance. Also, whether the model would just be a better model of the data instead of a model with a good predictive power for new unseen datasets.