Methods 3: Final Exam Portfolio

Ioana-Luisa Forcaș
Cognitive Science BSc, University of Aarhus
202106168@post.au.dk

Ioana Luisa Forcas
au693527

Assignment 1

Part I

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?

Simulation

One of the most important parts of statistics is running simulations to appreciate how our data behaves regarding the analysis we want to conduct.

I will therefore describe what the process of simulation taught me and how it helped me:

- It made me consider the particularities of the design: how many participants would make the best number for a proper analysis, how many conditions and observations of participants/ condition, and what is the best way to capture all of these

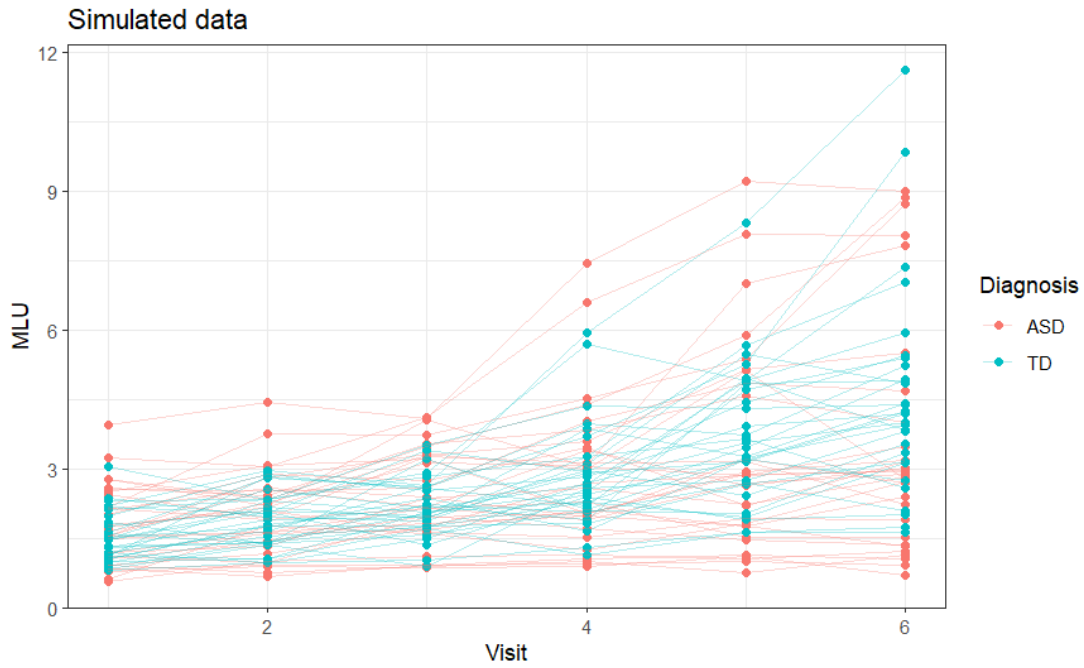
- It made me take a closer look to the distributions sampled, which helped me understand the real data

- It improved my intuition on the best tools to use, the outcomes and values the process should generate in the real data analysis

- Writing the analysis code on the simulated data made my job easier when I moved to the real data, and because repeating is the best practice so learn, I got a better picture of the goal of the analysis and how it should be performed.

Ioana Luisa Forcas
au693527

Plot of the simulated data



The goal of the plot was to show that the MLU (Mean Length of Utterance) value we simulated seems to capture the event we are interested in. This phenomenon shown in it provides plausible trajectories of the MLU development for both conditions (ASD and TD) consistent with the visit number.

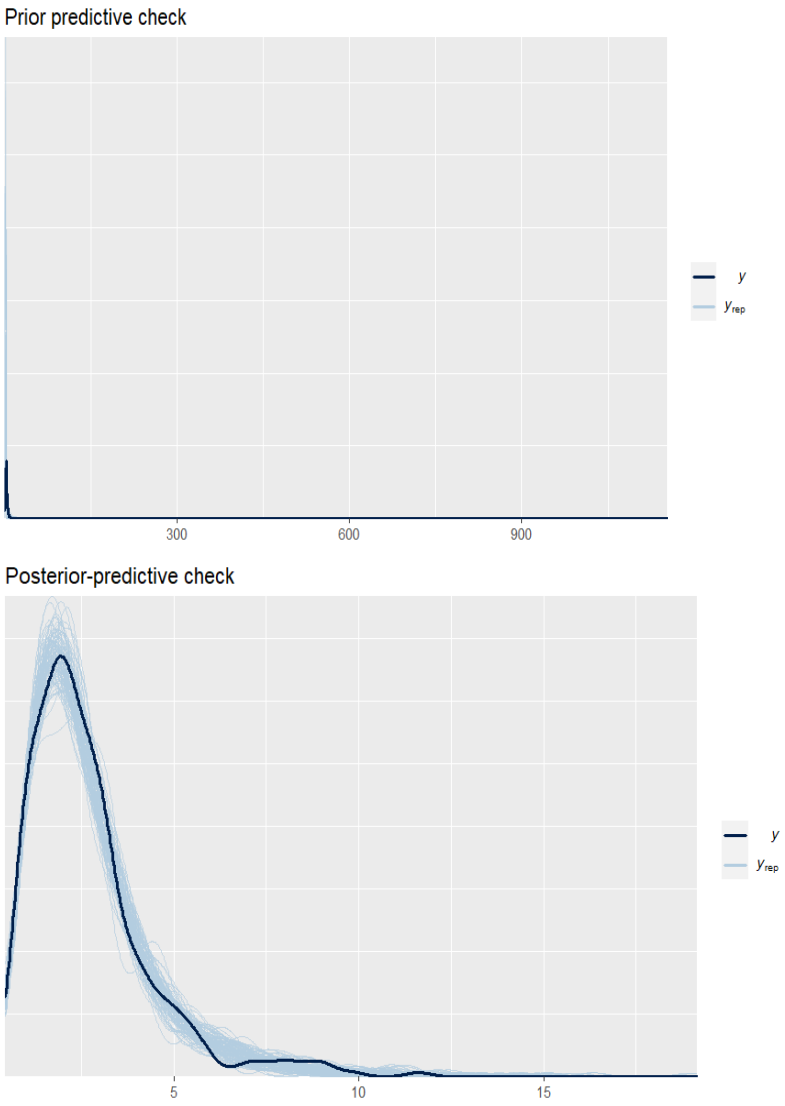
We can deduce the differences between the populations and how individual-level characteristics influence the anticipated evolution of both groups from the graphic.

Since we are interested in the development of MLU, I defined the formula in the following way:

MLU ~ 0 + Diagnosis + Diagnosis: Visit + (1 + Visit | ID)

This formula suggested that each diagnosis category (ASD and TD) should have a unique estimate for the intercept with the same degree of uncertainty, that both groups would have a different slope at each visit depending on the diagnosis, and that all the parameters should be distinctive for each child, who would also have a unique intercept and slope.

Ioana Luisa Forcas
au693527

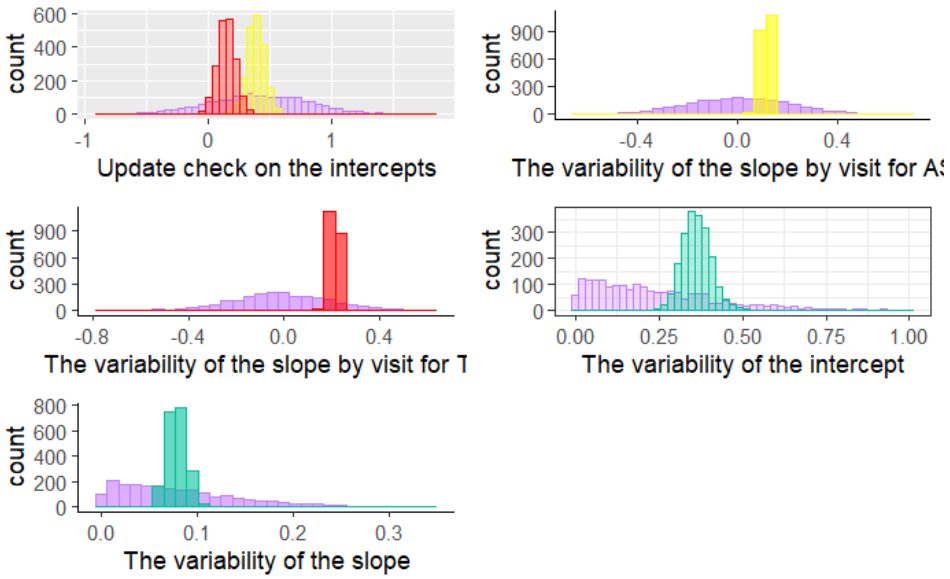


As all individual variants were incorporated in this figure, a prior prediction check reveals that MLU values of 200 and above are potential result values. The model performed on the simulated data shows that the range of potential values is reduced, and a satisfactory match is demonstrated.

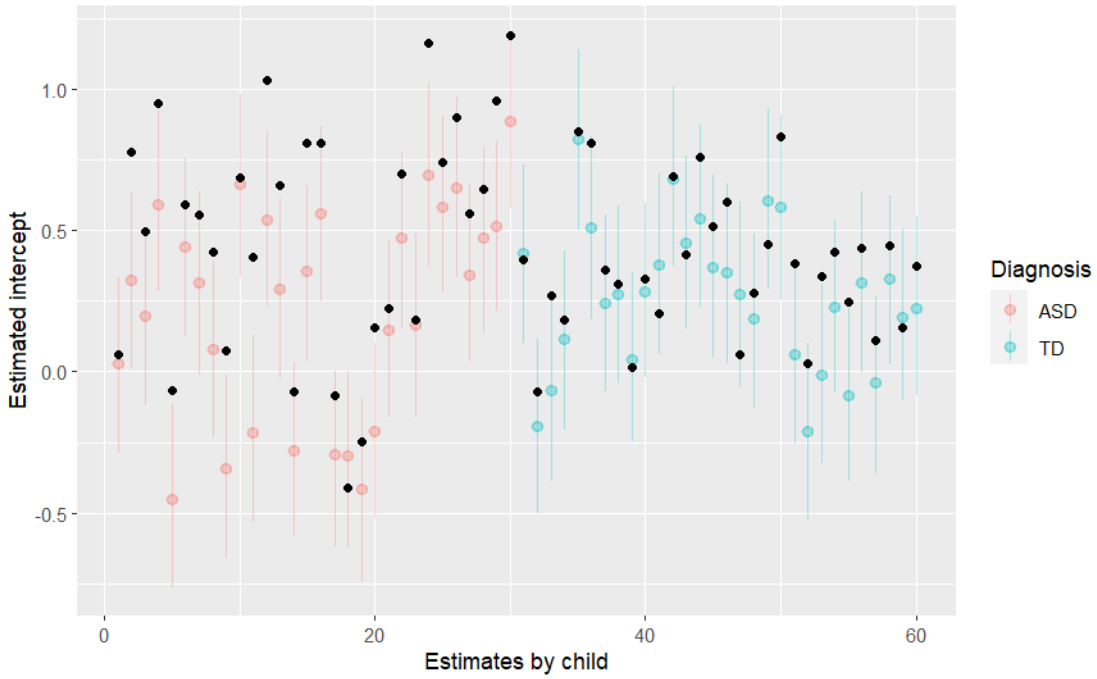
Posterior-predictive checks help us see if the priors we set on the model fit. We can see the discrepancies of the prior and posterior predictive checks in the following plots:

- Prior is shown in the purple color
- The distribution of ASD and TD is shown in the yellow and red colors (Yellow for ASD and Red for TD)
- Both groups are shown in the blue color

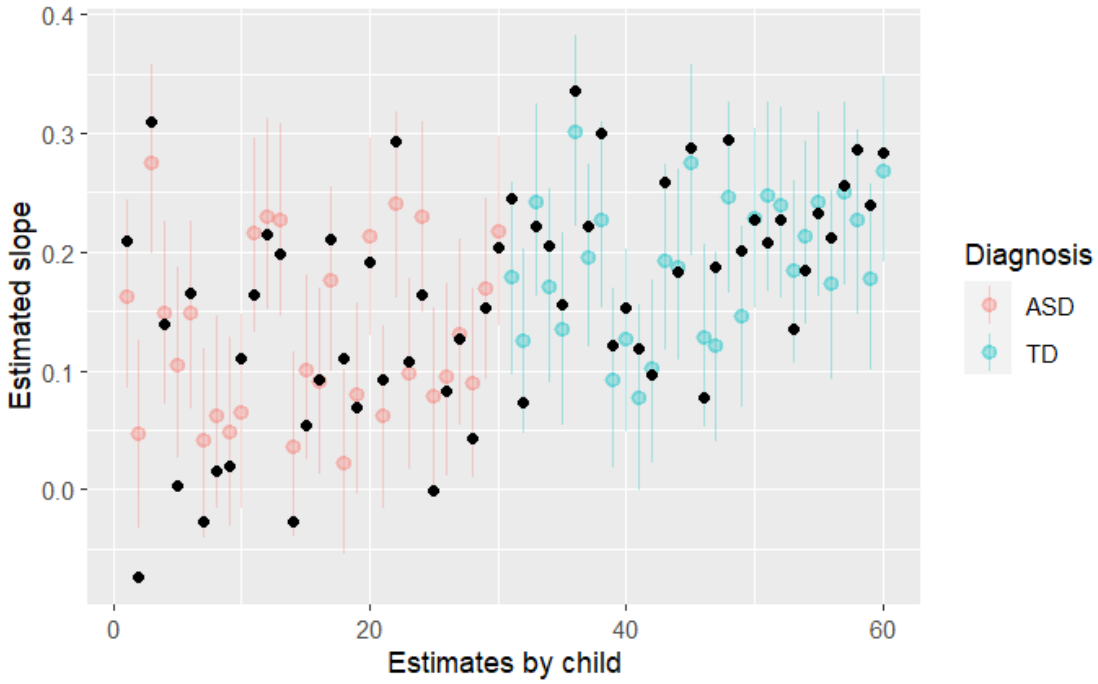
Ioana Luisa Forcas
au693527



By looking at the distributions we can conclude that the prior distribution "incorporates" the posteriors, the model has learned from the data, showing that the prior-setting goal was reached.



Ioana Luisa Forcas
au693527

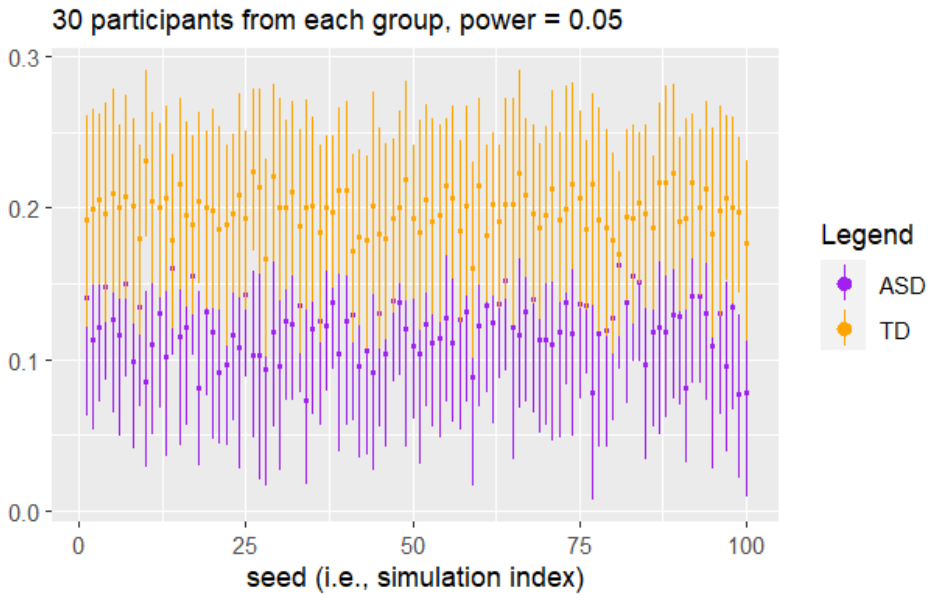


In the first plot we see that Intercept is wider for ASD, rather than TD and in the second one, the slope shows faster evolution for the TD children, which is very logical.

Power analysis

Since the real data is fairly equal to the simulated data, then this power analysis and subsequent visualization of it demonstrated that a sample size of 30 participants from each one of the two groups is reasonable. The power analysis for 0.05 is quite poor. The plot shows an effect size of 0.1 for ASD and 0.2 for TD.

Ioana Luisa Forcas
au693527



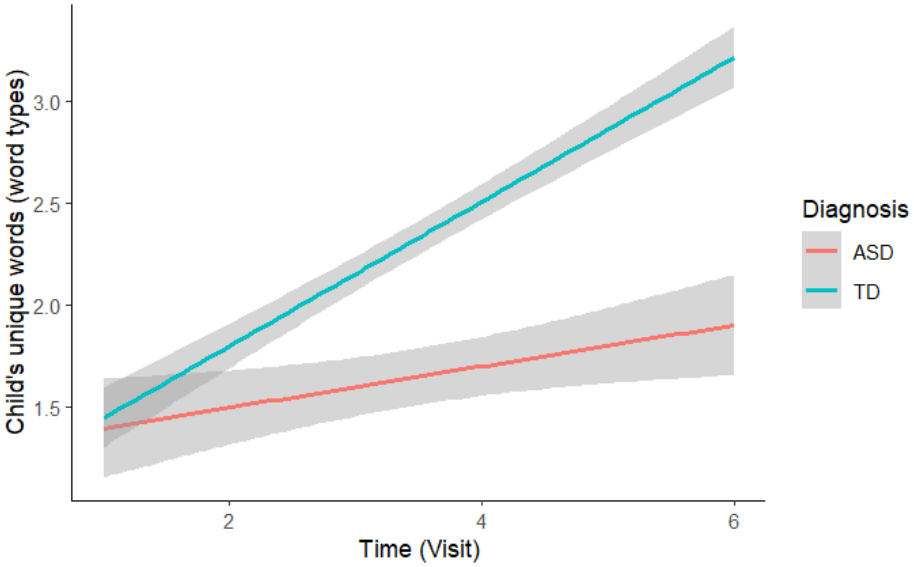
Part II

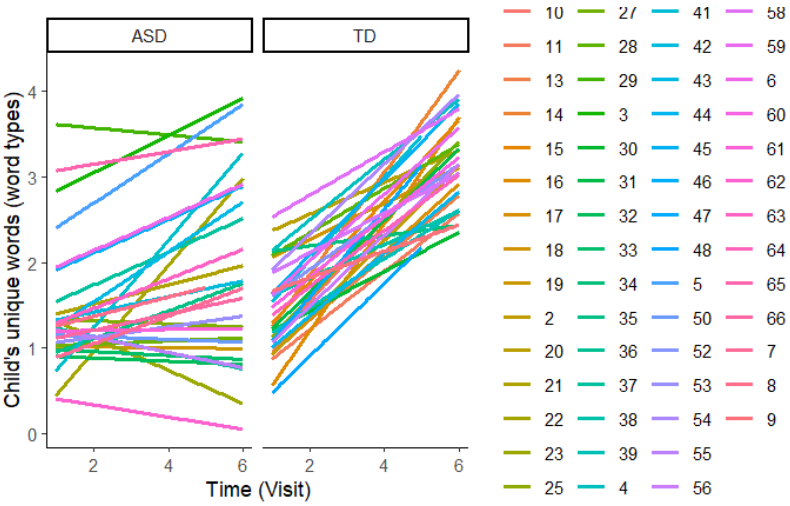
Q2: Briefly describe the empirical data and how they compare to what you learned from the simulation (what can you learn from them?). Briefly describe your model(s) and model quality. Report the findings: how does development differ between autistic and neurotypical children (N.B. remember to report both population and individual level findings)? Which additional factors should be included in the model? Add at least one plot showcasing your findings.

Description of the Empirical Data

The 66 individuals in the sample that made up the empirical data were split into 55 female and 11 male participants; however, not all participants were present for each visit, so the participant count changes regarding to the number of the visit, with 55 participants in the first visit and only 50 in the last visit. Participants' ages ranged from 26.4 months on average at visit 1 to 47 months on average at visit 6, which served as the last visit. The study's participants underwent socialization, verbal, and nonverbal IQ tests. One of the problems with the data is that the groups are not balanced, the female ratio being significantly higher than men, so this may conclude into non-universal results.

Ioana Luisa Forcas
au693527



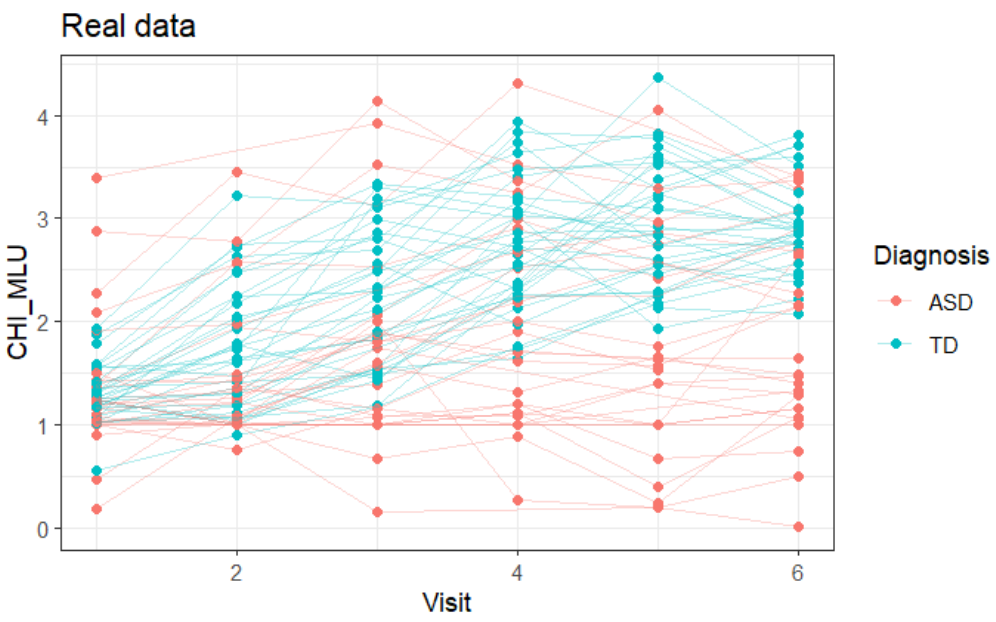


The plots above show that TD children had a better performance when tested (which was also to be expected).

Even if the number of participants in the simulated data was almost the same as in the empirical data, a lot of uncertainty in the empirical data was caused primarily by the heterogeneity in the number of visits per child. This is one of the significant differences between the two data sets. This could be solved in the future by considering a higher number of participants in the real experiment.

Plot of the Empirical Data

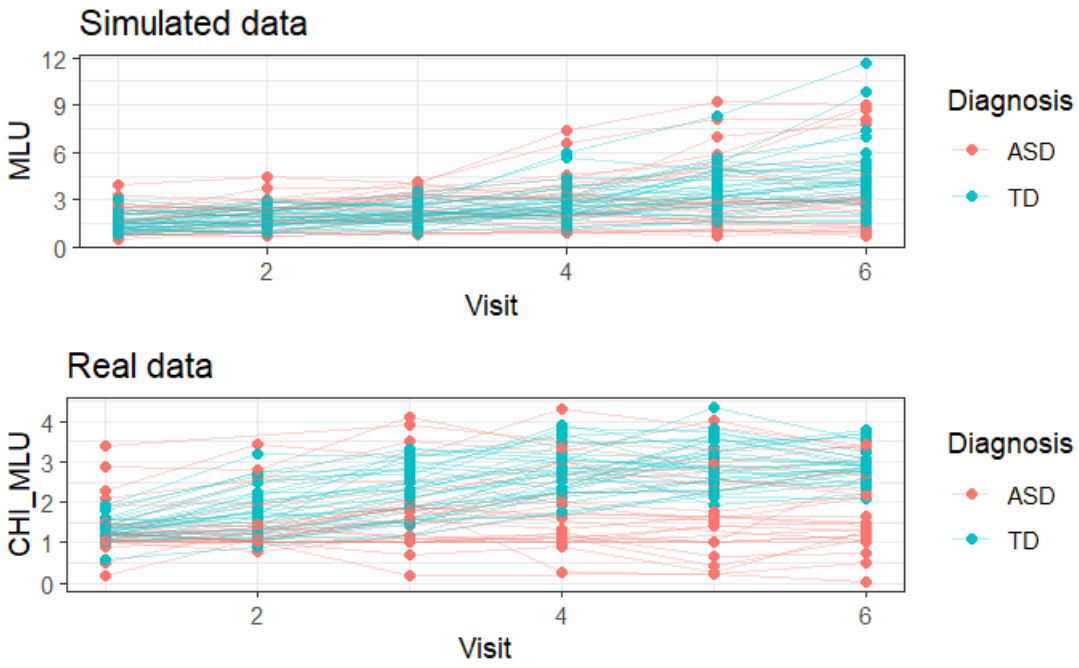
Ioana Luisa Forcas
au693527



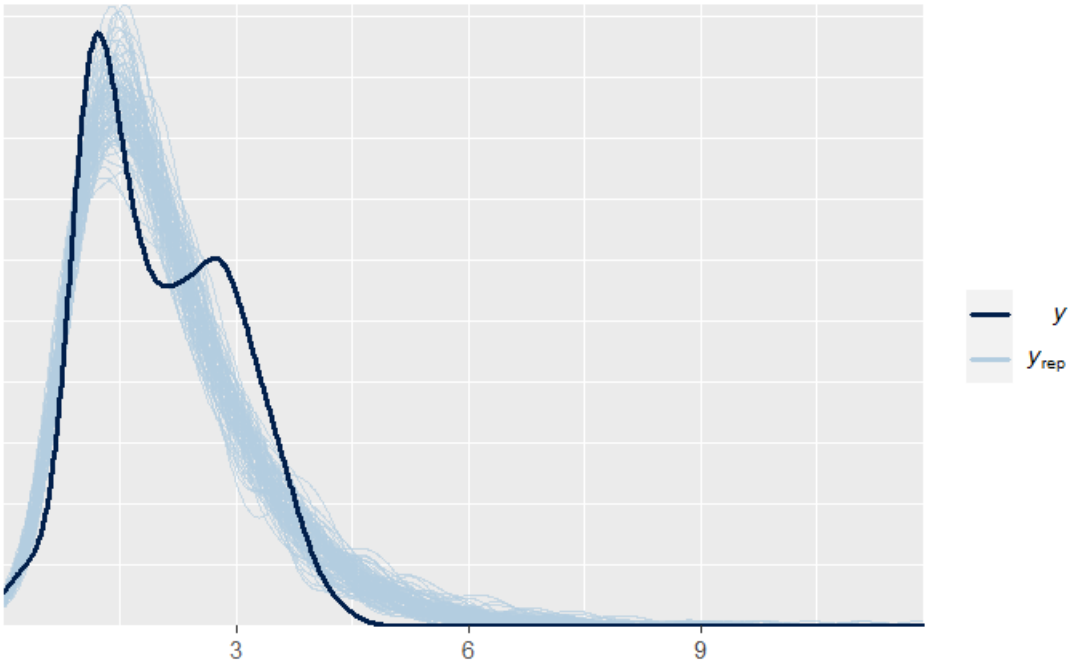
Plotting the simulated data shows that both groups appear to steadily increase in terms of the value of MLU. Additionally, while not as much as could be anticipated, the ASD group exhibits increased variability when compared to the TD group. It is evident at visit 4 (and subsequent visits) that the TD group is mostly above the ASD group, which makes sense given that TD should grow more rapidly than ASD.

Although real data shows the TD group's evolution over time, it does not indicate levels as high as in the simulation. The starting position at visit 1 is roughly the same for both the simulated and real data populations. Although the latter one does not provide as much evidence in the development as the former does, it is still evident that the ASD group has significantly more variability than the TD group. These two graphs show that in fact there is far more variety and less dramatic development in both groups, demonstrating the influence of our parameter choices that attempt to generate real-world data.

Ioana Luisa Forcas
au693527



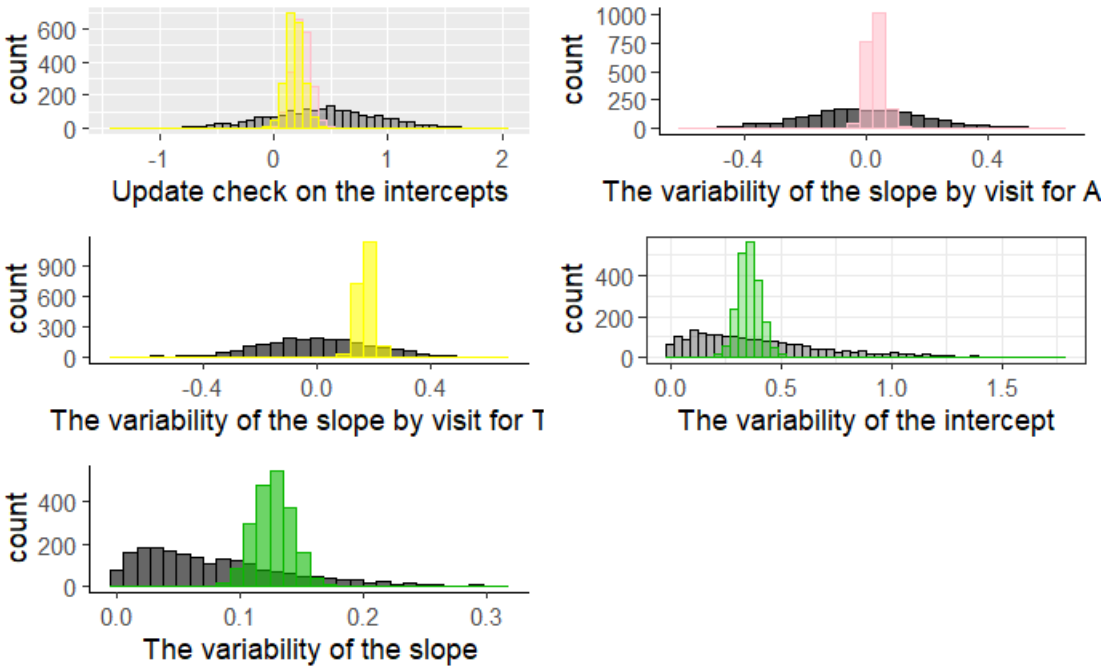
The model I trained on the simulated data I applied on the empirical data. To avoid being overly limited given that we are working with noisier data, priors were also designed to be roughly identical but with a larger intercept standard deviation.



As the curve is smoothed, the model's fit to the real data reveals that certain children may be overstated, and others underestimated.

Ioana Luisa Forcas
au693527

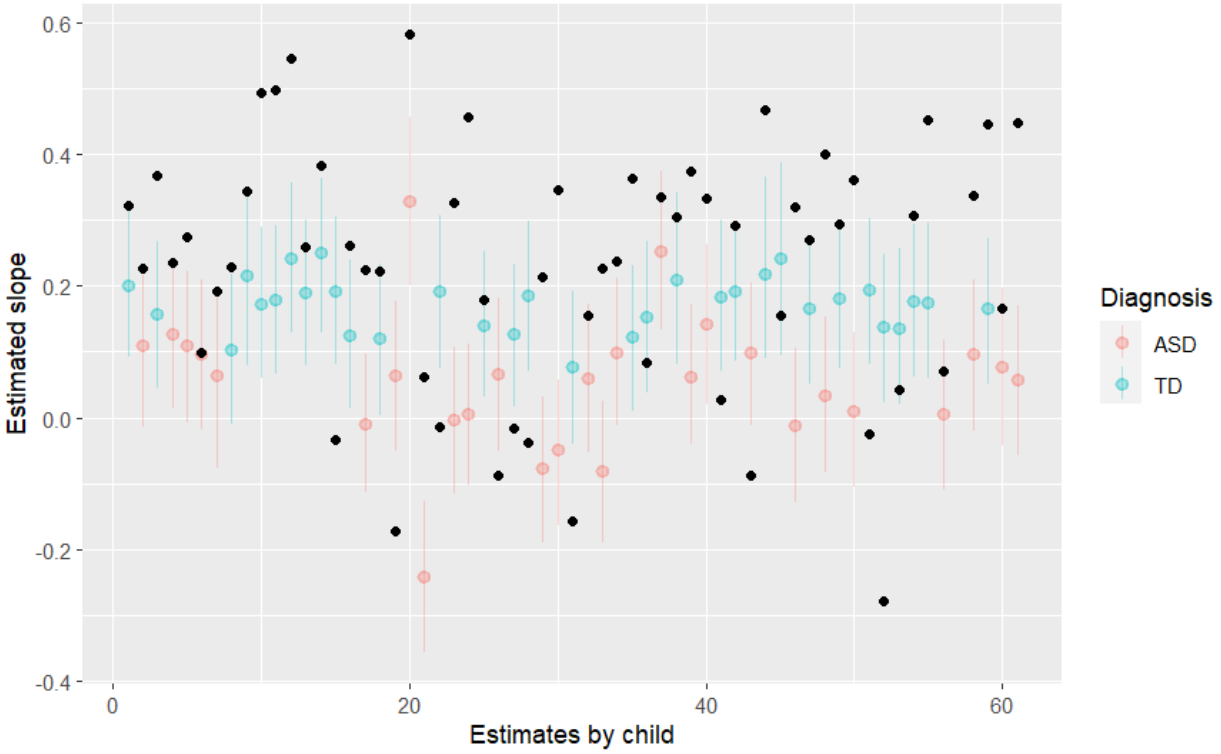
Prior-posterior check



The posterior distributions do not appear to push against the prior, because they are not near the ends of the prior distribution. The posterior distributions that can be seen also seem to have grown more certain and to have learnt from the model. This suggests that the model is adequate for the available data.

Ioana Luisa Forcas
au693527

Estimates



The model employed in part 1 appears to perform admirably when doing prior-posterior predictive tests since it learns from the priors, but it doesn't capture successfully the estimates of the empirical data.

Additional Factors

After doing a bit of research, concluded that one of the most important factors implied in child's development speech is mother's mean length of utterance (MOT_MLU).

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

Also, another possibility would be to have the verbal IQ:

Ioana Luisa Forcas
au693527

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