# Portfolio 1

## Experimental Methods 3: Multilevel models and machine learning

E22.147201U013.A

Submission: 12. October 2022

Student: Sara Kjær Kristensen

Email: au702259@uni.au.dk

Student no: 202105320

School of Communication and Cognition, University of Aarhus

Nordre Ringgade 104 4 th, 8200 Aarhus N, Denmark

Parts of this assignment have been made in a group consisting of Freddy Wulf (FW), Ida Møller (IM), Maria Mujemula (MM), Sabrina Zaki (SZ) & Sara Kjær Kristensen (SK).

E.g. Introduction (MM, SK)

Maria wrote the main part and Sara adapted it.

The main parts of the assignment have been discussed in the mentioned group, and written by SK. The group discussions will be addressed as 'we' or 'the group' hereafter.

Character count with spaces w/o plots: 9.050

## Table of content

## Introduction

This portfolio will cover handed out data[1] investigating autistic and neurotypical children's language development over a series of 6 visits with several months apart. At each visit, different measures were taken including the binary diagnostic grouping, either autistic or neurotypical, and the child's mean length of utterance during that visit. To investigate the dataset R (version 4.2.0 & 4.2.1) with RStudio (version 2022.07.1) and statistical packages for Bayesian workflow BRMS (Bürkner, 2021).

## Simulating the data at hand

To better understand the data and the models, that are going to describe it, a new dataset is simulated using the literature estimates as a guideline. The dataset is structured with the most important variables and leave the others for consideration; number of visits (6); number of participants in the two diagnostic conditions (50, sample size = 100); and mean length of utterance (MLU).

To estimate the priors, we used the values given in the assignment as a starting point for the MLU for each diagnostic group at visit 1, average individual variability in this initial MLU, the average change in MLU across visits for the diagnostic groups, and average individual variability for the diagnostic groups. To modify the priors, we make histograms to check if the values give a realistic impression of the data, by trial and error we estimate the priors to be informed priors as they assume a difference in the development of MLU. Generally, we use a lognormal distribution to exclude the possibility of the MLU going below zero.

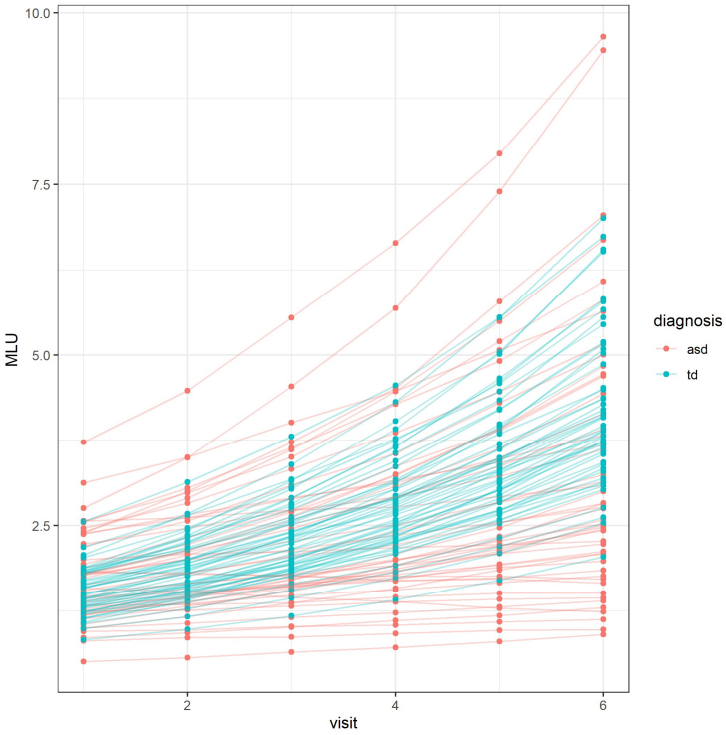



*Figure 1 - Simulated data showcasing MLU development over time*

[1] From Fusaroli, et al (2019)

To assess the simulated data and what predictors are best to describe the change in MLU, we formulate 3 models:

| Model no. | Equation | Description |
|---|---|---|
| 1 | $MLU \sim 0 + diagnosis$ | MLU modulated by diagnosis |
| 2 | $MLU \sim 0 + diagnosis + diagnosis{:}\,visit$ | MLU modulated by diagnosis with a difference in change per visit pending on diagnosis |
| 3 | $MLU \sim 0 + diagnosis + diagnosis{:}\,visit + (1 + visit \mid ID)$ | MLU modulated by diagnosis with a difference in change per visit pending on diagnosis and individual differences per visit |

The priors are specified as needed for the different models and fitted to their model:

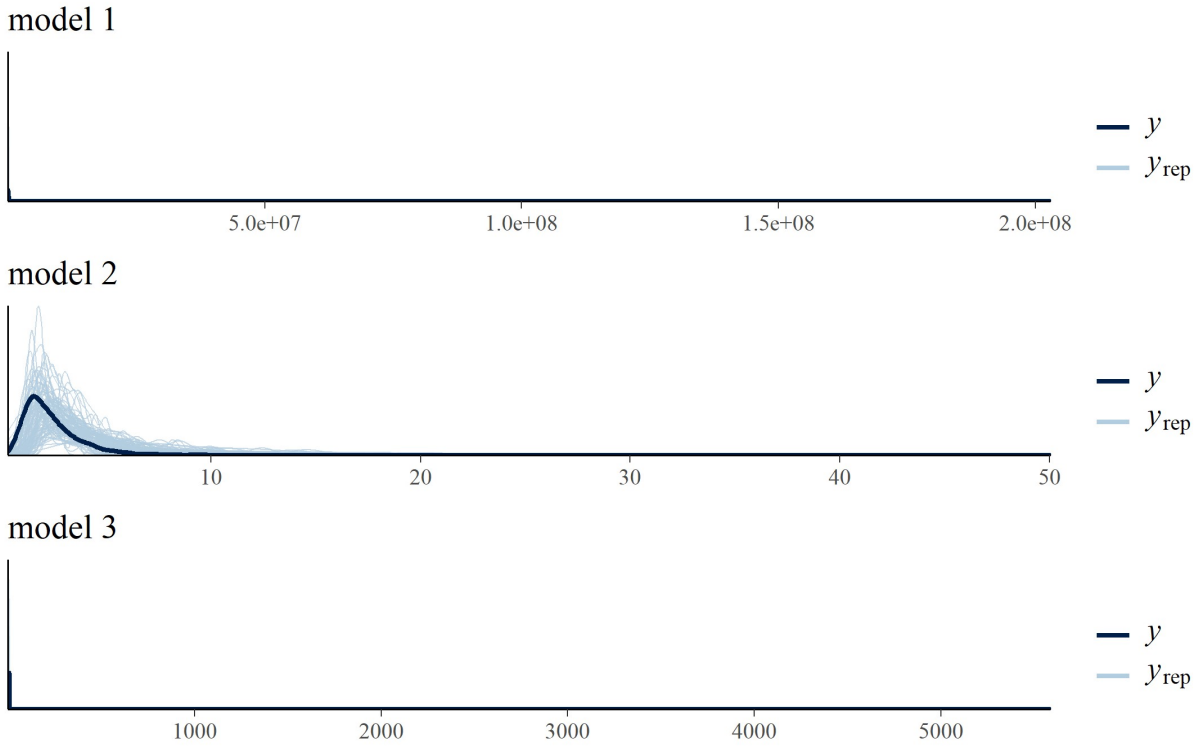



*Figure 2 - Informed priors*

Fitting the data to the posterior distributions gives these prior posterior update checks. This way the data's influence can be inspected.

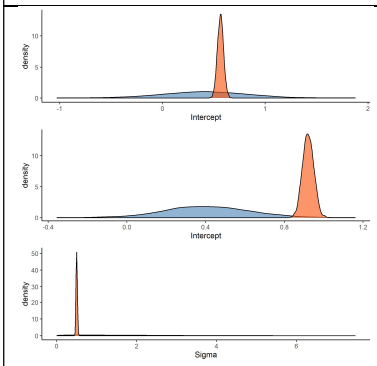
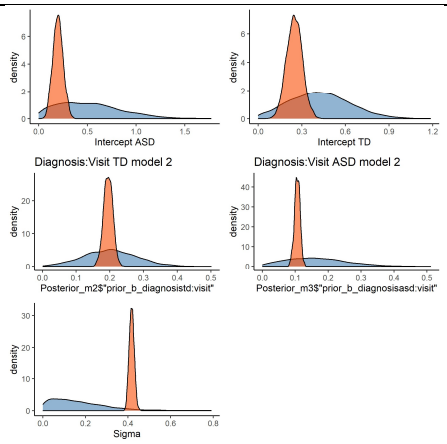
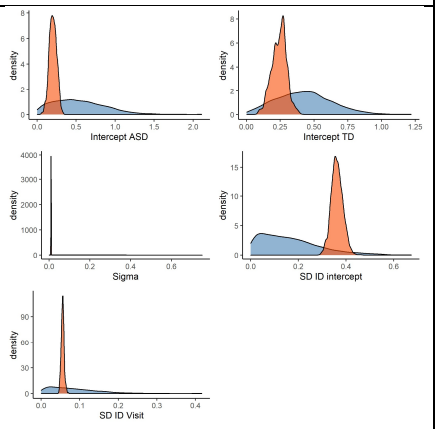
| Model 1 | Model 2 | Model 3 |
|---|---|---|
|  |  |  |

*Table 1 - blue for prior and orange for posterior*

Checking the quality of the models is assessed by looking for a ratio spread for the Markov Chain Monte Carlo (Rhat) value between 0.90 and 1 and as high values as possible for effective sample sizes for both the bulk (Bulk_ESS) and tail (Tail_ESS).

| Model no. | Summary output | Esti-mate | Est.Er-ror | l-95% CI | u-95% CI | Rhat | Bulk_ESS | Tail_ESS |
|---|---|---|---|---|---|---|---|---|
| | Population-Level Effect | | | | | | | |
| | diagno-sisasd | 0.56 | 0.03 | 0.51 | 0.62 | 1.00 | 2088 | 1534 |
| 1 | diagno-sistd | 0.92 | 0.03 | 0.8 | 0.98 | 1.00 | 1803 | 1845 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.50 | 0.01 | 0.47 | 0.53 | 1.00 | 1839 | 1667 |
| 2 | Population-Level Effect | | | | | | | |
| | diagno-sisasd | 0.20 | 0.05 | 0.10 | 0.30 | 1.00 | 1320 | 1262 |
| | diagno-sistd | 0.25 | 0.05 | 0.14 | 0.35 | 1.00 | 1207 | 1205 |
| | diagno-sisasd:visit | 0.11 | 0.01 | 0.08 | 0.13 | 1.00 | 1288 | 1887 |
| | diagno-sistd:visit | 0.20 | 0.01 | 0.17 | 0.22 | 1.00 | 1552 | 1374 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.42 | 0.01 | 0.40 | 0.44 | 1.00 | 2009 | 2251 |
| 3 | Population-Level Effects | | | | | | | |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | diagno-sisasd | 0.20 | 0.05 | 0.11 | 0.29 | 1.01 | 103 | 239 |
| | diagno-sistd | 0.24 | 0.05 | 0.14 | 0.35 | 1.05 | 44 | 78 |
| | diagno-sisasd:visit | 0.11 | 0.01 | 0.09 | 0.12 | 1.01 | 117 | 132 |
| | diagno-sistd:visit | 0.20 | 0.01 | 0.18 | 0.21 | 1.13 | 16 | 152 |
| | Group-Level Effects (~ID) | | | | | | | |
| | sd(Inter-cept) | 0.36 | 0.02 | 0.31 | 0.41 | 1.02 | 105 | 167 |
| | sd(visit) | 0.06 | 0.00 | 0.05 | 0.06 | 1.02 | 144 | 295 |
| | cor(Inter-cept:visit) | 0.03 | 0.09 | -0.14 | 0.23 | 1.02 | 114 | 179 |
| | Family Specific Parameters | | | | | | | |
| | Sigma | 0.01 | 0.00 | 0.01 | 0.01 | 1.00 | 754 | 1206 |

*Table 2 - Comparing model summary output*

Model 1's parameters is overall okay, but the estimates are very different from our priors. Model 2's parameters are better both for estimates and mcmc specific parameters. Lastly, model 3's estimates are fine, but the mcmc specific parameters are concerning; the Rhat is above 1, Bulk_ESS is in many case below a 100 samples same goes for the Tail_ESS. Keeping these values in mind, divergency plots are consulted:

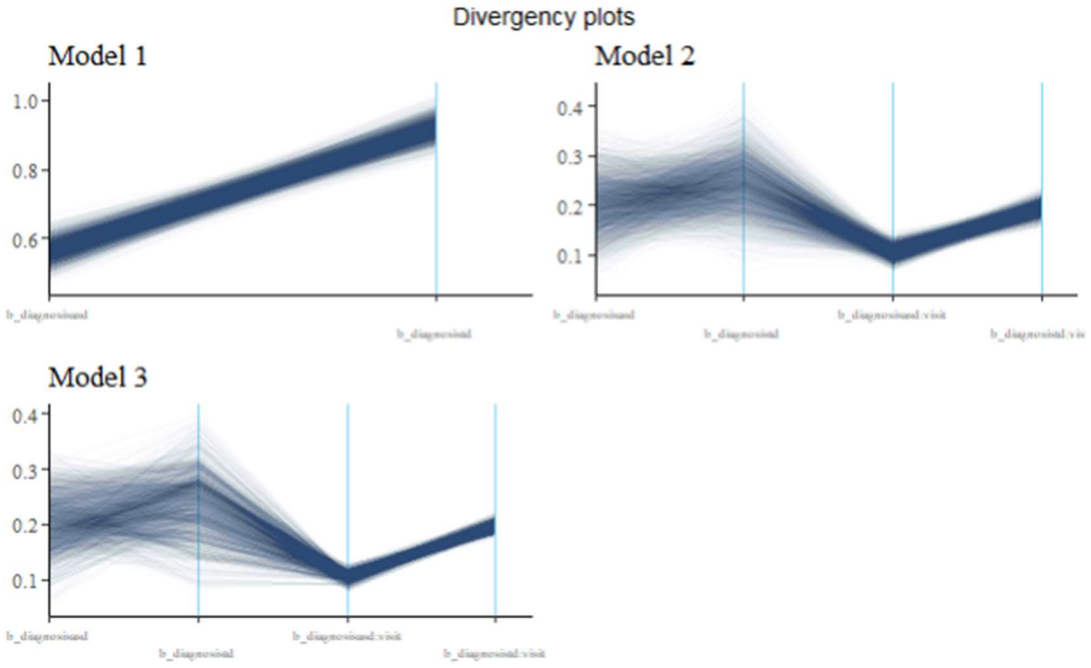



*Figure 3 - Divergency plots*

The divergency plots points towards model 2 instead of the other models because of the spread in the plots. However, consulting other seeds and my group gives another view on the models and by doing model comparison with the leave-one-out method, is a bit more in favour of model 3:

| Model no. | elpd_diff | se_diff | weight |
|---|---|---|---|
| 1 | -2270.2 | 23.8 | 0.000 |
| 2 | -2166.4 | 26.1 | 0.000 |
| 3 | 0.0 | 0.0 | 1.000 |

*Table 3 - Model comparison with LOO*

Expected log pointwise predictive density difference (elpd_diff) compares models in ascending order, and the negative numbers suggest that the function favours the previous. Logically follows that model 3 is the better one as the leave-one-out log score for predictive density (weight) is maximized by using this model solely. Taking all of this into account and adding the theoretical layer, it makes sense to have children within a certain diagnostic group developing at their own pace. On top of that, the group's general view of model 3 as being the best describing model. Model 3 is chosen for further empirical analysis.

Besides comparing models, the sample size might have a substantial effect on the model's estimate. As the simulated data is already using 3 times as many data points as the empirical data for practical reasons. However, this is an unrealistic picture to paint as real-world data collection would properly not contain 50 participants per condition:

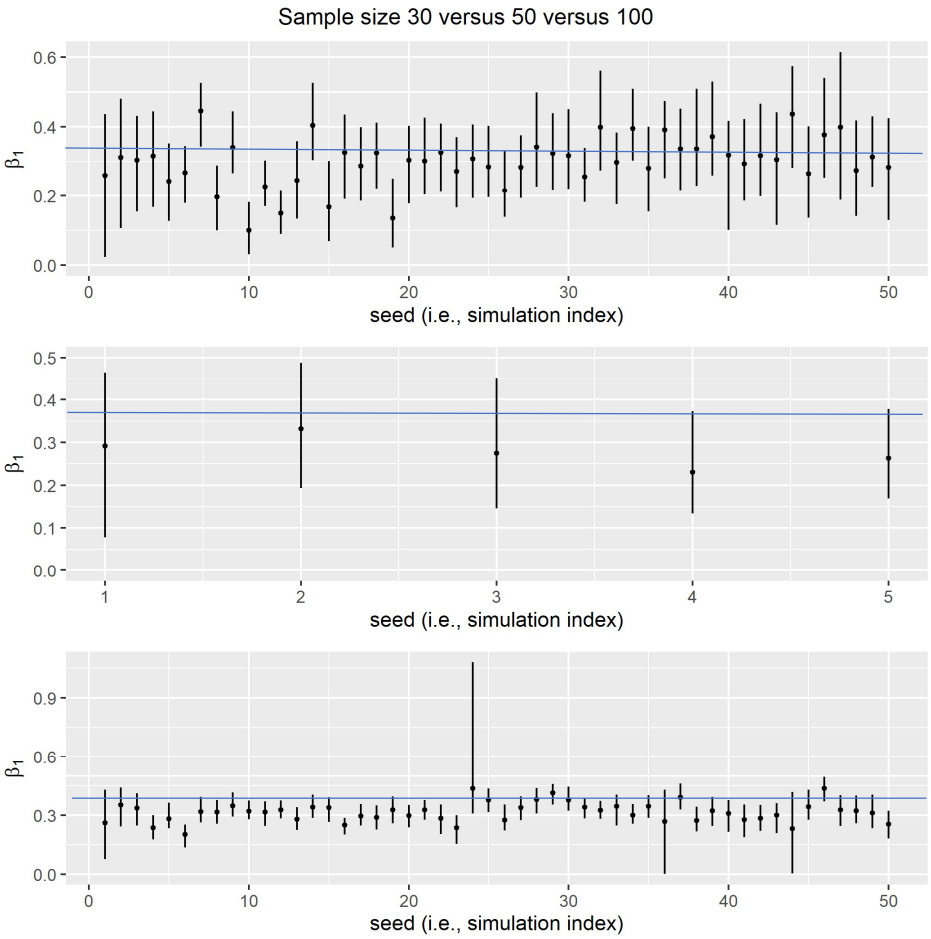



*Figure 4 - simulation of sample size and seed, the blue line resembles estimates for model 3's standard deviation of intercepts (0.36)*

These simulation plots give a slight insight into how different seeds for random sampling for 30, 50 and 100 participants per group gives different estimates of the different ASD intercepts. Overall, more data points give more certainty, but 30 participants per group seem to be a sufficient place to start. Looking at the power estimates does not give further enlightenment:

| Variable | Power estimate |
|---|---|
| Diagnosis ASD | 1 |
| Diagnosis TD | 1 |
| Diagnosis ASD : visit | 1 |
| Diagnosis TD : visit | 1 |

*Table 4 - Power estimates*

Besides simulating across seeds, changing the priors to weakly informed prior to make the data work more to make a convincing case of change and difference between visits and diagnostic groups. On that note, it is time to dive into the empirical data.

## Empirical data

Like many other cases, the simulation does not capture all the noise in the data set. Comparing the simulated and empirical data the difference is clear as is the general tendency in language development.

Again, our prior belief is updated to a posterior distribution of the estimates, this time model 3's parameters are updated on the empirical data. Prior-posterior update plots are as follows:

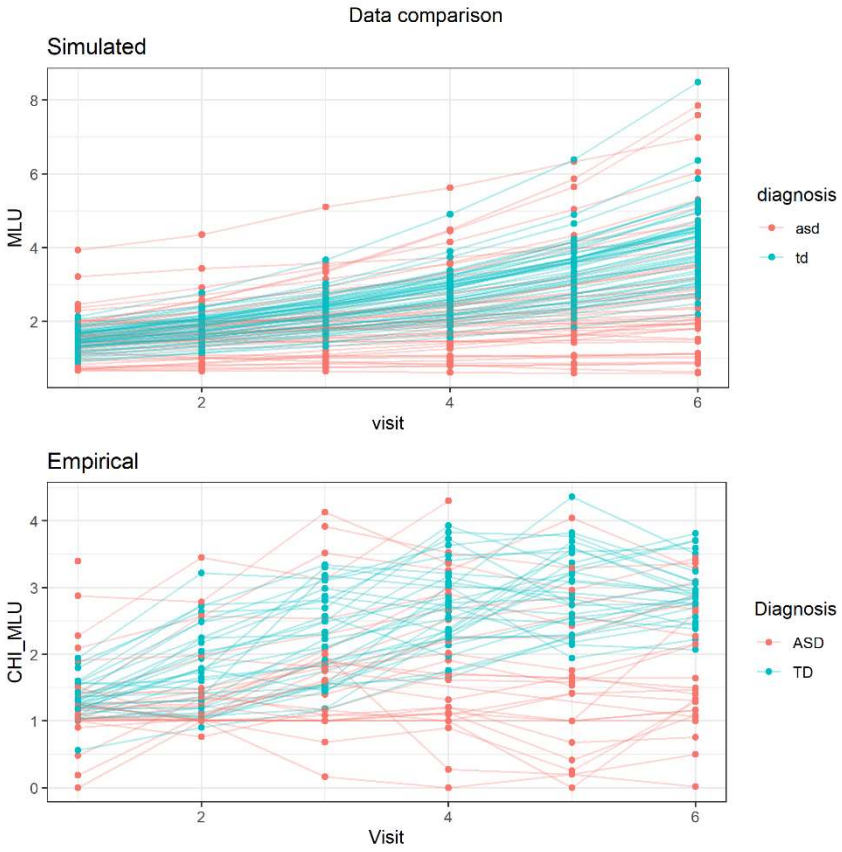



*Figure 5 - comparing data's MLU development over time*

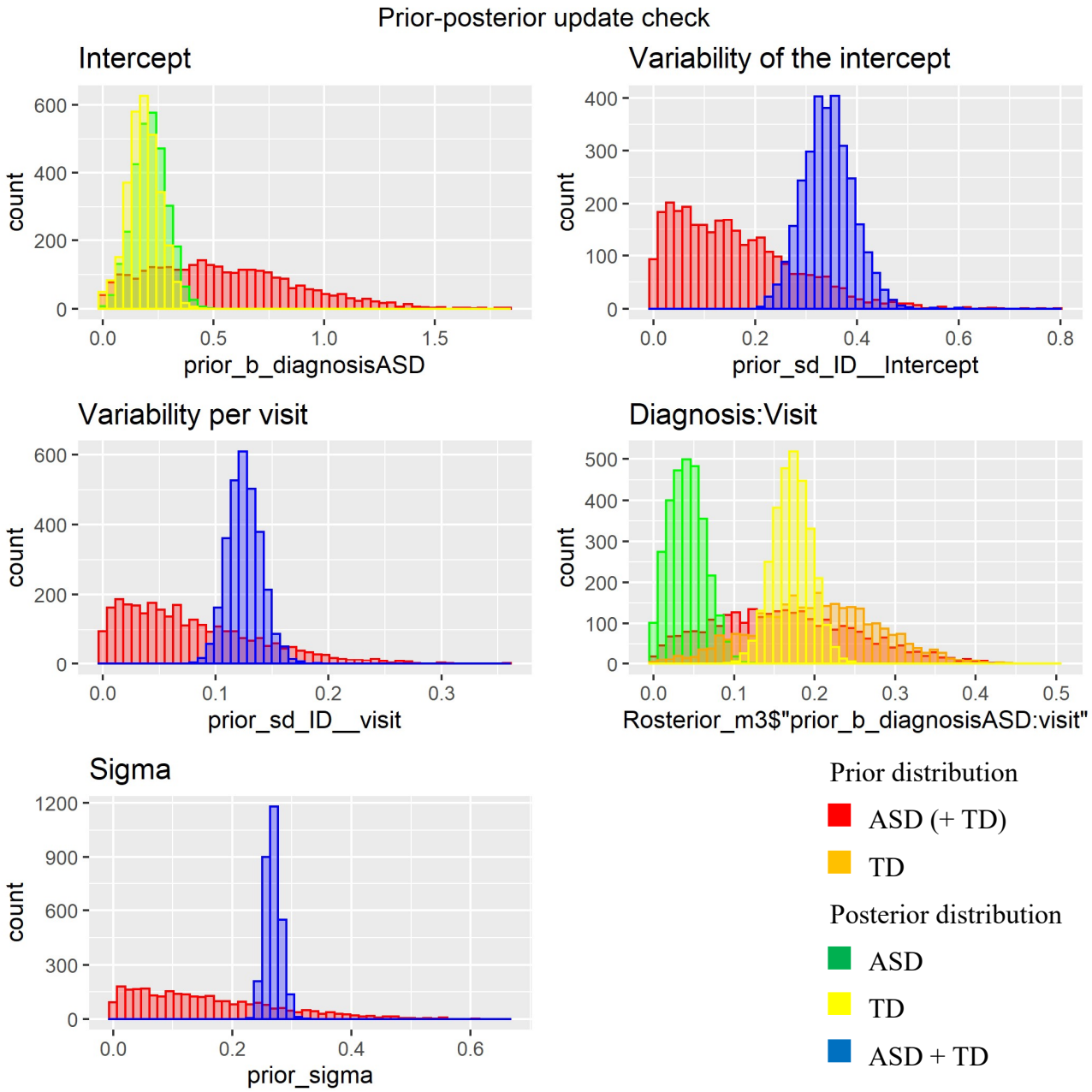



*Figure 6 - Prior-Posterior update plots with empirical data*

Before we dive into the exciting difference in change in MLU over time, as indicated in plot *Diagnosis:Visit*, we will take a look at the model's qualities. First of all, the model:

$$MLU \sim 0 + diagnosis + diagnosis{:}\,visit + (1 + visit|ID)$$

predicts the mean length of utterance (MLU), calculated as the mean of morphemes in the utterance divided by the number of words in the utterance during the visit (30 min), by the binary diagnostic grouping (autism spectrum disorder (ASD) or typically developed (TD)), interaction effect of diagnostic group and visit as participants in the different groups may vary differently in acquired language over time, and lastly with individual random effects for each participant over visits. This way the

participants' MLU can account for just bad days or overall smaller changes in MLU. The model's summary highlights the RHat, Bulk_ESS and Tail_ESS as mentioned before.

| Group-Level Effects | | | |
|---|---|---|---|
| Parameters | Rhat | Bulk_ESS | Tail_ESS |
| sd(Intercept) | 1 | 1023 | 1574 |
| sd(visit) | 1.01 | 344 | 823 |
| cor(Intercept,visit) | 1.01 | 285 | 689 |
| Population-Level Effects | | | |
| diagnosisASD | 1.00 | 897 | 1074 |
| diagnosisTD | 1.00 | 781 | 424 |
| diagnosisASD:visit | 1.00 | 580 | 721 |
| diagnosisTD:visit | 1.00 | 514 | 804 |
| Family Specific Parameters | | | |
| Sigma | 1.00 | 1590 | 1836 |

*Table 5 - Model summary*

The values do not give rise to much concern except the standard deviation for visit and correlation between intercept and visit in the group-level effects, as their Rhat-values are above 1. With this concern we consult sensitivity plots for the variability over time for the two diagnostic groups to inspect our prior's impact on the posterior estimates:

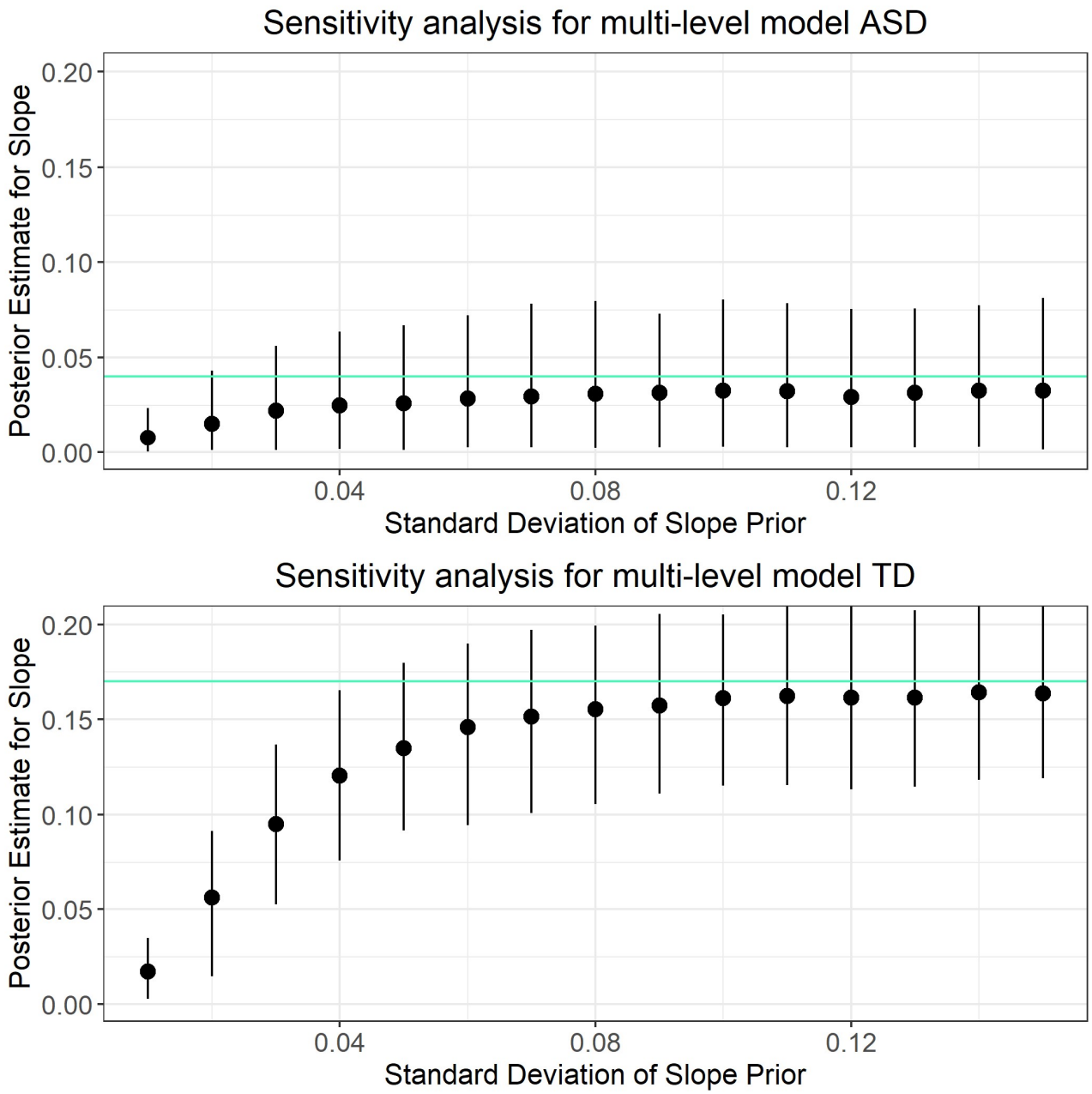



*Figure 7 - Sensitivity check of prior-posterior estimates for variability over time for ASD and TD (MU). Lines reflects model posterior estimate diagnosisASD:visit (0.04) and diagnosisTD:visit (0.17) (SK).*

The informed prior's standard deviation specified for our model and model summary above was set to 0.1 and 0.08 for ASD and TD respectively. In these plots, the posterior estimate for ASD (0.04) and TD (0.17) is better estimated with optimistic priors. For this analysis, the prior slope standard deviation was set to 0.1 for ASD and 0.08 for TD. Given the prior for slope standard deviation was

set to 0.1 for ASD the model gets a good estimate with a coherent credible interval. As do the prior standard deviation for TD (0.08) gives a decent interval of the posterior estimate of the slope.

Now returning to plot *Diagnosis:Visit* in the prior posterior update checks, the model summary is investigated.

| Prior | | | Posterior | | | |
|---|---|---|---|---|---|---|
| Estimate | Error | Variable | Estimate | Error | l-95% CI | u-95% CI |
| Population-level | | | | | | |
| 0.41 | 0.41 | diagnosisASD | 0.21 | 0.08 | 0.07 | 0.36 |
| 0.41 | 0.22 | diagnosisTD | 0.18 | 0.08 | 0.03 | 0.32 |
| 0.15 | 0.1 | diagnosisASD:visit | 0.04 | 0.02 | 0.00 | 0.09 |
| 0.2 | 0.08 | diagnosisTD:visit | 0.17 | 0.02 | 0.13 | 0.22 |
| Group level | | | | | | |
| 0 | 0.2 | sd(Intercept) | 0.34 | 0.05 | 0.25 | 0.44 |
| 0 | 0.1 | sd(visit) | 0.13 | 0.01 | 0.10 | 0.16 |
| | | cor(Intercept,visit) | -0.46 | 0.14 | -0.68 | -0.15 |

*Table 6 - Prior and posterior estimates for population and group level effects*

To test the first hypothesis about whether the development across diagnostic groups, the hypothesis() function is used for "diagnosisTD:visit is greater than diagnosisASD:visit". The output gives an estimate of 0.13 with a 95% credible interval of 0.08-0.18, and posterior probability of a 100%. The posterior probability reflects sampling 1 random participant from each group and finding that the TD has a higher MLU is 100 % likely. This result is reflected in the prior-posterior update plots for diagnosis(ASD/TD):visit above and in the conditional effect plot below.

To test the second hypothesis for the starting value across diagnostic groups, same function for "diagnosisASD < diagnosisTD". The output of 0.03 with a 95% credible interval of -0.13-0.20, and posterior probability of 40%. Formulating last the hypothesis as the individual development for ASD is smaller than the development for TD. The output gives an estimate of 0.13 with a 95% credible interval of 0.08-0.18, and posterior probability of a 100% across all children. Which funny enough is the same as for the population-level. However, all in all indicating that the diagnostic slopes are statistically different at the population-level:

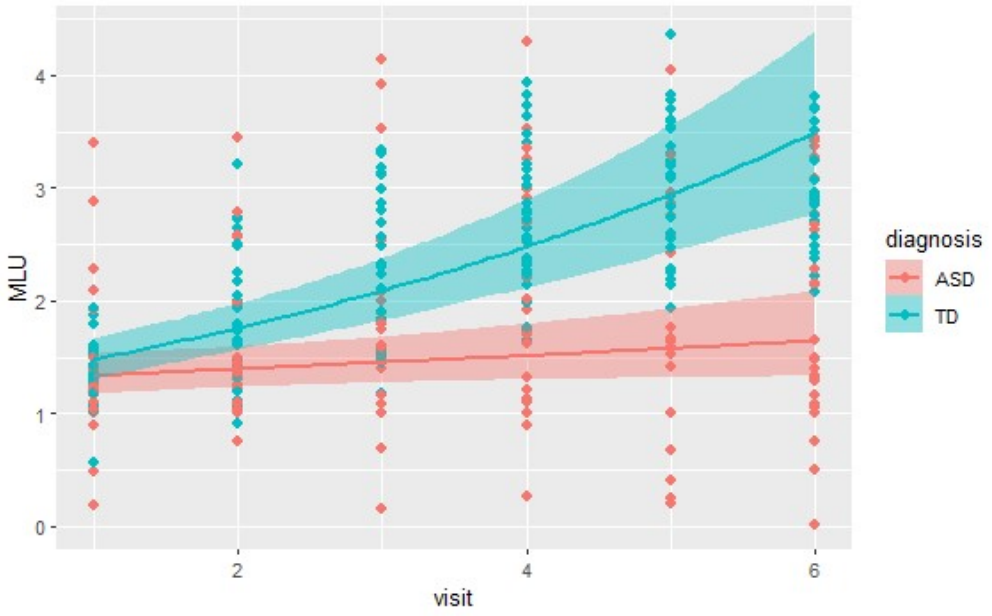



*Figure 8 - conditional effects of the development of MLU over time depending on the diagnostic group*

Some additional factors would be interesting to add to the model, as the child's own effect on her language can hardly account for her linguistic development. E.g., the mother's MLU would be interesting to consider as well as it we know that children learn from their environment and if the environment or environment controller (in this case the mother) decides what the child is exposed to and therefore decides the syllabus on the upbringings linguistic course. If the child is inhibited from exposure to more diverse linguistic environments, the child could quickly reach a plateau in the linguistic acquirement or make up her own language. The latter part is probably more likely if a 3-year-old could make up her own language and syntax. On the other hand, if the child's environment is stimulating and has a lot to offer, then the development in MLU would reflect more on the child's ability to acquire linguistic concepts rather than it being a lack of resources.

## Reference

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2019). Hearing me hearing you: Reciprocal effects between child and parent language in autism and typical development. *Cognition*, *183*, 1–18. https://doi.org/10.1016/j.cognition.2018.10.022

## Materials

RStudio Team (2022). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Bürkner P (2021). "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software*, **100**(5), 1–54. doi:10.18637/jss.v100.i05.