Portfolio 1

Frederieke Wullf – 202105369

Q1:

Simulating data in advance of analyzing the real data, can have many benefits. Generally, simulations help to better understand the collected data, evaluating ideal sample sizes, which conditions and how many observations to include. Furthermore, our goal was also to create a model that can sufficiently analyze the real-world data. Simulations have many more benefits, such as helping to prove a hypothesis or understanding a connection between different variables.

To simulate the data, we firstly had to define multiple values for the variables in the dataset. The dependant variable, mean length of utterance (MLU), was our main target in that the goal is to predict the MLU, based on, if the child has an autism spectrum disorder (ASD) diagnosis or if it is typically developing (TD). The MLU was recorded over the course of 6 visits.

To set up the simulation, we firstly needed to define the data type, which we set to be a gaussian distribution, which we set up to be on a log-scale, as the MLU cannot contain any negative values. For the simulation, we started to define the mean (μ) and the standard deviation (σ) of the MLU for the two groups of children. Furthermore, the increase/decrease of MLU based on the visits, needed to be defined. Finally, we set the error. We defined the values as follows:

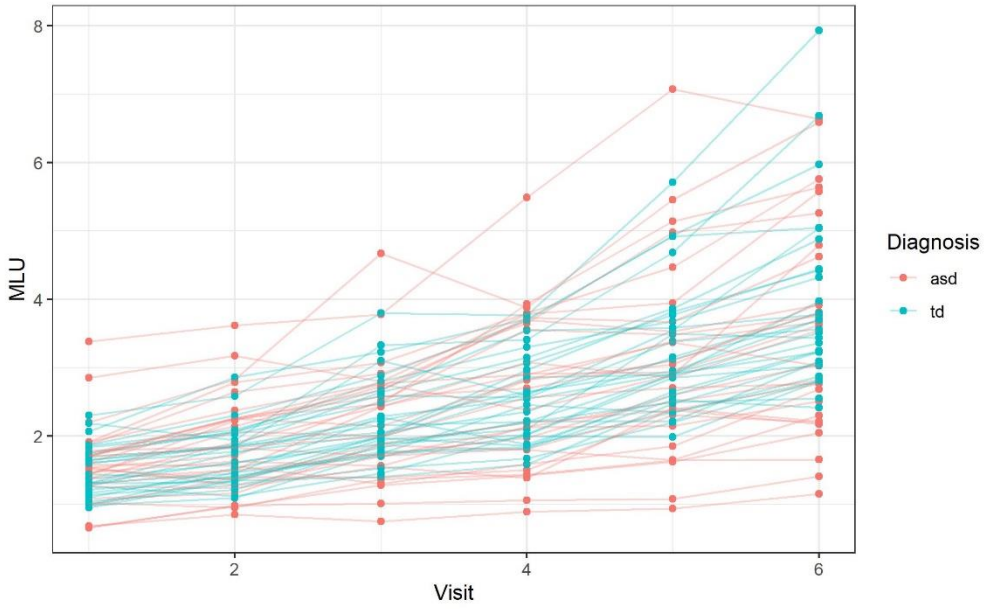
```
mu_asd <- log(1.5)

sigma_asd <- log(1.5) - log(1.5-0.5)

mu_td <- log(1.5)

sigma_td <- log(1.5) - log(1.5-0.3)


mu_visit_asd <- 0.14

sigma_visit_asd <- 0.06

mu_visit_td <- 0.19

sigma_visit_td <- 0.02

error <- 0.1
```

The values of the parameters that included visit, were chosen by looking at the distribution and spread of simulated data in a histogram.

The sample size was set to be 30 per group to begin with, as the real dataset contains around 30 children per group. The analysis of the best sample size was conducted later and subjected to change to figure out, which sample size would give the most precise output. Another way to change the precision of the model estimates would be to change the predictors in the model, to account for possible conditional effects etc.

The simulation was initiated by creating an empty tibble and then including a function that would loop through the parameters the function rnorm(), which creates normal random variables in a defined space. Individual intercepts and individual slopes were created for each ID datapoint, which takes the variability in between the children into account.

The output of the simulation was captivated in a plot, showcasing the development in MLU for each child over the course of the 6 visits. The plot shows that the autistic children generally have a bigger population sd, which explains why the individual lines, one representing each child, being spread out more. The typically developing children seem to lie closer to each other, i.e., they have a smaller sd, while simultaneously having a higher change in MLU for each visit, compared to the autistic group.

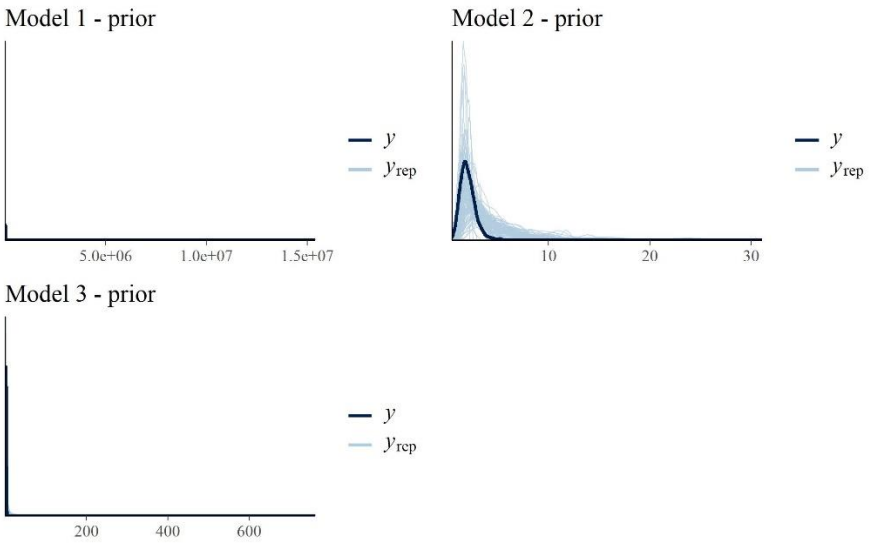


After the simulation, the formula of the model was defined. Three models of different complexities were considered:

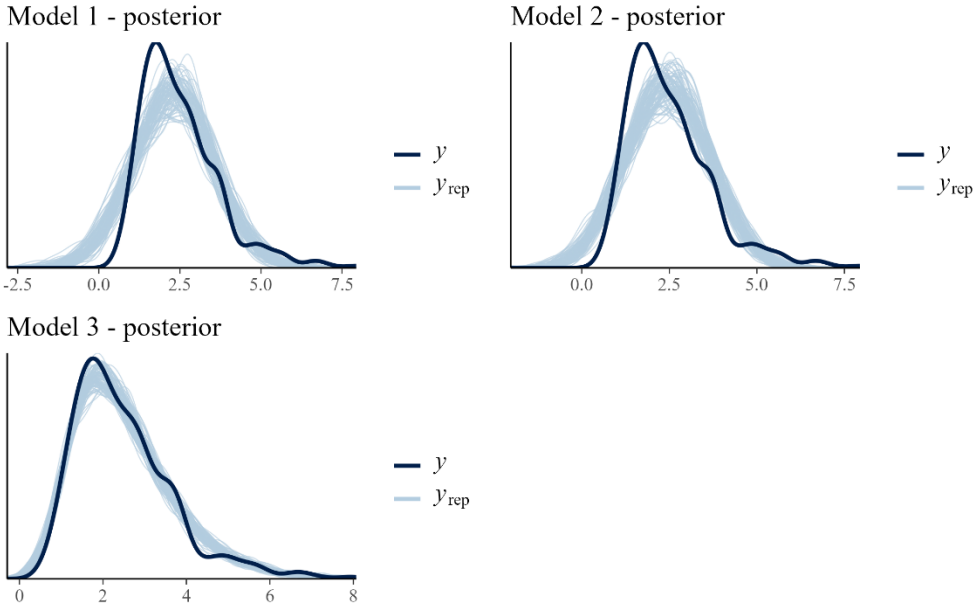
Model 1 → (MLU ~ 0 + Diagnosis)	MLU only predicted by Diagnosis
Model 2 → (MLU ~ 0 + Diagnosis + Diagnosis:Visit)	MLU predicted by Diagnosis, where the change in MLU by visit, is given by diagnosis
Model 3 → (MLU ~ 0 + Diagnosis + Diagnosis:Visit + (1 + Visit| ID))	MLU predicted by Diagnosis, where the change in MLU by visit, is given by diagnosis. Random slopes for and intercepts for each child are added as well.

For each model, priors were being estimated and set. In this setting, opting for uninformed (not very informed) priors, would be preferred as we want the data to determine the results of the analysis. By choosing priors that only favor a specific interval, some values are going to be selectively preferred, which often leads to not keeping inferences in a reasonable range.

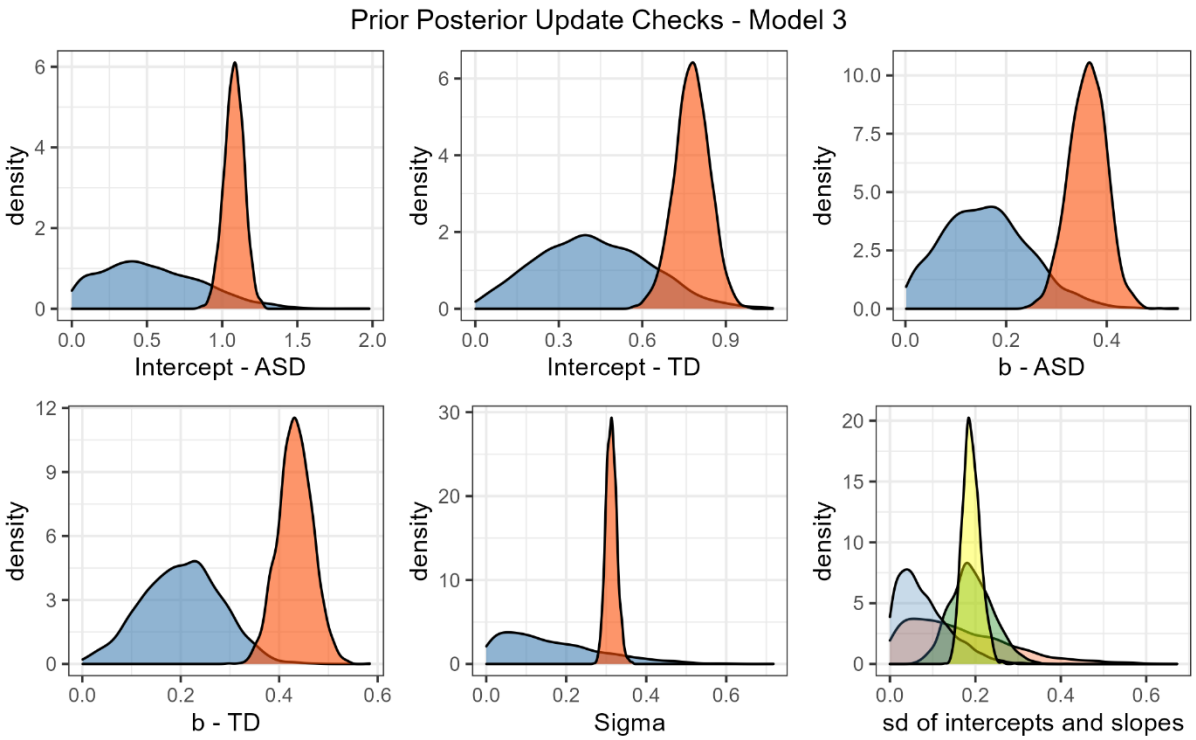
The priors have then been used to perform prior-predictive checks, in order to investigate whether the distribution is reasonable. Graph 2 shows the prior predictive checks for all 3 models.



Subsequently, the three different models were run with the pre-defined priors. To investigate, whether the model has captured the data, posterior-predictive checks were run. Those revealed that the model had learned from the data, as the simulations followed the overall distribution of the real data. Especially model 3 shows that the data follows the distribution of the real data, whereas the data for model 1 and model 2 are slightly off.

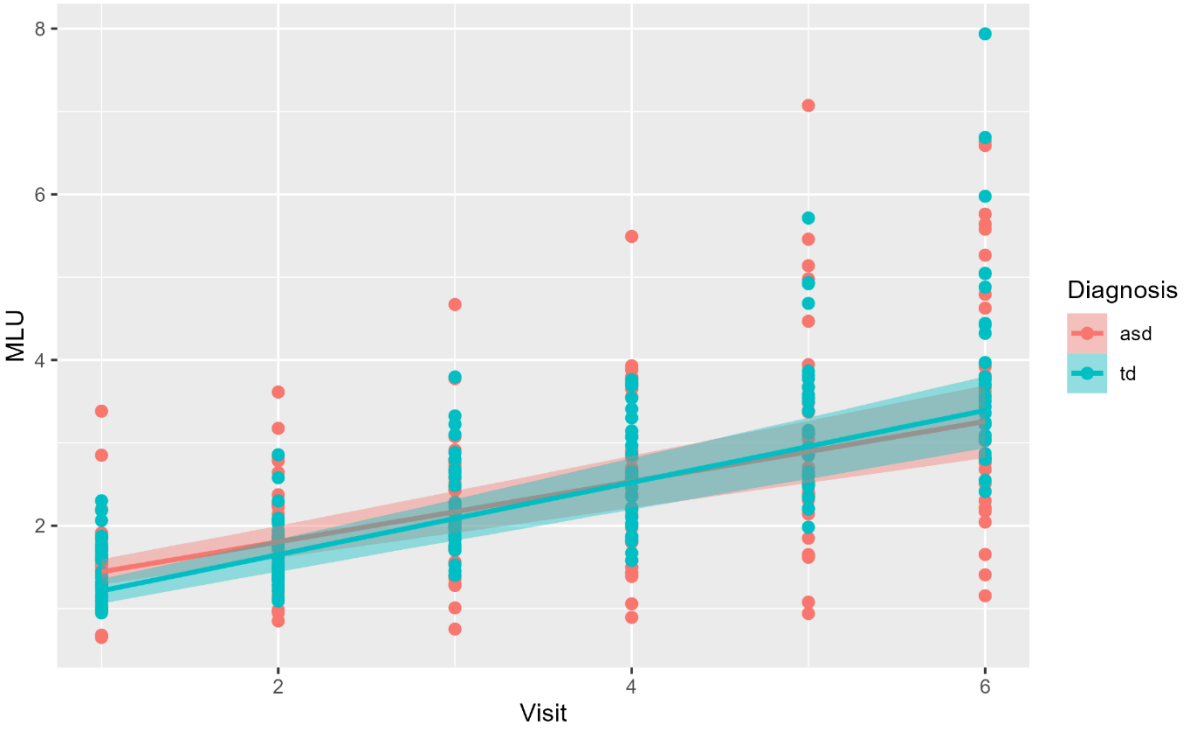


The “learning effect” that the models have shown have been visualized with prior-posterior update checks. The prior-posterior update check for model 3, depicts clearly that the model has learned from the data:



It can be observed that some of the posterior, such as the slope "b" are bordering on the prior, which could be a sign to redefine the priors, so that they fit the data better.

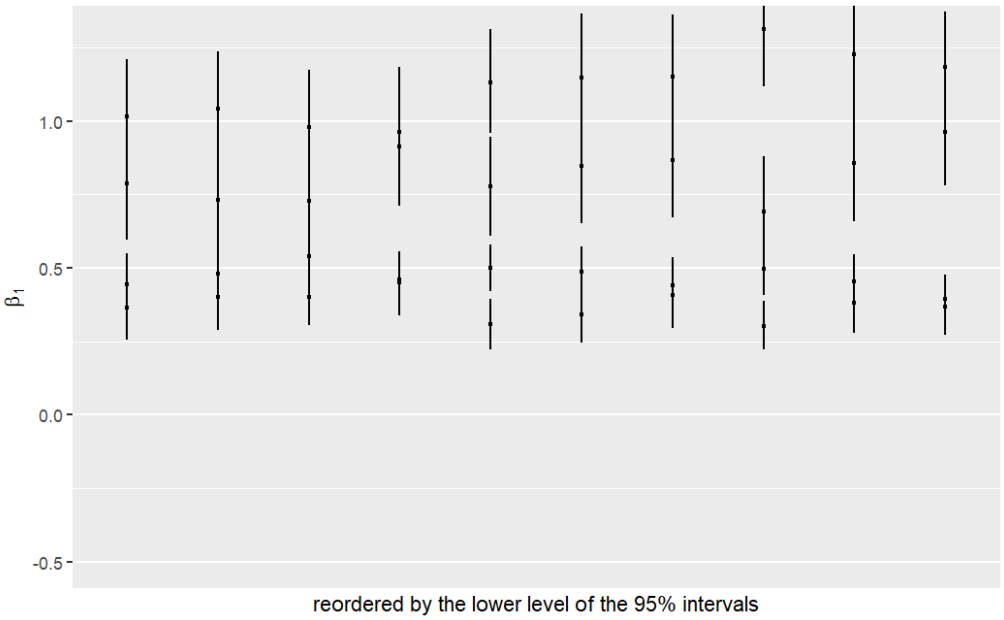
Lastly conditional effects for model 3 was computed to visualize whether the conditional effects of this model had any influence on the outcome of the data, meaning that the diagnosis had an influence on the slope/development of a child, as we hypothesized that typically developing children will generally have a steeper learning curve.



It can definitely be observed that the two groups look very similar, although the typically developing children seem to have a slightly steeper development in MLU. This closeness of the two groups could be due to the priors not being precise enough.

Sample size:

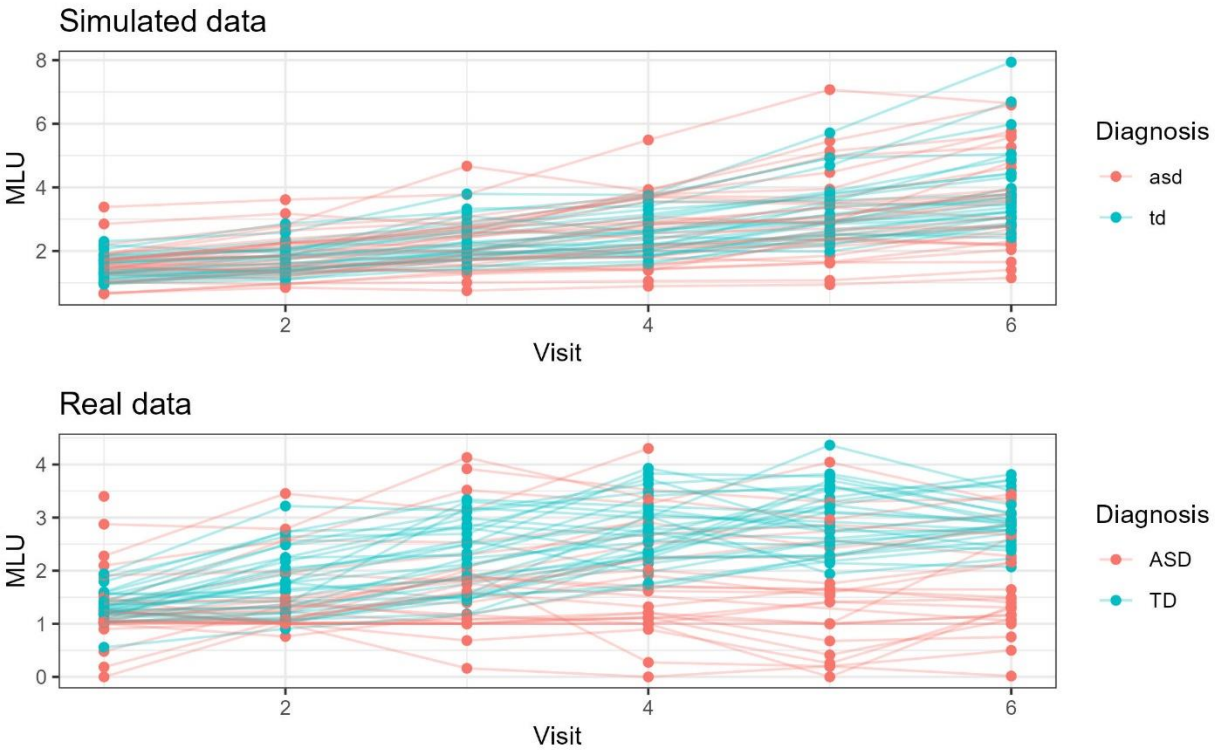
Lastly, the sample size was considered. In the real data, the sample size was 30 children per group, which is what was used during the simulation process. Other than that, simulating more datasets with varying sample sizes, can give an insight into which sample size can give the best estimates from the model. 20, 50 and 70 were sample sizes, which were simulated.



This shows the simulated data, using a sample size if 50. This is also the sample size that seems to be the most effective considering our goal, as 20 is far too low, with a very uneven distribution (*cannot be visualized, as R was not functioning properly*). Also the simulation with sample size 70 did not show results that were as even with many intervals being further away from the true effect size.

Q2:

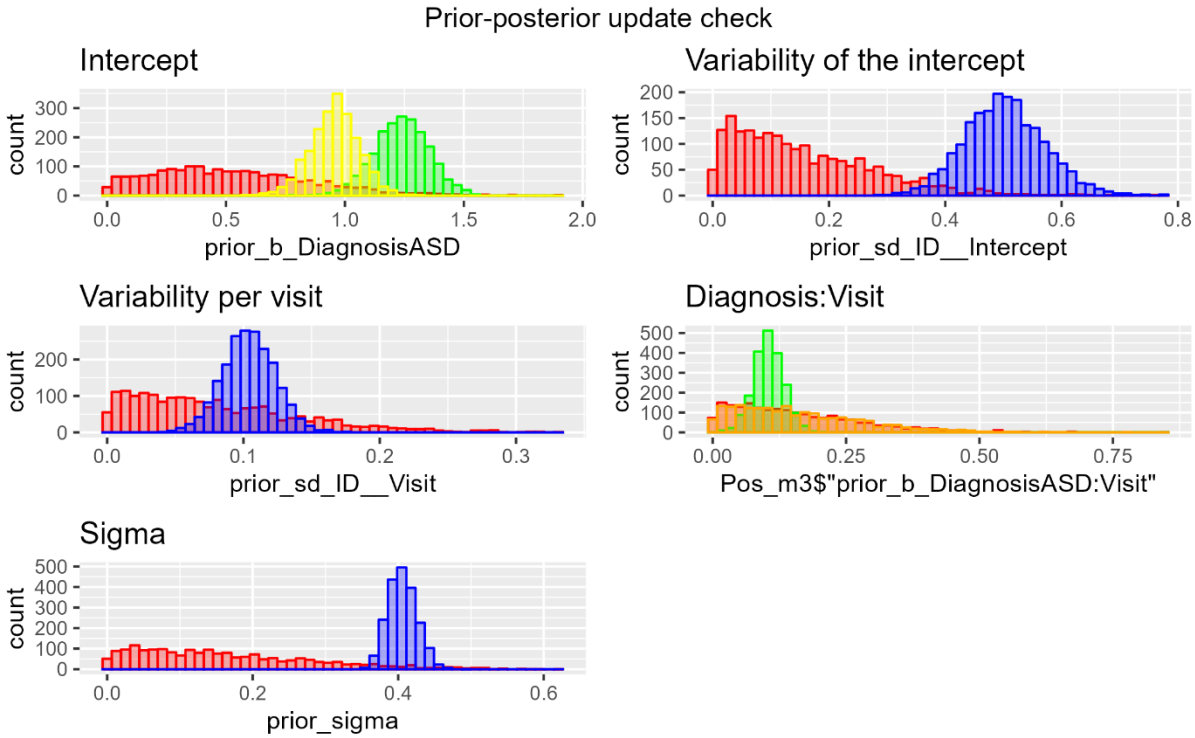
Counting the participants in *R*, by grouping them into their conditions by ID, shows that there are 31 children in the autism spectrum disorder (ASD) group and 35 children in the typically developing (TD) group. All in all, this makes up for 66 participants across the groups. For gender, the distribution is more unequal, with 55 of the 66 participants being male, while only 11 are female.



When comparing the data from the simulation and the real data collection, one can clearly observe that the real-world data is far messier than the simulated data, as there will always be some form of noise present while collecting data and also possible variables/predictors that have not been considered, such as the mood of the mother/child etc. The overall distribution of the simulation seems to match the real data though. The typically developing children seem to have a steeper development in MLU from visit to visit, whereas the development of MLU in autistic children is not as steep and the individual differences in the autistic group are generally greater than in the typically developing group.

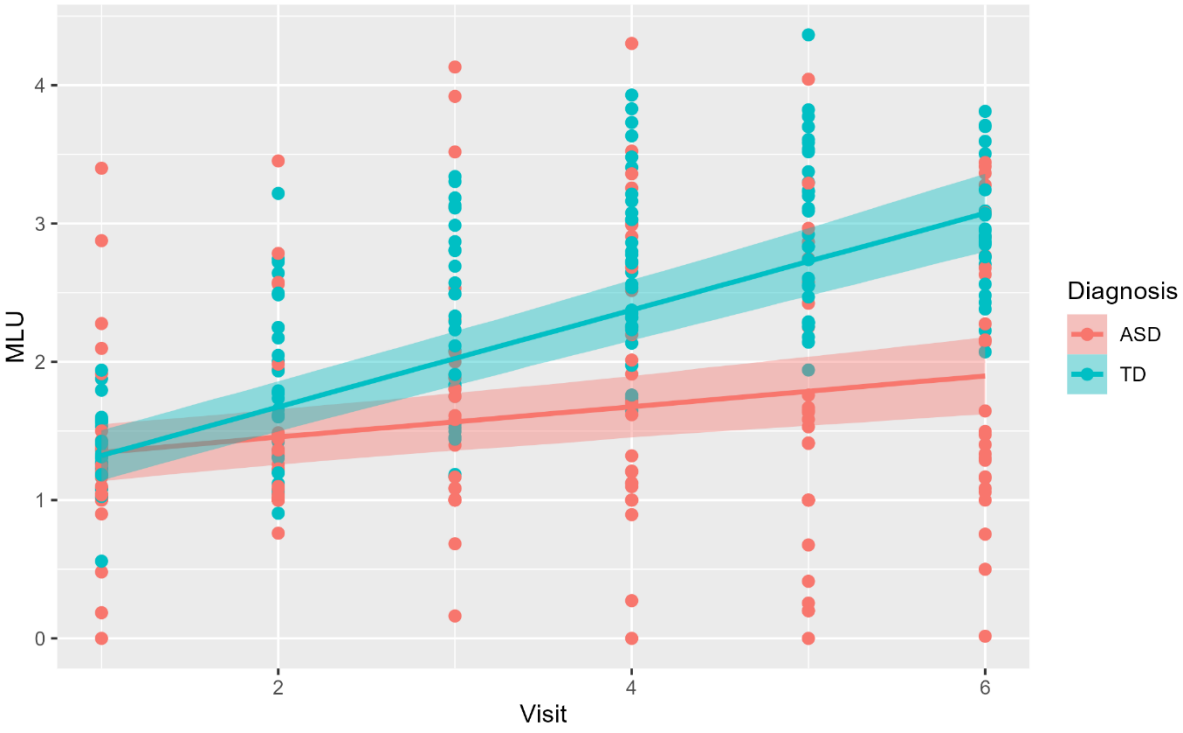
Prior-posterior update checks:

To begin with, the priors were specified again, and the models were set to run. After that, the posterior distributions were computed and compared to the prior distributions to ensure that our model has learnt something from the data:



We can clearly see a difference in the prior-posterior update checks.

For model 3, the conditional effects were visualized:



Now, we can compare the model outputs, to choose the best fitting model.

Model comparison:

The model outputs were summarized and compared, to find the best model fit for the data. Firstly, convergence and efficiency for the Markov Chains were regarded and compared. While comparing the R-hat (Rhat) values, which tell us something about how much the different chain estimates (between and within) for the model parameters agree, all three models almost exclusively share values of Rhat being 1.00. While sd(Visit) in model 3 is 1.01, this can still be accepted as a general rule of thumb is that a sample can be used, if the Rhat value is less than 1.05.

Furthermore, comparing Bulk Effective Sample Sizes (bulk-ESS) and Tail Effective Sample Sizes (Tail-ESS) values for each model, can convey information about the sampling efficiency, i.e., the minimum of effective sample sizes for 5% and 95% quantiles.

Model 3 contains the highest bulk- and tail values overall:

Model 1:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: MLU ~ 0 + Diagnosis
   Data: df (Number of observations: 352)
  Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 2000

Population-Level Effects:
             Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
DiagnosisASD     1.61      0.07     1.48     1.74 1.00     1864     1506
DiagnosisTD      2.16      0.06     2.04     2.28 1.00     2037     1561

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.88      0.03     0.82     0.95 1.00     1891     1222

Draws were sampled using sampling(NUTS). 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).
```

Model 2:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: MLU ~ 0 + Diagnosis + Diagnosis:Visit
   Data: df (Number of observations: 352)
  Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 2000

Population-Level Effects:
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
DiagnosisASD          1.20      0.13     0.95     1.44 1.00     1213      832
DiagnosisTD           1.00      0.10     0.81     1.20 1.00     1071      924
DiagnosisASD:Visit    0.12      0.03     0.06     0.19 1.00     1122      866
DiagnosisTD:Visit     0.37      0.03     0.32     0.42 1.00     1109     1283

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.74      0.03     0.69     0.80 1.00     1117     1140

Draws were sampled using sampling(NUTS). 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).
```

Model 3:

```
 Family: gaussian
  Links: mu = identity; sigma = identity
Formula: MLU ~ 0 + Diagnosis + Diagnosis:Visit + (1 + Visit | ID)
   Data: df (Number of observations: 352)
  Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;
         total post-warmup draws = 2000

Group-Level Effects:
~ID (Number of levels: 61)
                    Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sd(Intercept)           0.49      0.07     0.36     0.63 1.00      820     1277
sd(Visit)               0.10      0.02     0.06     0.14 1.00      483      869
cor(Intercept,Visit)    0.03      0.25    -0.36     0.63 1.00      494      651

Population-Level Effects:
                  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
DiagnosisASD          1.23      0.11     1.01     1.46 1.00     1470     1267
DiagnosisTD           0.97      0.10     0.76     1.16 1.00     1330     1295
DiagnosisASD:Visit    0.11      0.03     0.06     0.16 1.00     1944     1558
DiagnosisTD:Visit     0.35      0.02     0.30     0.40 1.00     1691     1231

Family Specific Parameters:
      Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
sigma     0.41      0.02     0.37     0.45 1.00      807     1241

Draws were sampled using sampling(NUTS). 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).
```

Cross validation with the leave-one-out comparison:

	Elpd_diff	**Se_diff**
Model 3	0.0	0.0
Model 2	-167.2	16.6
Model 1	-223.0	14.5

Including model weights:

	Weight
Model 1	0.000
Model 2	0.011
Model 3	0.989

Performing cross validation with the leave-one-out method is especially significant, as it can provide a much less biased outcome as compared to only splitting the data up into a test- and a training-set once. For the function loo_compare, the most favoured model will be positioned in the most upper row and have 0 in the elpd_diff column, as this column shows the difference between the preferred model and the model in said row, which in this case is subtracted by itself. Furthermore, the loo_model_weights() function also seems to favour the third model compared to the other two, as it seems to give it the most weight. This confirms the model outputs from above in that model 3 is the best fit for our data.

Looking back at our posterior predictive checks, we can confirm that model 3 has the best overall fit for the data.

Hypothesis testing:

Finally, we tested our hypotheses:

Our general hypothesis states that the development of MLU per visit in typically developing children will be greater than the development of MLU per visit for autistic children. Looking at the output this is being confirmed by the posterior probability being 1, and the evidence ratio being set to infinity (Inf). This tells us that it is very likely that the hypothesis is true.

Looking at the hypothesis that includes the fact that there is a great individual variability amongst the participants, gives another story. Only 9 out of 60 datapoints (ID), have a high enough posterior probability to be significant.

All in all, this gives the impression that on a population level, autistic children tend to be slower and experience lesser progress of speech development, compared to typically developing children, but on an individual level, there is a lot of variability. This variability could of course be due to underlying factors that were not considered during the research, but that cannot be concluded, by only observing the existing data.