Justina Razanauskaite
202109601

Assignment 1: Language development

Part 1

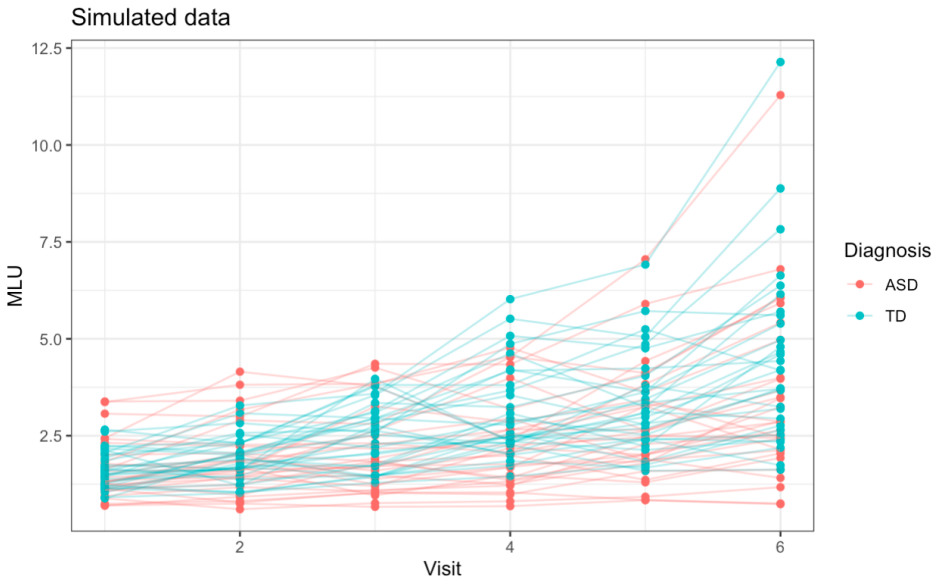
Question 1:

- *Briefly describe your simulation process, its goals, and what you have learned from the simulation.*
- *Add at least a plot showcasing the results of the simulation.*
- *Make a special note on sample size considerations: how much data do you think you will need? What else could you do to increase the precision of your estimates?*

Simulation: goals and process

One of the goals of the simulation is to identify whether our created model could recover true parameter values/estimates of the phenomena we are interested in. It would allow us to try different things on the simulated data, and let us to indicate the best tools we could use in order to analyze and perform the right inference on the real data. Moreover, simulation helps us to understand the process of inference, gives a better picture of the outcome that is generated by certain values, and gives opportunity to think about the conditions in which the known true values can be captured or not.

To begin with, the given initial values were used to simulate the data. As the outcome values were not the ones that are expected, those initial values were adjusted until the outcome approximately captured the phenomena (MLU) of interest, keeping the overall difference between the populations in place.

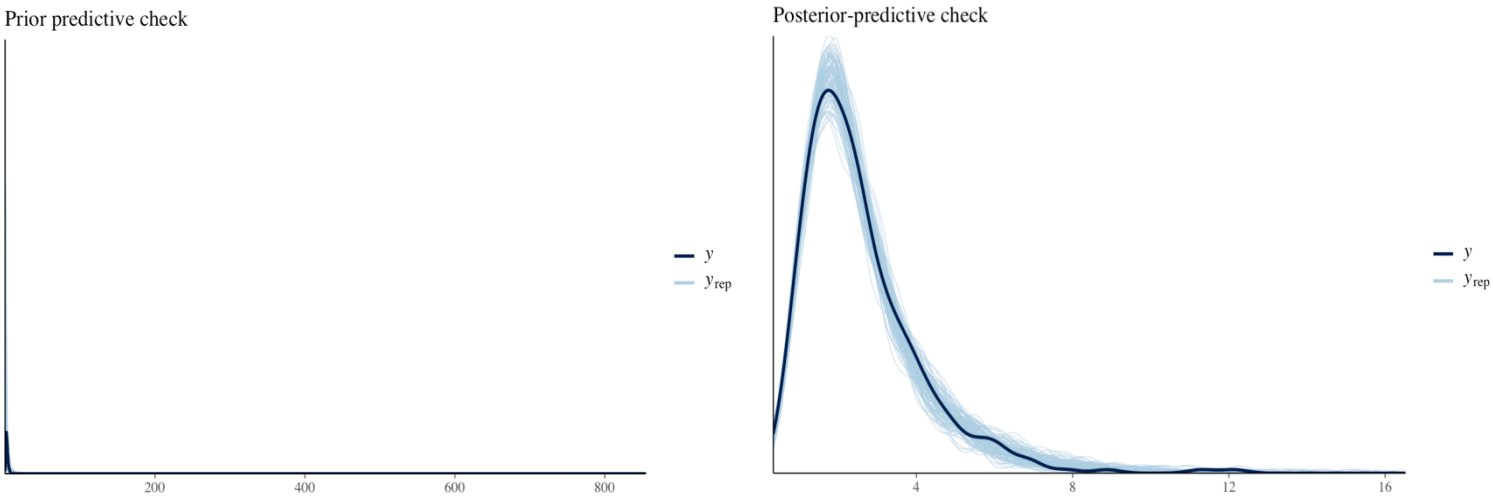


The MLU values plotted in the figure above seem to be logical enough to capture the phenomena. It does not provide any extreme values and captures possible trajectory of MLU "growth" for both groups, namely, autistic disorder (ASD) and typically developing (TD) children. From the visualization we can infer the differences between the populations, and how individual-level parameters affect the expected development of both groups.

In order to set up the model for our data analysis, I defined a formula that describes our outcome of interest (MLU). The formula that has been used in the model is:

$$MLU \sim 0 + Diagnosis + Diagnosis: Visit + (1 + Visit \mid ID).$$

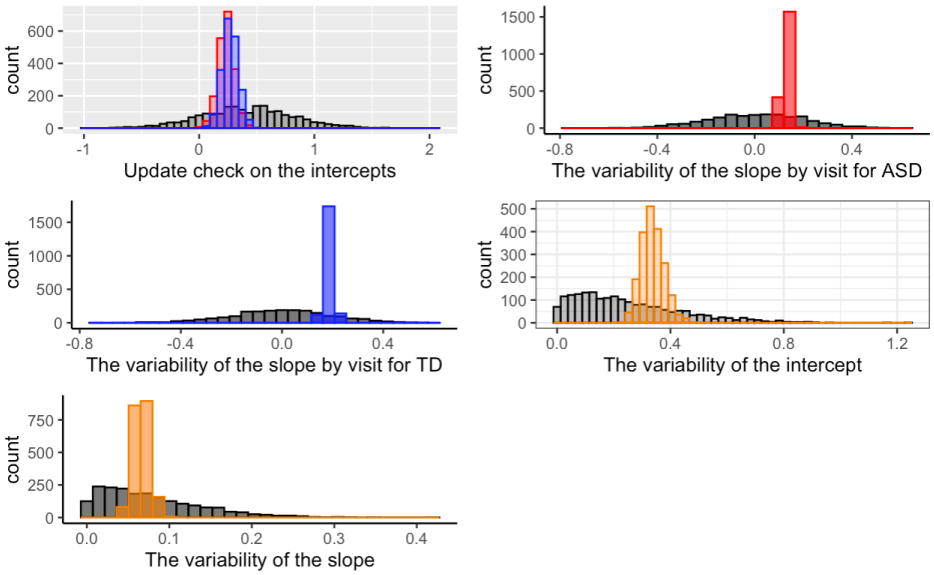
In short, with this formula it was indicated that each category of diagnosis (ASD and TD) should have own estimate for the intercept with the same amount of uncertainty, that both groups would have their own slope at each visit which would depend on the diagnosis, and that all of the parameters should be different by child, who would also have its own intercept and slope.



In this figure, prior predictive check indicates that the MLU values of 200 and above are possible outcome values, as all individual variations were added. Allowing model to be "exposed" to the simulated data, it narrows down the range of possible values, and shows pretty good fit.

Going further with the simulation process, prior-posterior update checks were performed to see whether the defined priors fit. The prior distribution is indicated in grey, and the distribution of ASD and TD is shown in red and blue, respectively. Orange distribution indicates both groups.

Justina Razanauskaite
202109601

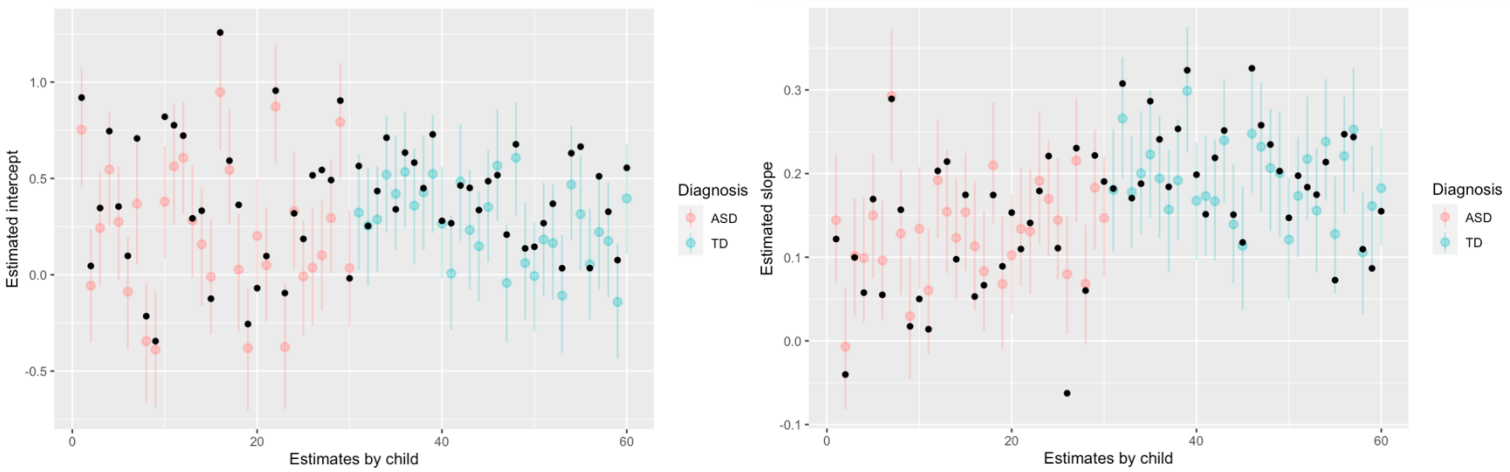


By looking at the plots, it does not seem like posterior is pushing against the prior, posterior distributions are withing the range of the prior, and they are way "pointier" than the prior, which indicates that the model has learned from the data and is more confident.

```
 Family: lognormal
  Links: mu = identity; sigma = identity
Formula: MLU ~ 0 + Diagnosis + Diagnosis:Visit + (1 + Visit | ID)
   Data: df (Number of observations: 360)
  Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 2000

Group-Level Effects:
~ID (Number of levels: 60)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)           0.34      0.04     0.27     0.42 1.00     1011     1379
sd(Visit)               0.07      0.01     0.05     0.09 1.00      914      985
cor(Intercept,Visit)   -0.03      0.18    -0.36     0.34 1.00      830      935

Population-Level Effects:
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
DiagnosisASD          0.23      0.07     0.10     0.37 1.00      915     1223
DiagnosisTD           0.27      0.07     0.14     0.40 1.00      817     1308
DiagnosisASD:Visit    0.13      0.01     0.10     0.16 1.00     1533     1727
DiagnosisTD:Visit     0.19      0.02     0.16     0.22 1.00     1411     1607

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.19      0.01     0.18     0.21 1.00     1325     1398

Draws were sampled using sample(hmc). For each parameter, Bulk_ESS
and Tail_ESS are effective sample size measures, and Rhat is the potential
scale reduction factor on split chains (at convergence, Rhat = 1).
```

Looking at the model summary for ASD and TD slopes per visit, it seems like it successfully recovered true parameter values, which are in the range of CIs for both groups.

Justina Razanauskaite
202109601



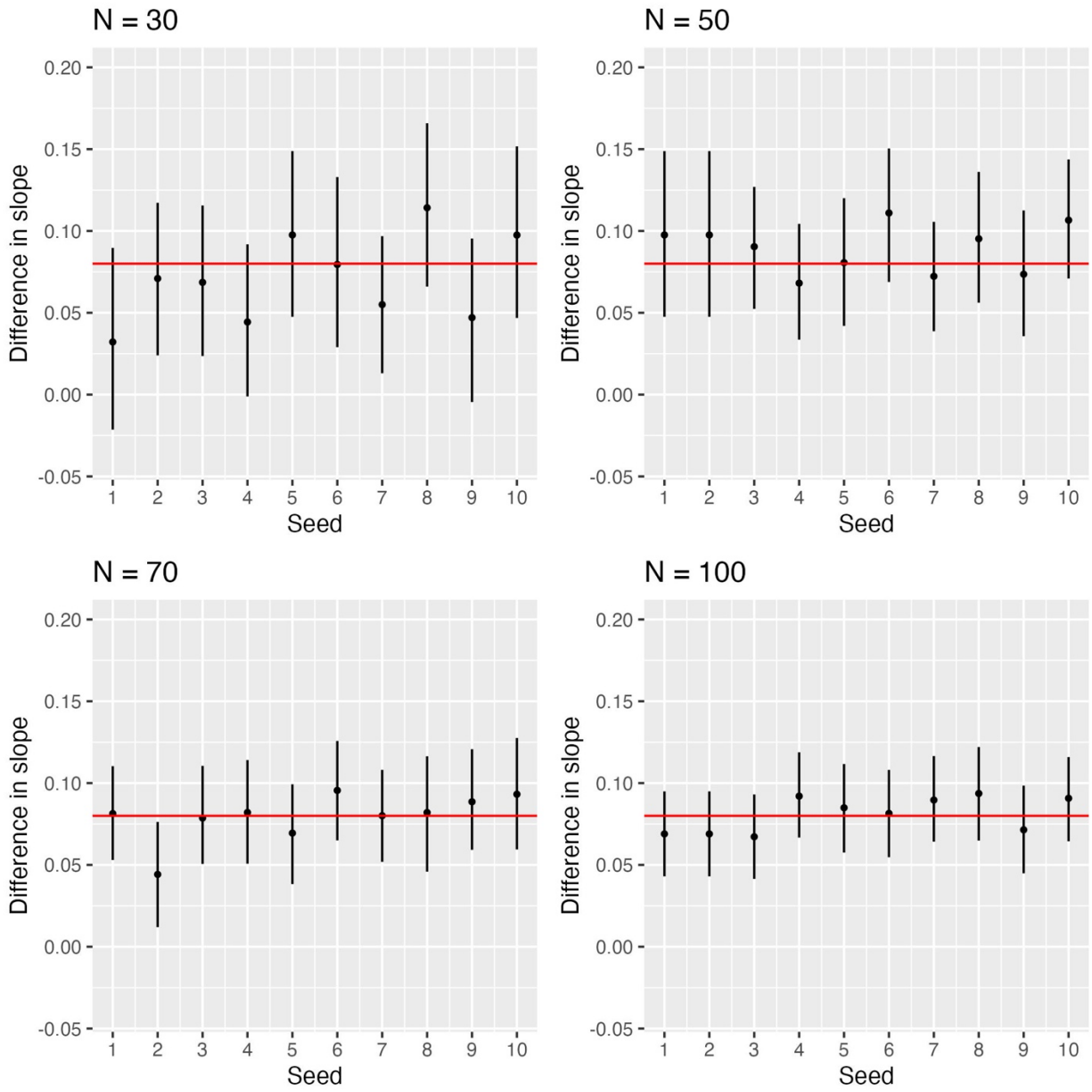
By looking at these plots, it seems like compatibility intervals almost always include the true value, and overall, the model captures the general trend of MLU across both groups. Intercept is broader for ASD group than TD, and TD has a greater slope (faster development) than ASD.

Power analysis

To investigate the effect that the sample size might have on precision of the estimates, the power analysis was conducted. For the simplicity, I ran only 10 simulations for each chosen sample size. Running 100 and more simulations would give more precise results, however, I chose to stick to 10.

Here, I simulated data of sample size of 30, 50, 70 and 100. For this task, I found it hard to come up with a formula that would perform the power analysis automatically, therefore I went through the steps one by one. I am aware that when running more simulations (for instance, 100) it would take too long to adjust all of the code that is needed, however, for this specific case, I chose to do it. Therefore, the results of the power analysis are shown below.

Justina Razanauskaite
202109601



Looking at these plots, it seems like almost all estimates capture the true effect (the difference in slopes between TD and ASD) within their compatibility interval. The only instance where it is not captured, is in N=70 (2). It looks like a sample size of 50 (per each group) would be the most optimal size, as it does not include 0 in their compatibility intervals (whereas N=30 does).

Justina Razanauskaite
202109601

In short, the simulations could be summarized this way:

Sample size	Include 0	Include true effect size	Power
N = 30	3	10	70% - 0.7
N = 50	0	10	100% - 1
N = 70	0	9	90% - 0.9
N = 100	0	10	100% - 1

As the “traditional” threshold for power is 80% (0.8), according to these results, the sample size of 50, 70, and 100 is enough to capture the phenomena. It looks like with more and more samples the confidence becomes greater, nevertheless, in real life setting, it might be difficult to have big sample sizes.

To ensure as much power as possible, one might want to include additional factors to the model. One of them might be more dense frequency of the visits at which the development of the children is assessed. Another thing that could be done during the study is thinking more about the set-up of how and in what way the children are assessed. It might be the case that the real language development is not captured as accurately as it might be, as the child is exposed to unusual environment and might tend to speak less.

Justina Razanauskaite
202109601

Part 2

Question 2:

- *Briefly describe the empirical data and how they compare to what you learned from the simulation (What can you learn from them?).*
- *Briefly describe your model(s) and model quality.*
- *Report the findings: how does development differ between autistic and neurotypical children (N.B. remember to report both population and individual level findings)?*
- *Which additional factors should be included in the model?*
- *Add at least one plot showcasing your findings.*

Describing the sample of empirical data

The table below describes the sample at each visit:

	Visit 1	**Visit 2**	**Visit 3**	**Visit 4**	**Visit 5**	**Visit 6**
N (ASD)	28	28	28	26	25	28
N (TD)	32	32	31	32	31	28
Age (mean)						
ASD	32.7	37.5	41.4	45.3	50	53.3
TD	20.4	24.7	28.8	32.7	36.8	41.1
Gender						
ASD	3 F, 25 M	4 F, 24 M	4 F, 24 M	3 F, 23 M	4 F, 21 M	3 F, 25 M
TD	6 F, 26 M	6 F, 26 M	6 F, 25 M	6 F, 26 M	6 F, 25 M	4 F, 24M
MLU (mean)						
ASD	1.35	1.44	1.77	1.93	1.66	1.89
TD	1.31	1.76	2.2	2.73	2.97	2.92
LU (SD)						
ASD	0.646	0.76	1.04	1.25	1.17	1.33
TD	0.645	1.1	1.54	1.96	2.13	2.19
Socialization (mean)						
ASD	77.29	76.5	77.4	77.2	77.4	78.3
TD	100.8	99.3	104.2	103.4	103.4	99.9

Justina Razanauskaite
202109601

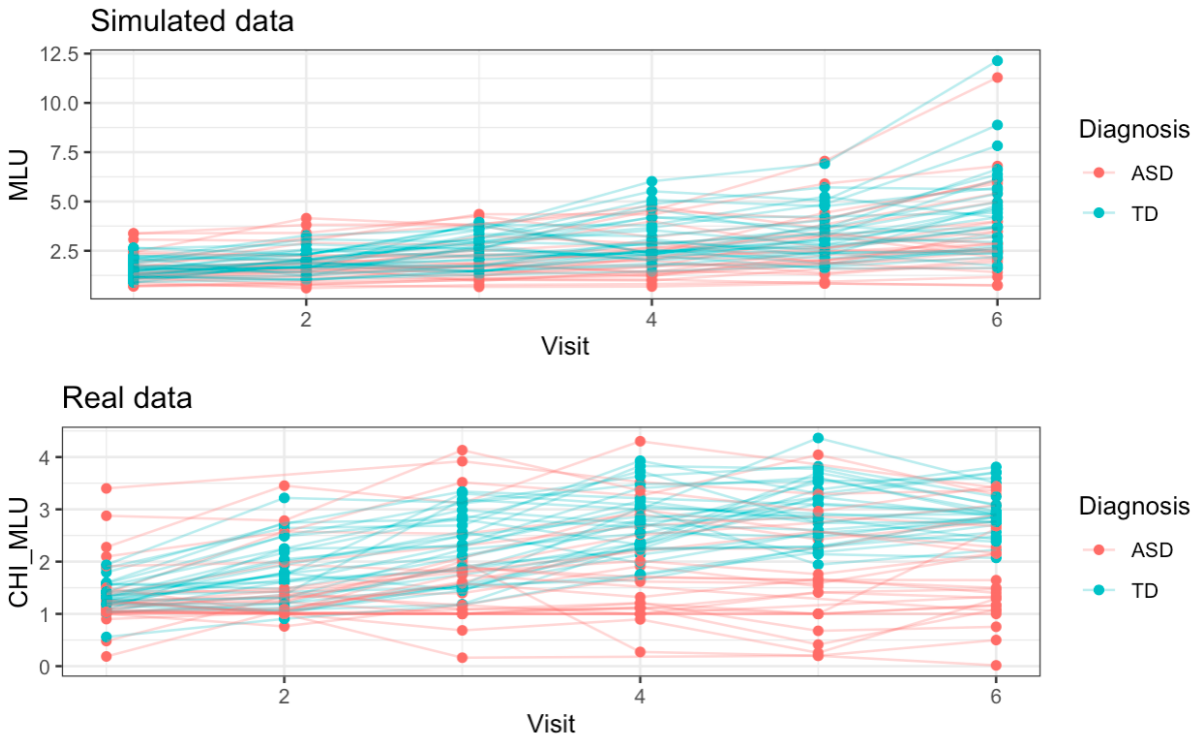
Just by looking at the summaries of the diagnosis per each visit, it seems like the groups are not balanced perfectly. There is a variation in sample size by each visit, and there is a clear variation in sample size by gender. It also might be better if the ASD and TD children could've been matched by age, as here, at each visit, TD group (on average) is always younger than ASD. Of course, it might not be a problem in capturing the degree of development.

At visit 1, on average, the MLU was higher for ASD than TD. This changed completely in the following visits, resulting in greater development for TD compared to ASD (even if TD children are younger than ASD). With regards to socialization skills, ASD group is generally lower score compared to TD group, however, it seems like skills on average increase for autistic group from Visit 1 to Visit 6, whereas neurotypical group has a lower mean score in Visit 6 compared to Visit 1.

With regards to the sample size and the power analysis described above, it might be better to have a bigger sample size in this case. Power analysis revealed that the size of 30 had power of 0.7, and the empirical data consists of 25-32 samples per each group, therefore, having a sample size of 40, or 50, might provide more accurate results.

Simulated versus empirical data

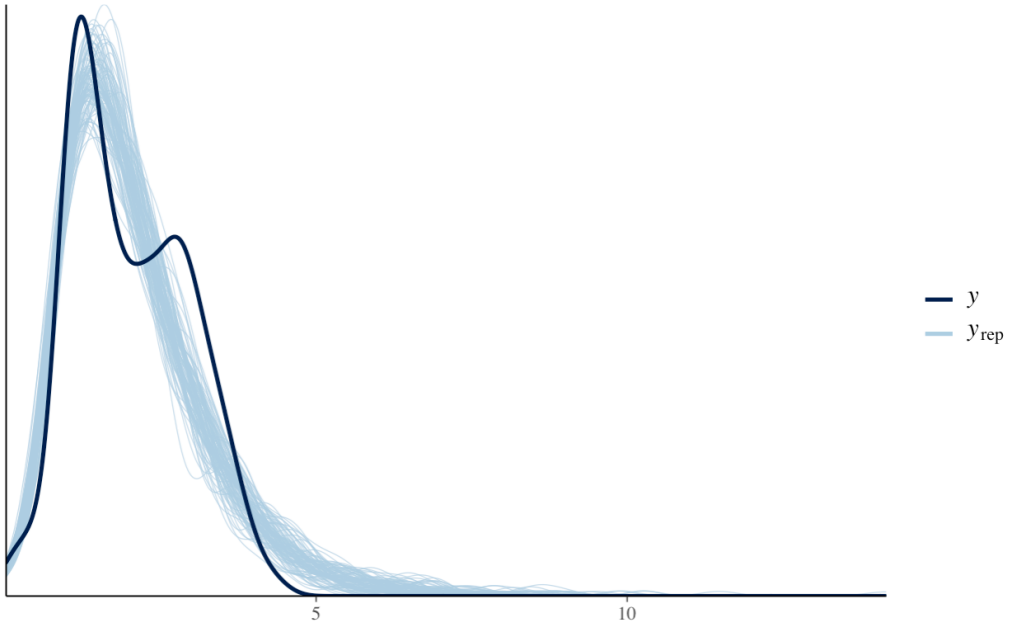
The plots of simulated and real empirical data look somewhat different:



Justina Razanauskaite
202109601

Looking at the plot of simulated data, both groups seem to gradually grow in regards to the value of MLU. In addition, the ASD group shows more variability compared to TD group, but not as much as it might be expected. At visit 4 (and later ones) it is clear that TD group is mostly above the ASD group, which is logical as TD should develop faster than ASD. However, it is a clear difference in the outcome value at the final visit between the simulated and the real data. Real data captures the gradual development of the TD group, nevertheless, it does not indicate values as high as in the simulation. In both, the simulated and real data, the starting point at visit 1 is approximately the same across both populations. In the later one it is also clear that ASD group has a way greater variability than the TD group, however, does not show as much evidence in the development as it is indicated in the simulation. These two plots help us to understand the impact of our chosen parameter values that try to generate real-world data, and indicate that in reality there is much more variability and less extreme development in both populations.

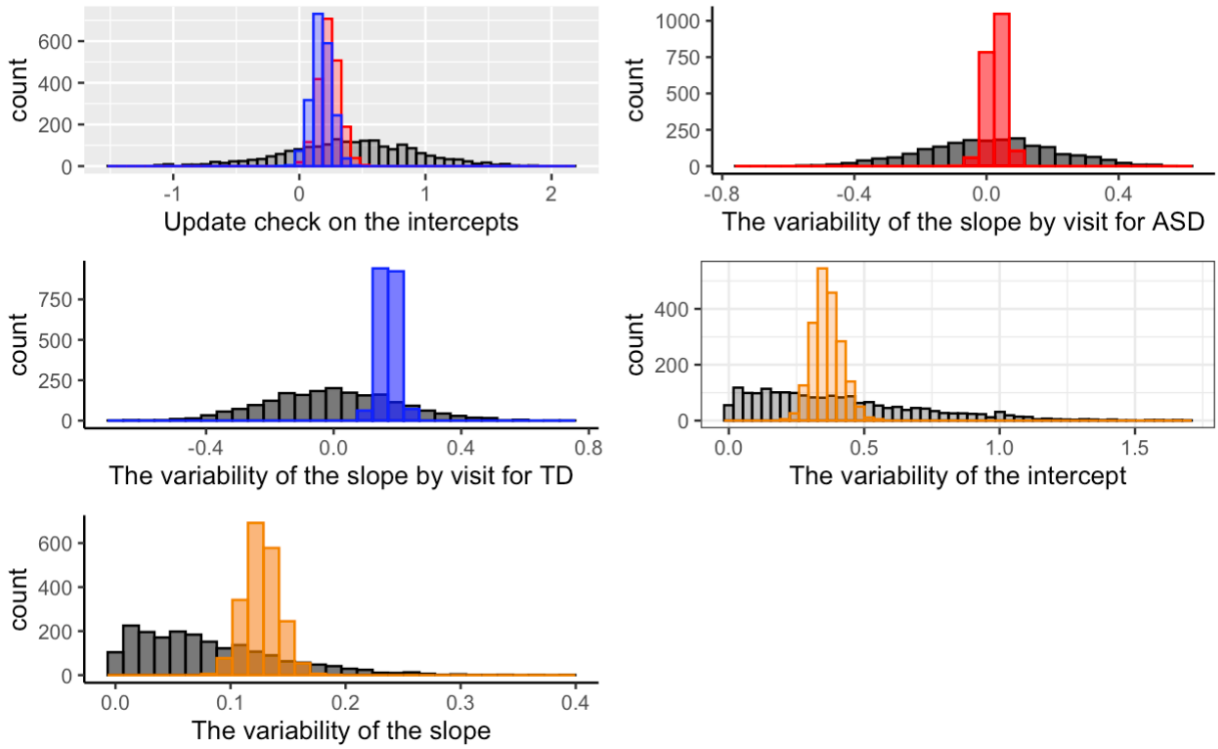
The model, that was applied to simulated data, was also used with given real data. Priors were also set to be almost the same, just with broader intercept standard deviation, so it would not be too constrained, as we are working with more noisy data.



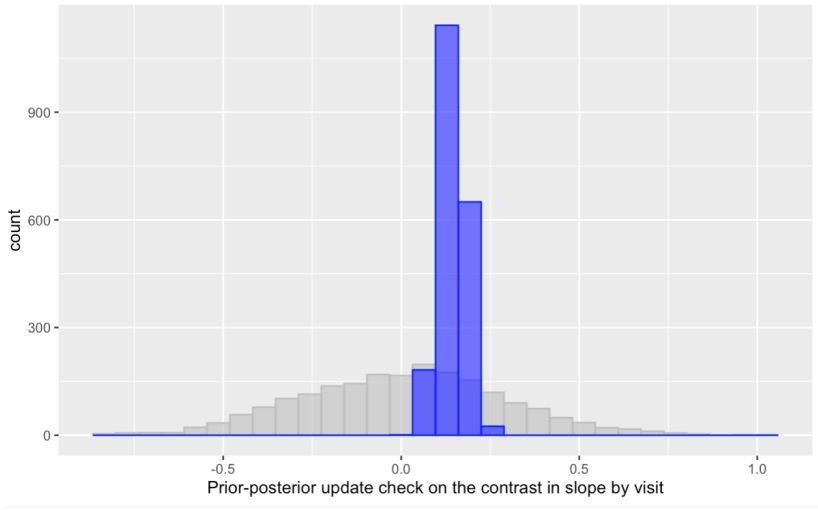
After fitting the model and doing the prior-posterior check, the result is not as clean and overlapping as it was with the simulated data. The model fit to the real data shows that some kids might be overestimated, and some underestimated by the model, as it smooths-out the curve.

Justina Razanauskaite
202109601

Prior – posterior update check



Looking at the prior-posterior update plots it seems that the posterior distributions do not push against the prior, as they are not at the tails of the prior distribution. It also looks like the visualized posterior distributions have become more confident and learned from the model. From this it seems like model is good enough for the data at hand.

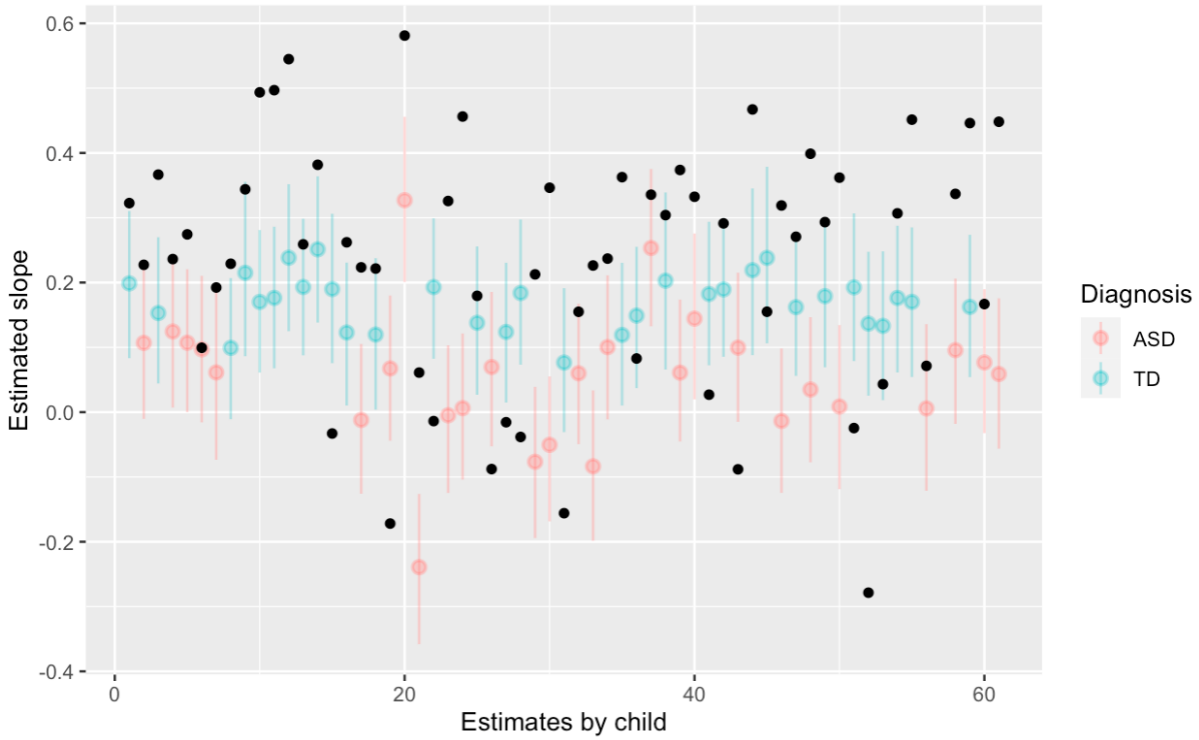


The plot above indicates the difference in slope between TD and ASD. The grey distribution captures the difference in sample from the prior, the blue one – the difference in posterior. Again, the posterior became much more confident in the difference of the slope between the two groups, and is contained well in the boundaries of the prior.

Justina Razanauskaite
202109601

Comparison of the estimates

Check for the estimates of empirical data and the ones estimated by the model can be find below. They were performed on both, intercept and slope. However, the estimated empirical intercept was not captured at all by the model, it was estimated to be way lower than the actual empirical data, therefore, the plot is not included here. The similar effect is seen with regards to the slope as well. Here, that black dots represent the average difference between any two visits and across all six visits for each child. The colored ones are the estimates by the model. It seems like only a few model estimates capture the estimate of the real data. Most of the instances are estimated (by the model) to be lower than they actually are (empirical estimates), and some are estimated to be higher than they actually are.



It seems like the model, which is used in part 1, works pretty well when doing the prior-posterior predictive checks, it learns from the priors, however, it poorly captures the estimates of the empirical data.

Justina Razanauskaite
202109601

Possible additional factors

It might be a good idea to include more factors in the model. I have run another model consisting of one additional factor – MOT_MLU (mother's mean length of utterance). The formula I used can be found below:

CHI_MLU ~ 0 + Diagnosis + Diagnosis:Visit + Diagnosis:MOT_MLU + (1 + Visit|ID),

which indicates that now, the slope depends on the mothers' language as well. In order to compare the models (the one used before, and the one, consisting of MOT_MLU), leave-one-out (LOOIC) model comparison was performed. The results can be found below:

loo_compare( )			**loo_model_weights( )**	
	elpd_diff	se_diff		weight
m_1_add	0.0	0.0	MLU_f_posterior	0.126
MLU_f_posterior	-5.9	3.9	m_1_add	0.874

Based on these results, it seems like the new model (m_1_add) is better than MLU_f_posterior, as both ways of comparing the models show the same outcome.