Emilie, Gergana, Jakob & Natasha

Portfolio 1 text

Part 1

Q1 - 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?

To better understand the model we want to build before applying it to our actual data, we simulate data based on knowledge from the literature. Next, we analyze it and see how well our model performs on the simulated data and test the quality of the model on different sample sizes before moving on to the collected data.

The data simulation process was based on the following parameters from earlier studies:

	Value	**Parameter**
Average MLU for ASD- and TD-children	1.5	Intercept for ASD and TD
Average individual variability in MLU for ASD-children	0.5	Standard deviation for the ASD-intercept
Average individual variability in MLU for TD-children	0.3	Standard deviation for the TD-intercept
Average change in MLU for ASD-children	0.4	Slope for the ASD-kids
Average change in MLU for TD-children	0.6	Slope for the TD-kids
Average individual variability in change of MLU in ASD-children	0.4	Standard deviation for the ASD-slope
Average individual variability in change of MLU in TD-children	0.2	Standard deviation for the TD-slope
Overall error	0.2	Error

The mean changes in MLU are made relative to the (population) mean MLU's for both groups. Respectively, the standard deviations of the changes are recalculated to fit the relative mean changes.

As length of utterances cannot be a negative value, the mean and standard deviations of MLU are log-transformed - i.e. scaled to only taking positive values. The slopes are not log-transformed, as these can take negative values - i.e. it is realistic that a child might produce shorter utterances/speak less compared to the first visit.

The simulated data is then generated by random sampling based on the above information. Based on the plot of the data, our simulation process seems reasonable with a highest MLU value of 4.7.

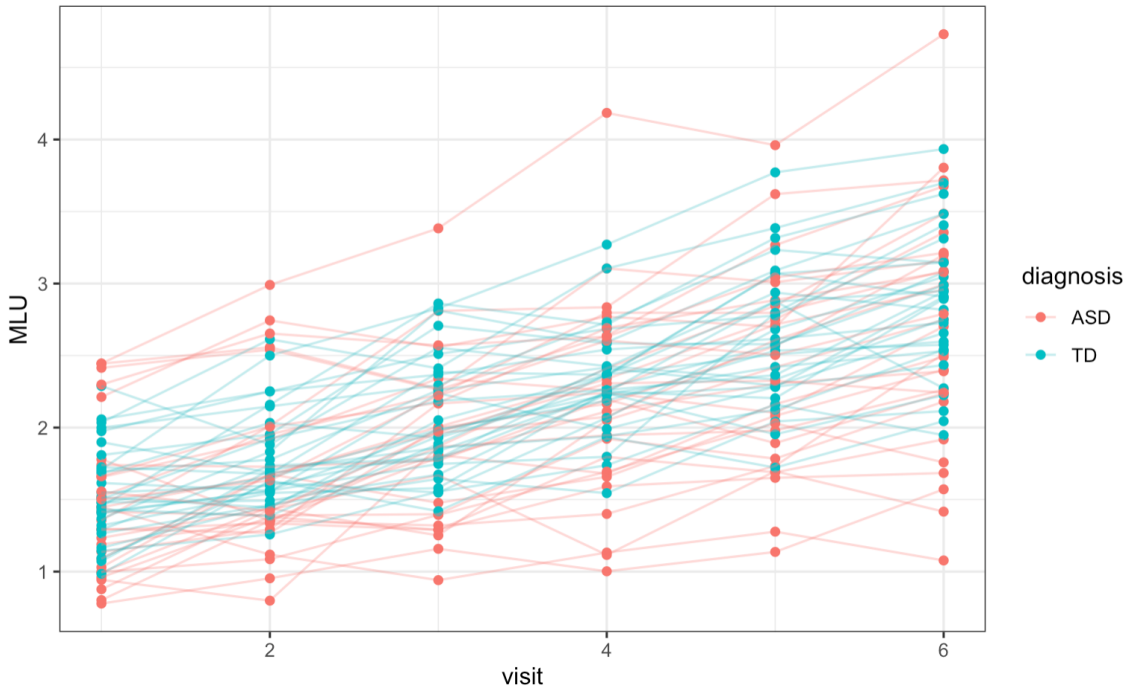

Figure 1.1 - Plot of the simulated data

Now that we have our (simulated) data, we set our formula, define our priors based on our expectation that ASD will be varying more in MLUs than TDs. Then we run a model and see how well our priors fit our (simulated) data.

The formula is set to : *MLU ~ 0 + diagnosis + diagnosis:visit + (1 + visit|ID)*
Since we are interested in the difference between the two types of children, each variable is set to have its own estimate of the intercept. The slope is set to be differentiated by diagnosis (diagnosis:visit) and individual intercepts (1+) and slopes are set by ID (visit|ID).

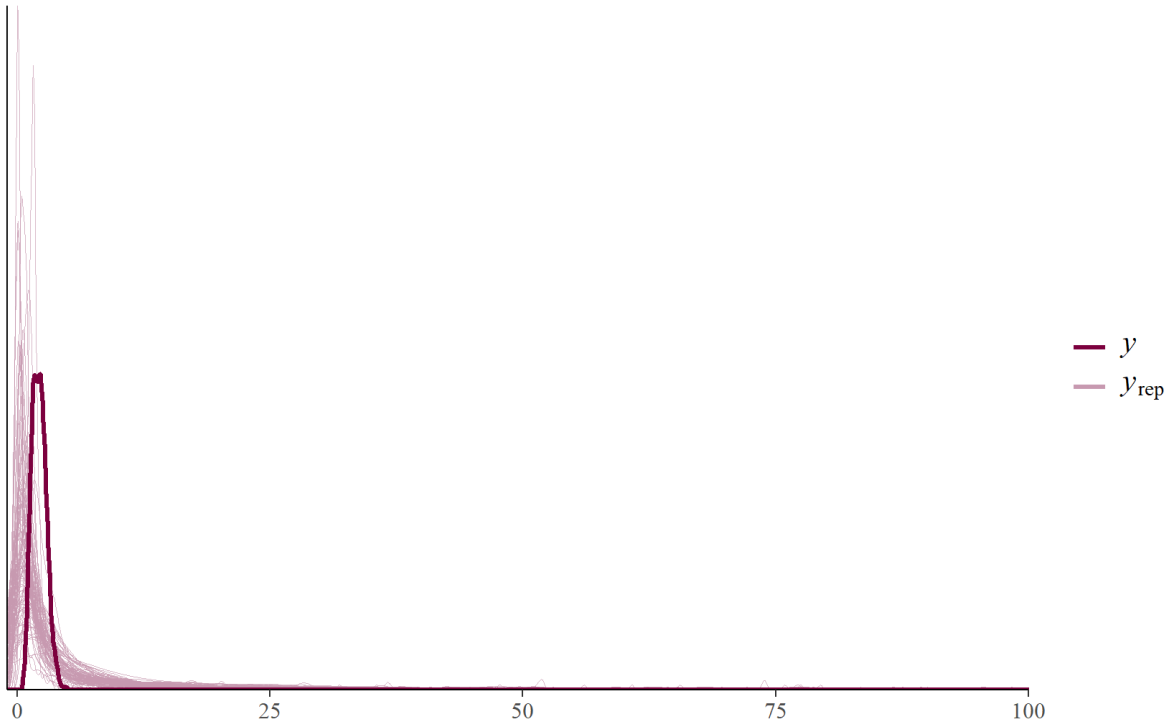

Figure 2.1 - Plot of our prior predictive check

Our prior predictions seem to have a larger and very unrealistic range than that shown by the simulated data. However, we continue with fitting the model of our simulated data and making our posterior predictions. We run a posterior predictive check on our model, which indicates two distributions within our simulated data.

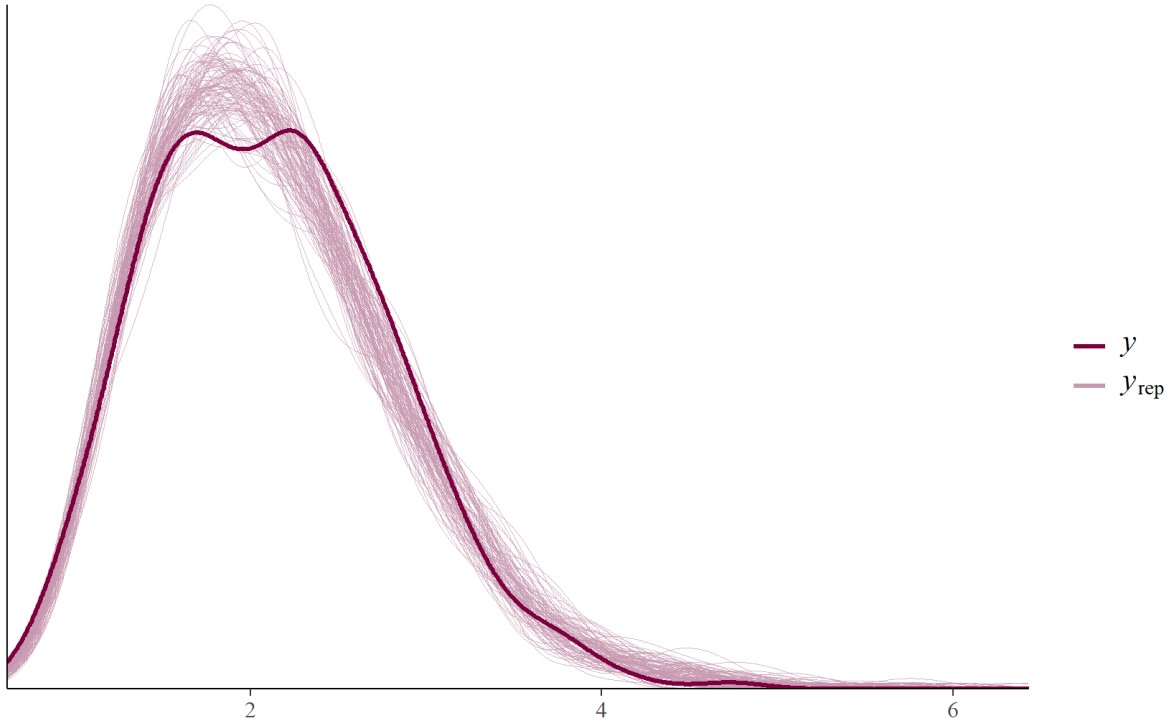

Figure 2.2 - Plot of the prior posterior check

Following this, we firstly visually inspect the prior-posterior update checks. Then, to check the quality of the model we look at the model output, effective sample sizes, rhat.

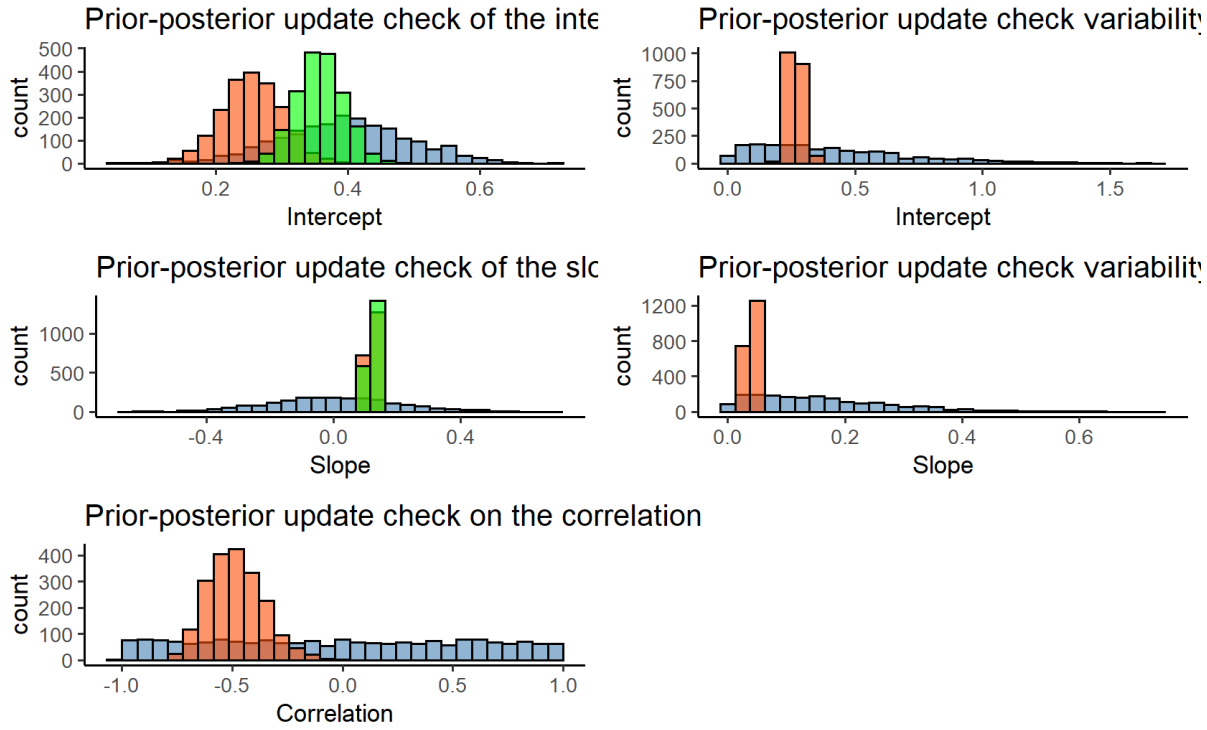

Figure 2.3 - Prior posterior update check plots of parameters. Priors are in blue. Where posteriors are different in the two groups, ASD is in red and TD is in green

The first prior-posterior update check of the intercepts shows that the ASD (red) and somewhat the TD (green) posteriors are most likely being constrained by the set priors, as we can see that they are pushing against it. The posterior distributions are slightly narrower than the prior, which shows that our model has learned somewhat from the data.

In the second plot, individual variability on the intercept, we can see that the posterior (red) is within the limits of the prior and it has gained a lot of confidence, hence the pointy shape.

The prior-posterior update check of the slopes by visit shows that both ASD (red) and TD (green) are within the prior. Furthermore, they show to be much more confident in their estimates.

In the fourth plot, individual variability on the slope, we can see that the posterior (red) is slightly within the limits of the prior. Moreover, it has seemingly gained confidence in its predictions.

Finally, the prior-posterior update check on the correlation shows that our posterior (red) is well contained within the set prior. Furthermore, it is negative and confident.

CI-intervals:
The intercept of MLU for the ASD children ranges from 0.16 to 0.34, and for the TD children it ranges from 0.29 to 0.43, indicating that our model did not capture our underlying value of 0.41 for the ASD children and narrowly for the TD children.

Estimated visit effects are 0.12 for the ASD population (CI: 0.10-0.14) and 0.12 for the TD-children (CI:0.11-0.14). Compared to the literature values of 0.27 for mean change of visit for ASD and 0.4 for mean change of the TD children, the model did not excel in capturing the underlying values.

The rhat is not larger than 1, which suggests that our chains have mixed well. All Bulk_ESS and Tail_ESS values are above 200 (should be at least 100 per Markov chain to be reliable), which suggests that the sampling efficiency in the bulk and tail of the distribution are both alright.

To sum up, from this it seems that this model suggests that both groups have different starting points, but develop at the same rate. A hypothesis analysis shows that we cannot be very certain about the difference between the two groups, as the estimate is 0.

SAMPLE SIZE CONSIDERATION
NB! Our power analysis will not run currently, both because 'R' keeps breaking and because we are trying to optimise/finish the code.

Conceptually, the power analysis shows how many simulations include a difference in slopes above 0 between the two groups at a given sample size.Thus the power analysis helps us estimate a reasonable sample size that includes the underlying effect.

Part 2

Q2 - 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.

Our sample contains 66 children in total (mean age in years ≈ 3, 55 males) divided into two groups (typically developing (TD) and autistic spectrum disorder (ASD)) depending on their clinical and cognitive features. 34 of the 66 children are assessed as TD, while the remaining 32 children belong to the ASD group. This distribution seems well balanced, and our sample

data resembles our simulated data (which was based on the literature) in regards to MLU values (see Figure 3.1 and Figure 3.2) However, we can see that they differ in development by visit.

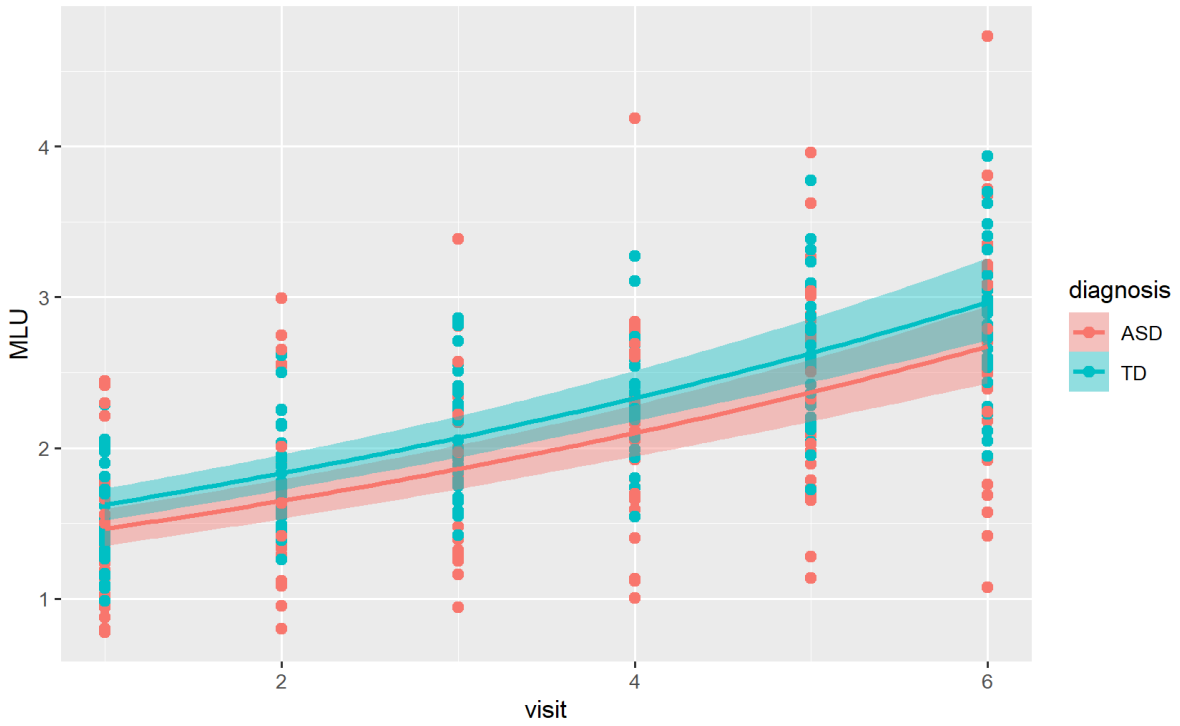

Figure 3.1 - Simulated data

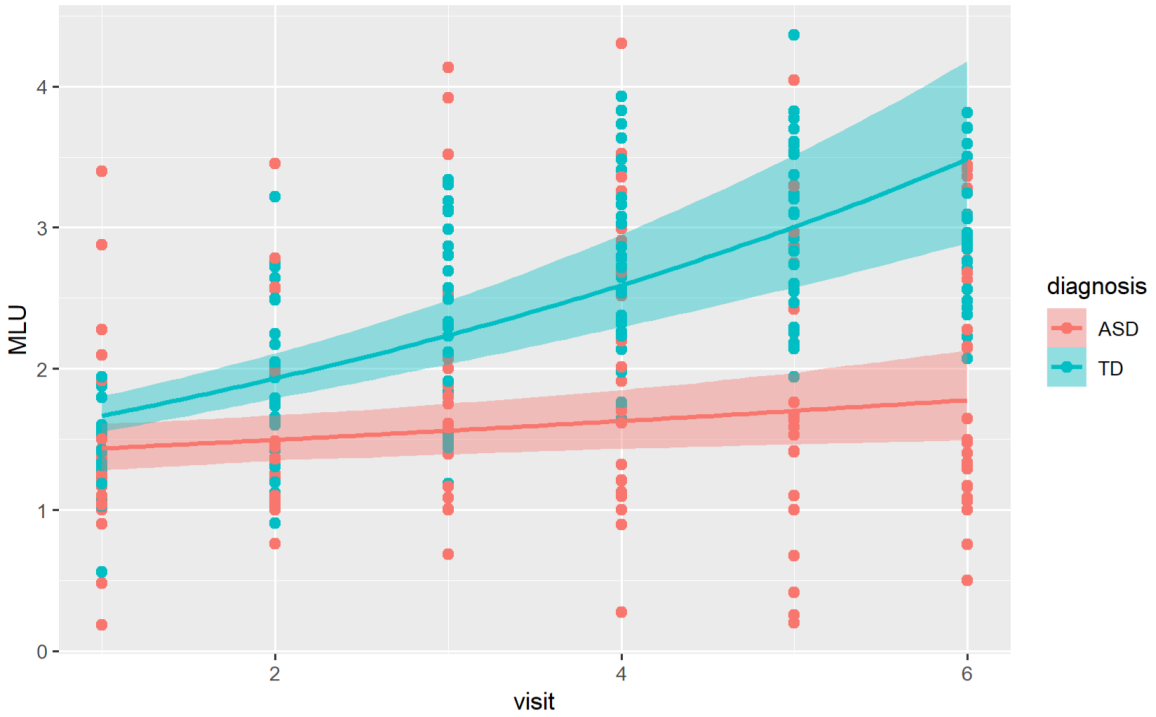

Figure 3.2 - Real data

To fit our coming models, 5 children were removed because of missing values. Moreover, 1 child was removed because they showed as nonverbal in the data. Therefore, when investigating language it does not make a lot of sense to include this child in the analysis.

Our baseline model is defined from the formula: *MLU ~ 0 + diagnosis + diagnosis:visit + (1 + visit|ID)*. From the model output, we can see that the ASD kids develop slower linguistically than the TD kids. The population-level slopes of their development are 0.04 and 0.15 respectively. Both groups have an estimated error of 0.02. The difference between the two groups is 0.1.

From the baseline model's output we can see that not all children's linguistic development is reflected by the general trend for their diagnosis group.The standard deviation of the slope for the individual child (across both groups) is 0.09. This means that the linguistic development for each child can vary with 0.09, meaning that within 1 standard deviation for ASD children this could result in negative linguistic development.

From the baseline model we see that diagnosis alone cannot explain the childrens' (individual?) linguistic development. Therefore, we fit other models with other predictors. Firstly, we fit a model to include/capture more of the childrens' linguistic environment: *MLU ~ 0 + diagnosis + diagnosis:visit + diagnosis:MOT_MLU + (1 + visit|ID)*. Then we fit a model to include/capture more of their cognitive abilities: *MLU ~ 0 + diagnosis + diagnosis:visit + diagnosis:verbalIQ1 + diagnosis:nonVerbalIQ1 + (1 + visit|ID)*.

Conceptually, we know that linguistic development is influenced by what all three models try to include/capture. Both the linguistic environment and the child's cognitive abilities play in towards their linguistic development, along with their diagnosis. As such, the optimal model would include all these as predictors.