Portfolio
Assignment 1

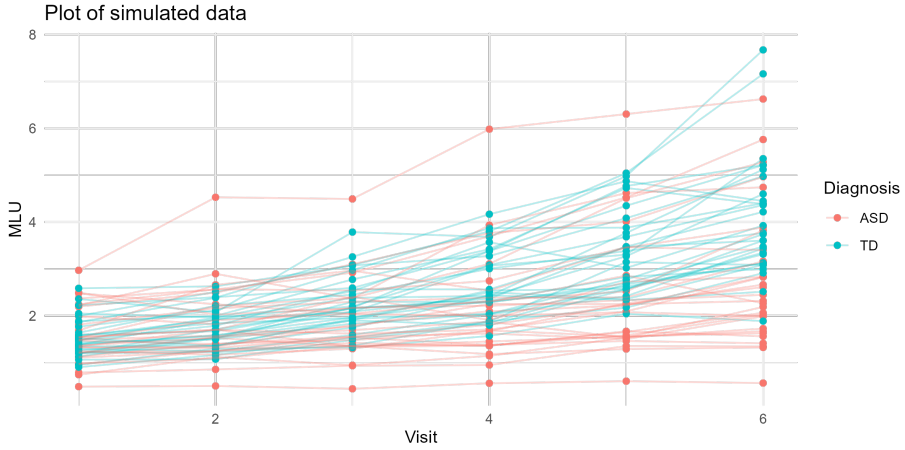
Sabrina Zaki Hansen (au693815)

Assignment Description

GitHub Repository: https://github.com/sabszh/Assignment1

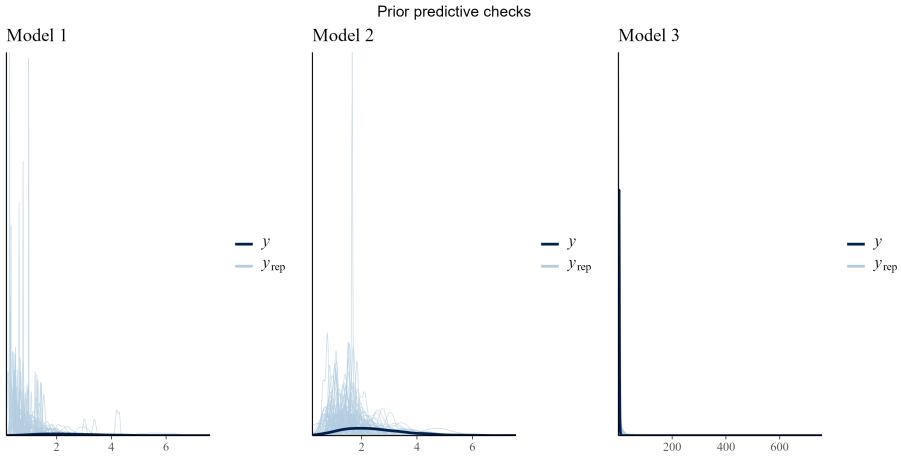
Q1

The goal for the simulation process for this assignment has been to structure a data-set that will be used to empirically asses whether Autism Spectrum Disorder is related to language impairment. The data was simulated with the given values from relevant literature on the average of mean length of utterances of two groups; neurotypical children and children with autism, throughout 6 visits in a clinical trial. In the following process and modelling, the data was log-normaled. This transformation ensured that we would only be working with positive values.

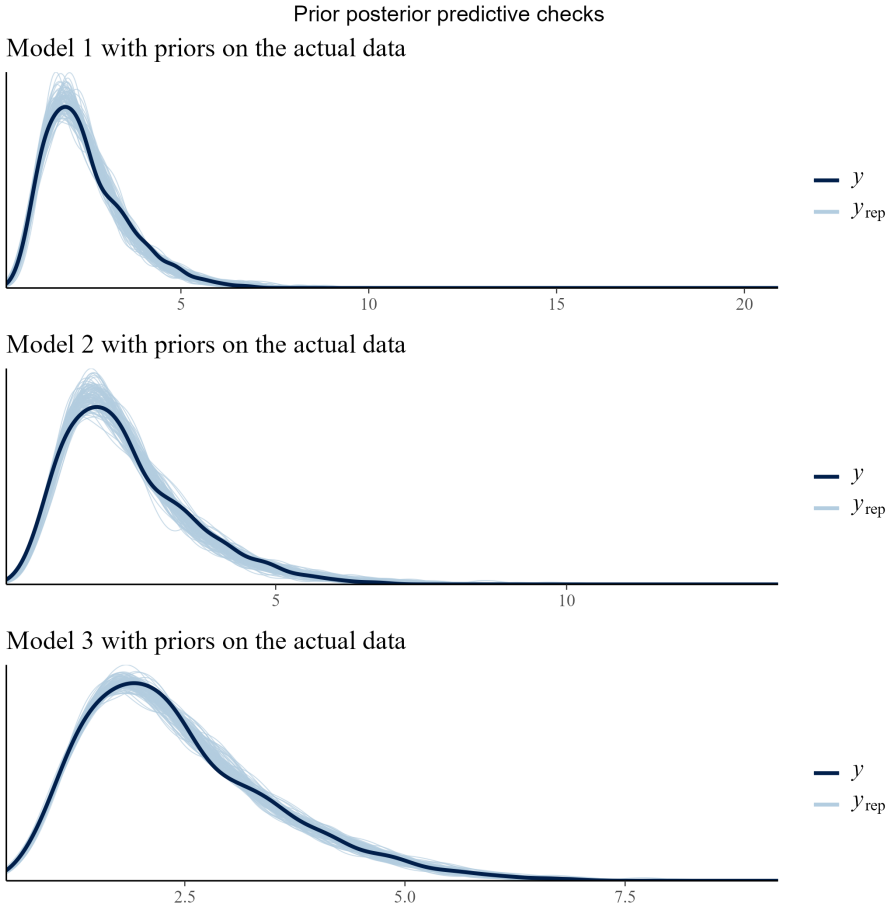


For the simulation three formulas were constructed based on plotting of the data, and then compared, the prediction is that a model fitting best would be that MLU for a child is predicted by their diagnosis, the interaction effect between their given diagnosis and visit, and the number of visits they have had, allowing for random slopes and intercept for Visit and their ID. However, this

had to be tested making sure having a complex model would not result in overfitting (thus comparing simpler models onto more complex ones).



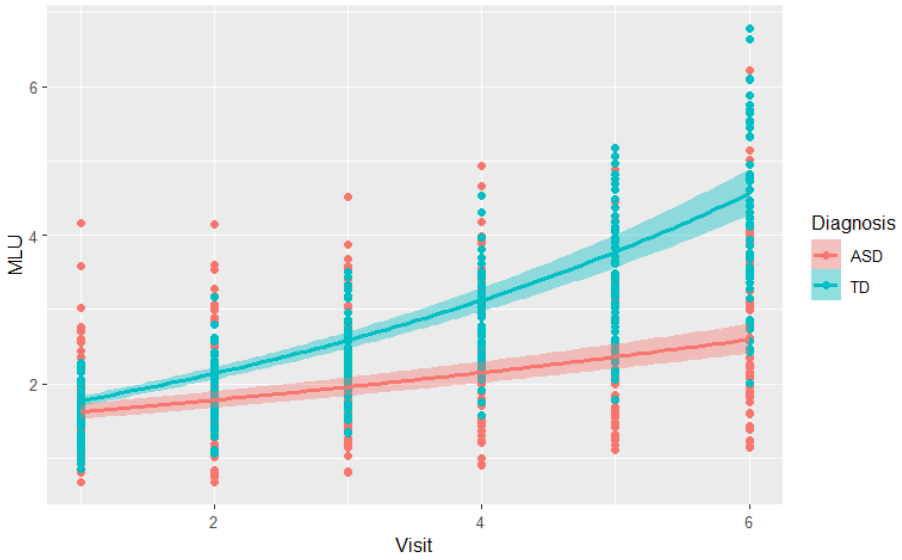
The plots above illustrate the priors predictive check, and is used to assess if our statistical models makes sense and is a good fit for our simulated data, and whether the model is able to make predictions about the data. Given these prior predictive checks we accept my weakly informed priors (as my μ is set as 0) reasonable and continue to use them.



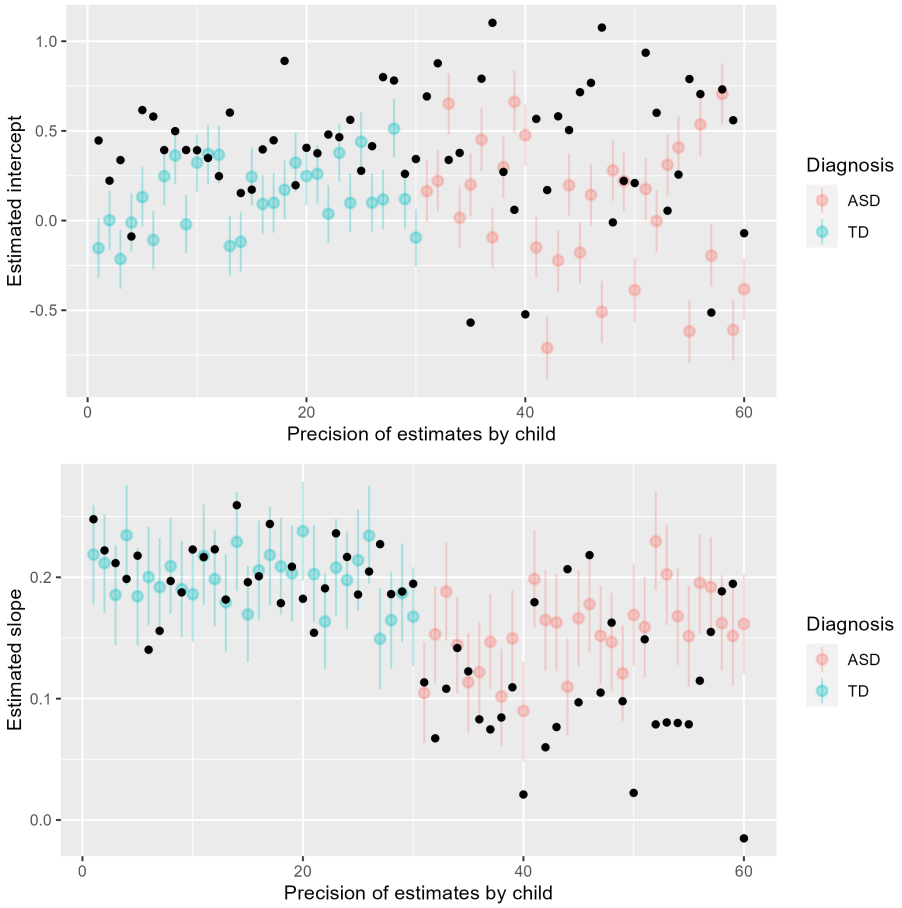
Accepting the priors, a Bayesian fit was performed to the simulated data using a Markov Chain Monte Carlo (MCMC) method. The posterior distribution of each parameter was thus obtained by sampling the likelihood of the data given each parameter to then update the prior distribution in light of the simulated data. Through the Prior posterior predictive checks we inspected whether the posterior distribution of our models are reasonable, given our priors. We checked if the posterior distribution is close to the prior, and if they peak at the right values. This seems to be the case with all three models, however the third model seems to be updating a bit better, as y and y_{rep} align better.

Inspecting the Prior Posterior Update Plots for our different models and parameters, it is visible that the prior does not have to work so hard to update. Even with a weak prior, the posterior update is very effective.

We also did trace plots to verify that the chains were mixing properly, this seemed to be the case as there was nothing popping out when investigating the convergence of the chains.



By doing a conditional effects inspection on our multilevel model, we discovered that the effect of visit on MLU is different depended diagnosis-this was however as we would expect given our set parameters in our models. However, we were interested in investigating whether our models would replicate this expected effect, and as it seems to be doing so we feel confident in working further into our modelling.



In the plots above, we performed a parameter recovery. The goal of it, was to extract and figure out in which ways the model is assessing the individual level estimates. Thus asking the crucial question: how wrong is the model when it tries to reconstruct the specific intercept and slope of all the samples? The black dots are the data-points from the children.

On the top-plot, we can see that the model is alright at predicting the individual level estimates for the intercept, with a small standard deviation, (the average SD of the estimated intercept was 0.58). In the last plot, we can see that the model is better at predicting the individual level estimates for the slopes, with a reasonably small standard deviation (average SD of 0.4). However, the model seems to generally underpredict the intercepts and overpredict the slopes. Overall, we can conclude that the model is quite good at estimating the individual level parameter.

By investigating $\hat{R}$, $Bulk_{ess}$ and $Tail_{ess}$, we inspect how well the different models fit the data. Overall the third model including explanations of the interaction and each level performs best.

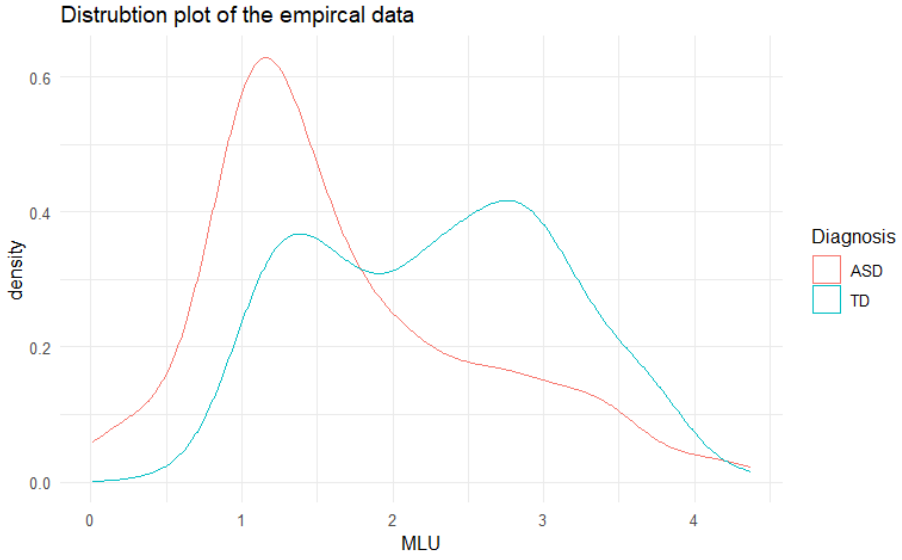
For this first draft of the assignment, a power and precision analysis has not been sufficiently performed as there have been issues with R. Therefore, I briefly just want to comment on what we expect from this analysis and how we would perform it.

The workflow would be to generate simulations with different sample sizes, to determine the sample size required for our study. The expected finding would be that a larger sample size than 30 would probably give a better estimated power, but we would then investigate for realistic standards how small we are able to make the sample size while still having a sufficient power. The same will be done for the precision analysis.

Q2

The empirical data that will be used for this part of the assignment, have the following descriptors:

Summary of descriptors of empirical data					
Gender	Diagnosis	n	Average age	mean of MLU	SD of MLU
F	ASD	5	33.07	0.88	0.60
F	TD	6	20.25	1.29	0.31
M	ASD	26	33.03	1.37	0.69
M	TD	29	20.41	1.31	0.27

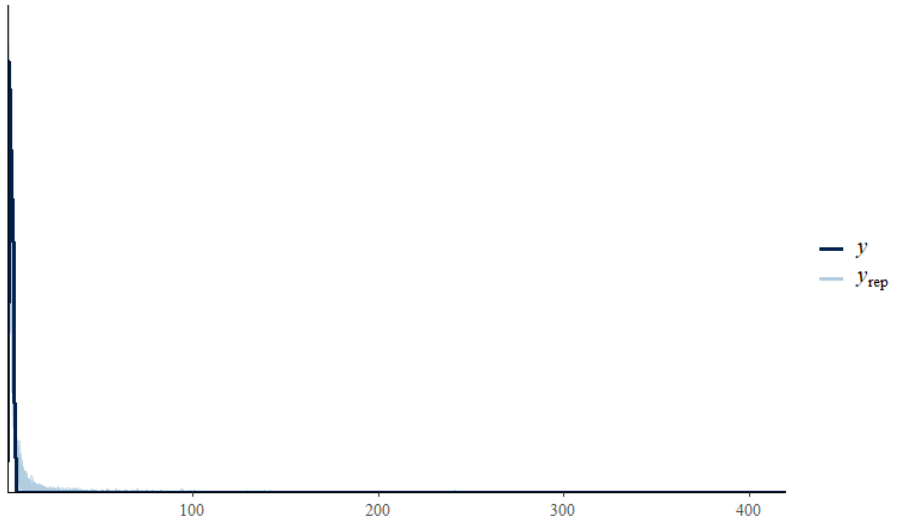


This plot shows how the empirical data is distributed. From looking at the plot there seems to be a clear bimodal distribution for the TD, which might be interesting to investigate. Inspecting the descriptors there does not seem to be an obvious explanation to this. However, our mean and SD estimates are rather

similar to those set for our simulated data.

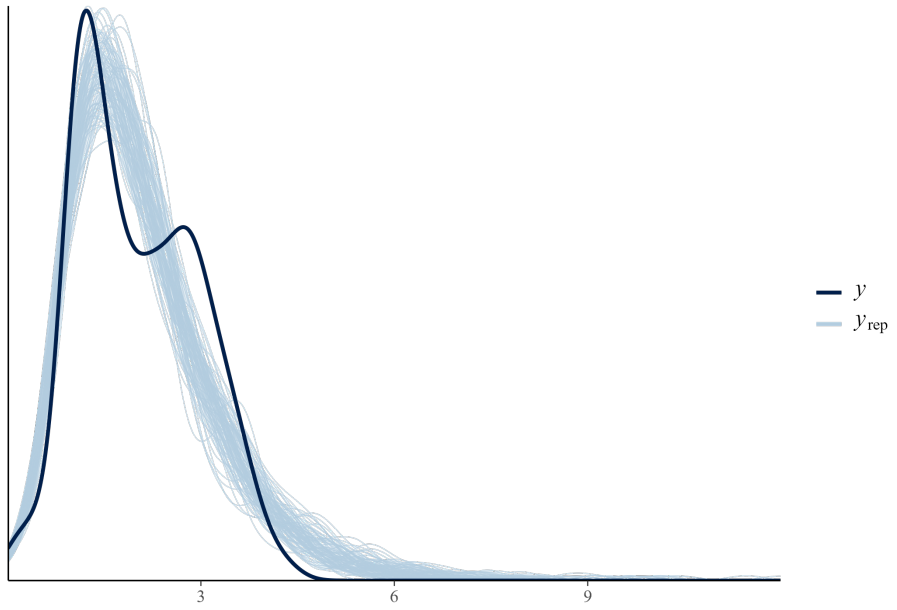
As the first part of this assignment, among other, concluded that a model with multilevel is doing the best job at explaining and predicting the data, it and its prior will therefor be assigned to the empirical data.

Model 3



The above prior predictive check suggest that there might be some issues with our modelled prior as they are long-tailed, however as the priors are weakly-informed we accept it, as it allover seems to allign with y.

Model 3 with priors on the actual data



Looking at the Prior Posterior Check we notice that the posterior distribution for μ is bimodal. This means that the data may be coming from two different sources. In this case, there are two possible values of μ that are equally likely according to the posterior distribution.

The most likely value of μ is around 1.5. However, there is also a second peak at around 3.0. This means that the data could be coming from two different underlying distributions, with one source having a mean of 1.5 and the other having a mean of 3.0. Investigation of the Prior Posterior Update plots tells us that the distributions for all the parameters are centered around the true value of the parameters.

Hypothesis testing was then performed, firstly to asses the population level findings, on how the development differ between neurotypical children and of those with autism. In this case, the evidence ratio was 3.76 and the posterior probability is 0.79. This means that the evidence favors the first hypothesis of the distribution of neurotypicals to be larger than those with ASD by a factor of 3.76, and that the first hypothesis is 79 % likely to be true after taking into account the evidence. For the individual level, an evidence ratio of 599 and posterior probability of 1 tells us that the probability of the hypothesis being true is extremely high.

To expand upon our model, additional factors such as socialization, as this might affect their performance in the study. This has however not been investigated further in this first draft.

Q3

For this part, it is anticipated that a cross validation will be executed to test the predicting power of the model(s). We will also critically look at how the model might be overfitting to the data, and describe improvements to our chosen models.