Assignment 1 - Language development in autistic and neurotypical children

**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 process

To simulate data we make a function to simulate MLU for baseline and condition that takes in sample size, number of visit, mean at visit one for ASD and TD, the individual deviation for the two groups and error (n, visit, mu\_asd, mu\_td, sigma\_asd, sigma\_td, error) with 6 rows per person – one per visit. Our parameters is defined using the provided information on the differences between the development, starting point and variation of lexical ability for children with and without autism and modified by visually investigating it through histograms. They’re on a logscale as we’re working with a lognormal distribution, so to get the MLU we exponentiate the individual intercept and individual slope for each visit.

The data then look like this:



We can see from the data that each “child” has its own starting point and own slope with an increased SD for the slope for ASD and therefore bigger variability. It’s apparent that ASD contains both the datapoints with the highest MLU and the lowest, whereas TD is less scattered and follow a more linear development.

Once we have the simulated data we define our formulas. We are investigating three different models The first is a very simple model predicting MLU based on the Diagnosis:

Model 1: MLU\_f1<-bf(MLU ~ 0 + Diag

The second model adds a cross-level interaction between Diagnosis and Visit:

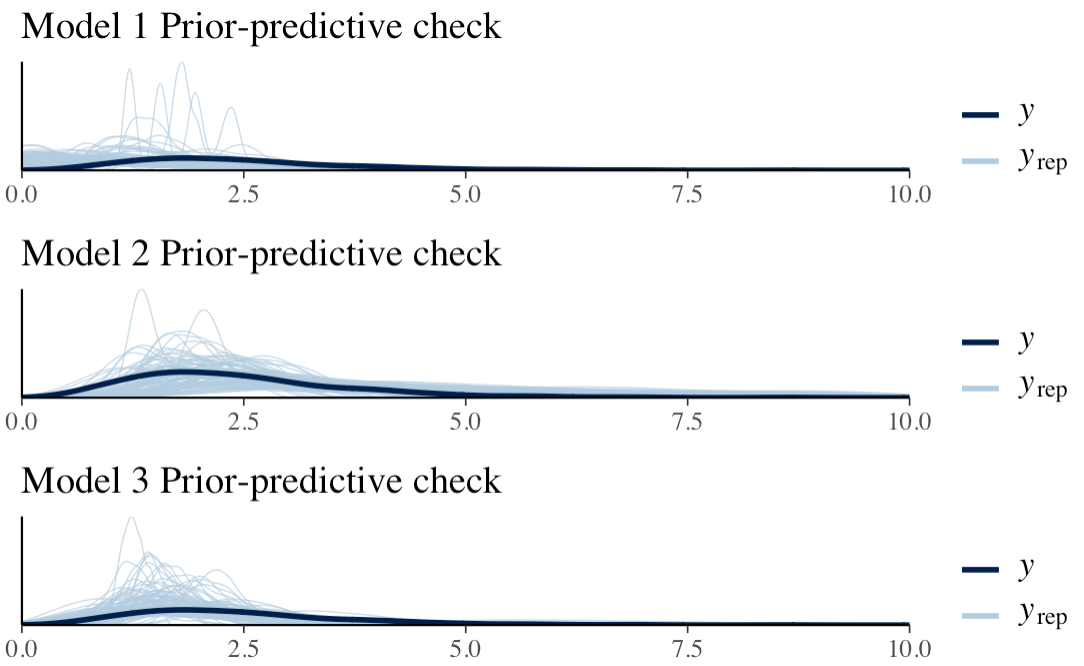
Model 2: MLU\_f2 <-bf(MLU ~ 0 + Diag + Diag:Visit)

And the third model includes random slopes per visit:

Model 3: MLU\_f3 <-bf(MLU ~ 0 + Diag + Diag:Visit + (1 + Visit|ID))

Next, we define some weekly informed priors with mu = 0 and the standard deviations from our simulation, so they are not too constraining on our posterior and let our data persuade the model that there is a significant difference.

The Prior predictive check:

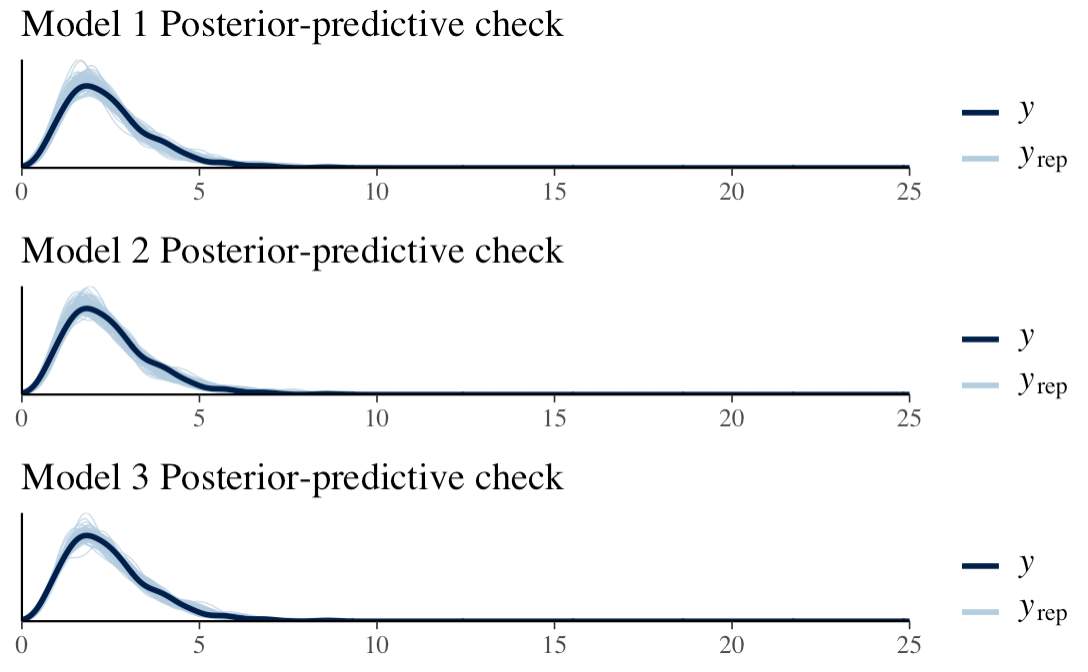


This plot shows the predictions from our prior distributions in blue and the actual data in black to check if our prior distribution are within a sensible range.

After running the pp\_check() function a couple of times, it seems as if our priors are appropriate for all three model, so we continue with these priors.

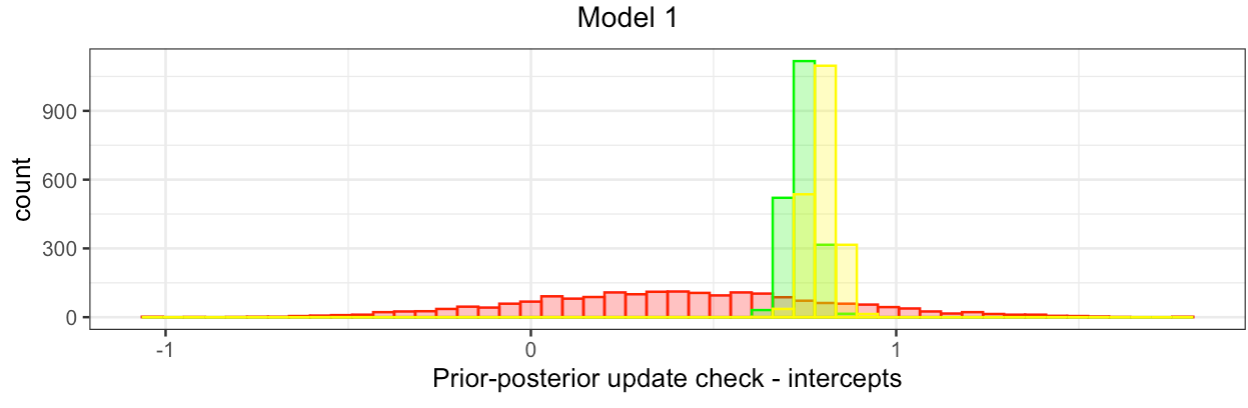
Then we fit the model to our simulated data and perform some posterior predictive checks to make sure that the model captures the data:

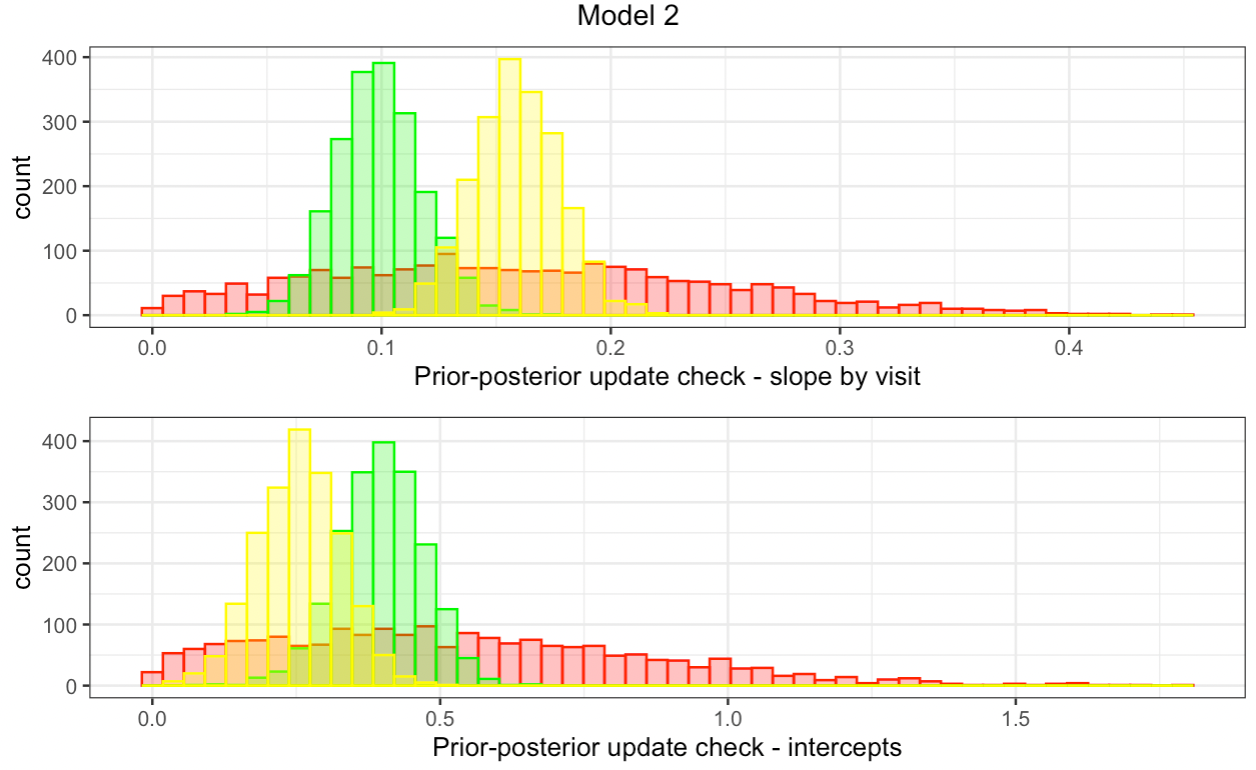
The posterior predictive check:

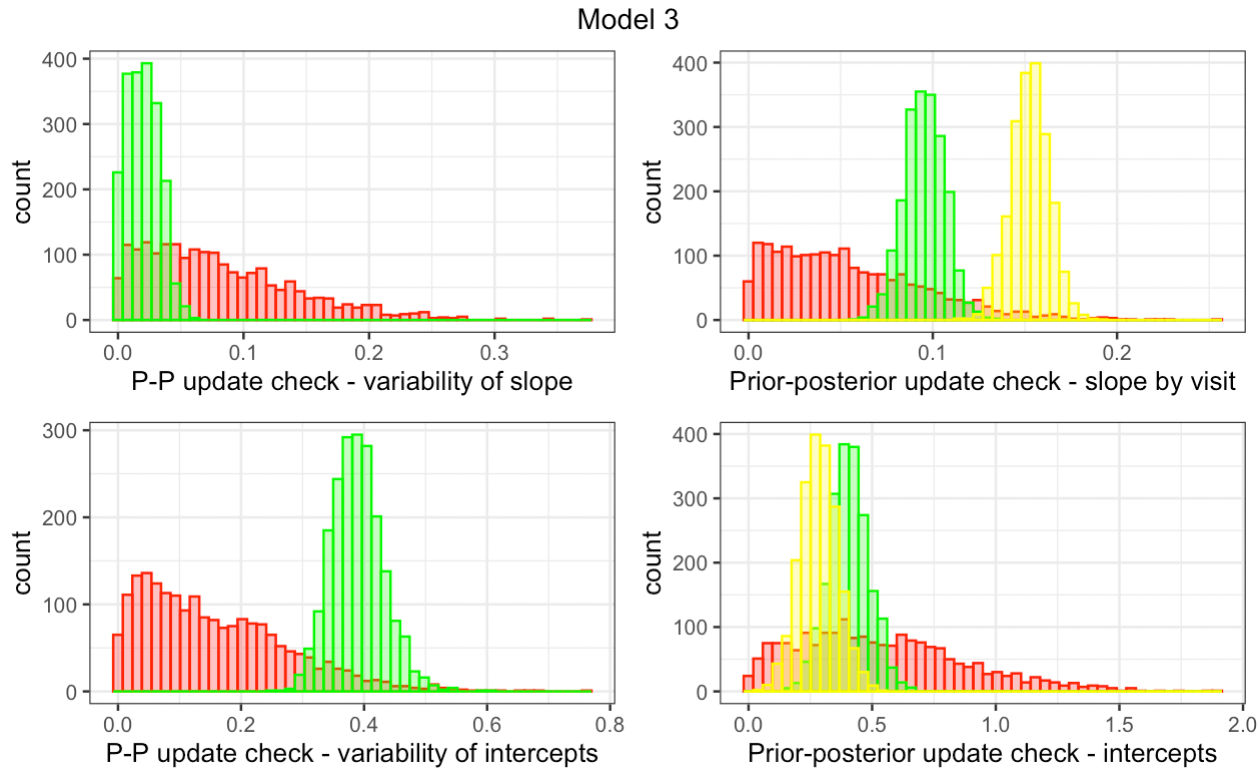


The posterior predictive checks look good for all the models and very similar.

We also make some prior-posterior update plots to visualize how the models updates after meeting the data.







Prior – posterior update check - Prior in red, ASD in green, TD in yellow

The prior posterior update check for model 1 for the intercept look good since the distributions are very narrow and confident in their estimate, but it shows a lot of overlap between the two groups, and hence similar intercepts.

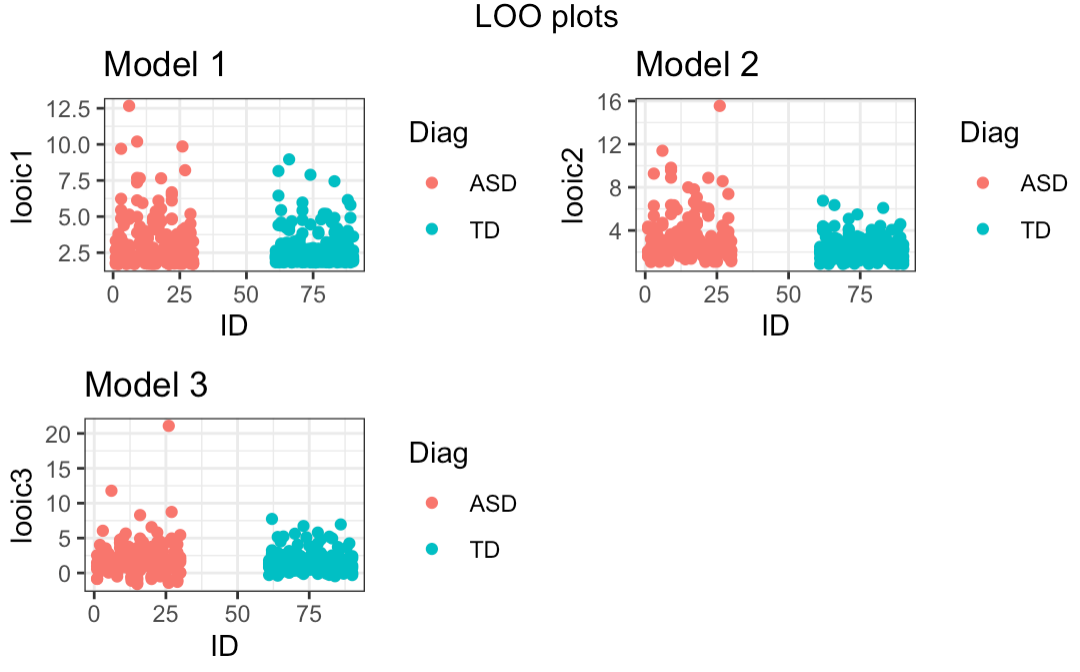
The posteriors are found nicely within the prior for both plots of model 2 with some overlap for both slope by visit and the intercept.

slope by visit in model 3 shows a great divide between ASD and TD with not much overlap and therefore possibly a significant difference between the development of lexical ability for the two conditions, whereas the plot for the intercepts has much more overlap and reveals a more similar starting point. Both distributions in the plots are pretty narrow, so we can accept our posterior as very confident.

Our prior posterior update check for variability of the slope shows a very pointy - and therefore confident posterior located well within the range of our prior, which is very nice. The prior posterior update check for variability of the intercept pushes towards the tail which can indicate that our priors might be a bit to narrow.

To assess what model to continue with for further empirical analysis, we conducted a leave-one-out cross validation with the following results:

|  |  |  |
| --- | --- | --- |
| Model | Elpd\_diff | Se\_diff |
| MLU\_m3 | 0.0 | 0.0 |
| MLU\_m2 | -202.4 | 14.6 |
| MLU\_m1 | -241.4 | 16.1 |



The Loo comparison points to Model 3 as having the best expected predictive accuracy, and

if we include conceptual reasoning about the model, it seems proper that the lexical ability of the children from the two groups develop at a different pace, so with this added to the results from the Loo-comparison, we will continue with Model 3.

It’s a Multilevel model with random slopes for each visit and add a cross-level interaction between Diagnosis and Visit: MLU\_f3 <-bf(MLU ~ 0 + Diag + Diag:Visit + (1 + Visit|ID))

In the model summary we can check the, Bulk\_ESS and Tail\_ESS to explore the sampling efficiency. Bulk Effective Samples Size assesses the efficiency of the posterior mean and medians estimates and the Tail Effective Samples Size looks at posterior variance and computes the minimum of effective samples sizes for 5% and 95% quantiles.

Our model has 2 chains and 2000 iterations, and a rule of thumb seems to be at least 100 for both Bulk and Tail per chain for it to be reliable. Our Model has values that that ranges from 701 to 2992 for the different parameter, so it seems we can trust the reliability of our model.

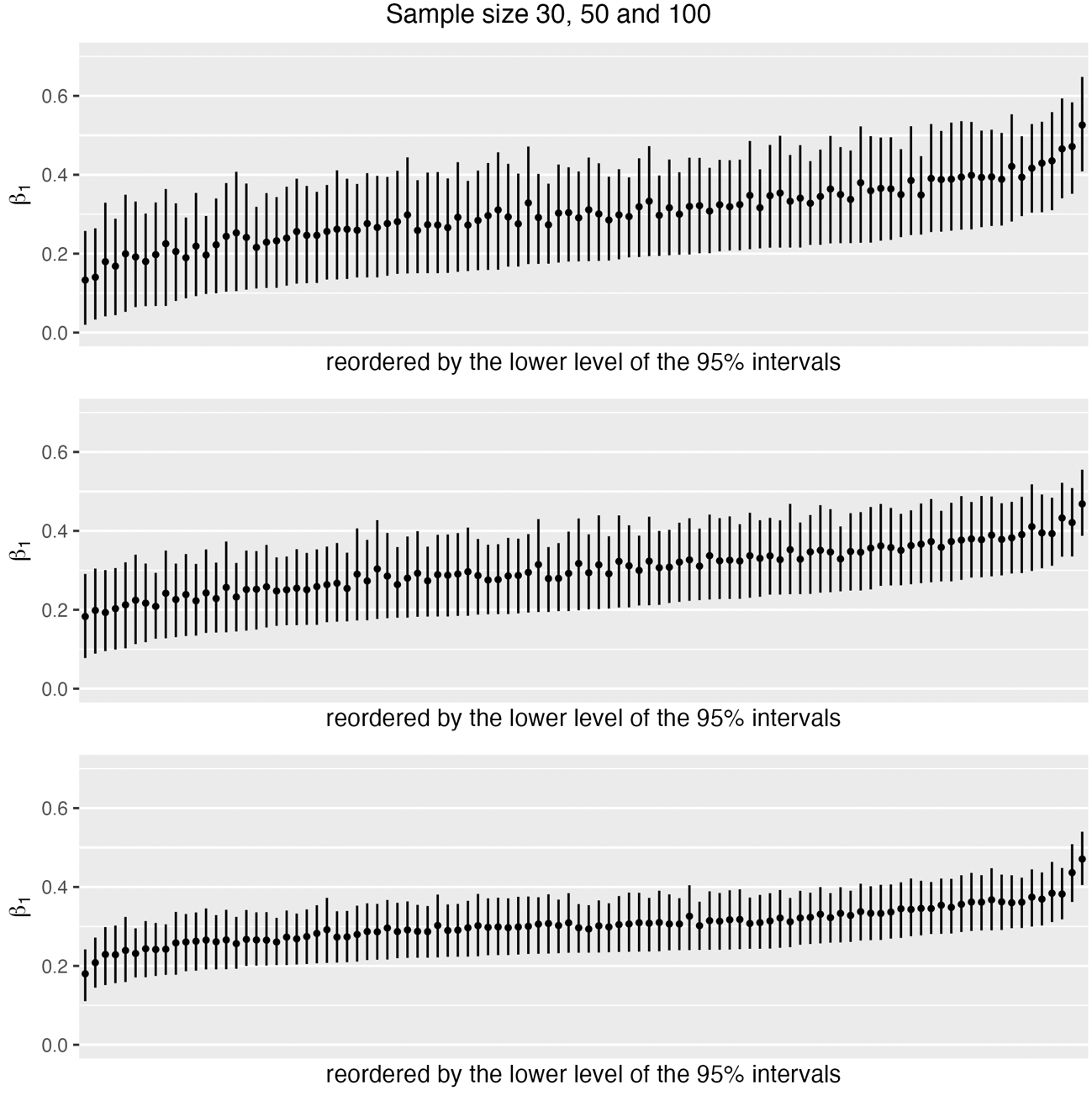
The Rhat value in our summary is 1, so suggesting with great confident that our model has converged.

Summary of output of our model:

|  |
| --- |
| Family: lognormal |
| Links: mu = identity; sigma = identity  Formula: MLU ~ 0 + Diag + Diag:Visit + (1 + Visit | ID)   Data: d (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) 0.39 0.04 0.31 0.49 1.00 823 1285 sd(Visit) 0.02 0.01 0.00 0.05 1.00 180 506 cor(Intercept,Visit) -0.15 0.44 -0.92 0.83 1.00 1508 720 |
| Population-Level Effects:  Estimate Est.Error. l-95% CI. u-95% CI. Rhat Bulk\_ESS. Tail\_ESS DiagASD 0.37 0.08 0.21 0.53 1.00 701 1233 DiagTD 0.22 0.08 0.06 0.37 1.01 626 1035 DiagASD:Visit 0.15 0.01 0.12 0.17 1.00 2992 1560 DiagTD:Visit 0.21 0.01 0.19 0.24 1.00 2897 1480 |
| Family Specific Parameters:   Estimate Est.Error. l-95% CI u-95% CI. Rhat. Bulk\_ESS Tail\_ESS sigma 0.28 0.01 0.26 0.30 1.00 970 1201 |
| 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). |

Now, we know that we can accept our model, so instead we look into the sample size. We do this by looping through 3 different amounts of “participants”: 30, 50 and 100 as described in the tutorial by Kurtz.

We base it on the function we made for simulating data but change it a bit to only take in seed and sample size as parameters. Then we simulate, fit our Bayesian model and iterate over 100 simulations to extract the β\_1 for each parameter in our model.



The β1 coefficient is the estimate of the difference between the means of ASD and TD and two of the lines surrounding our estimates is visualizing: our ideal effect size at 0.5 and the null hypothesis at 0.

We can see from the plot that the widths of our estimates that fall within a 95% probability range are with a few exceptions all captured by the interval between 0 and 0.5. This is a very narrow interval and therefore some extremely precise estimates, with an average width of 0.25 for a sample size of 30, 0.2 with a sample size of 50 and 0.14 with a sample size of 100. Kurz mentions an average width of 0.2 to be considered small within Psychology, so we can accept even the one with 30 participants as reliable, however the median effect within Psychology is 0.26, so if we were to be more certain of our results, it could be a good idea to have a sample size of 50.

We also computed the power. With a critical value set to 0.05, power is the probability that we will have a value that falls below this threshold, when indeed there is a difference between the two means. A power of 1 suggests that 100 of our 100 simulations will correctly reject our null hypothesis, so there is a significant difference between the two means. The power for the three different sample sizes were all 1, so n = 30 is very sufficient.

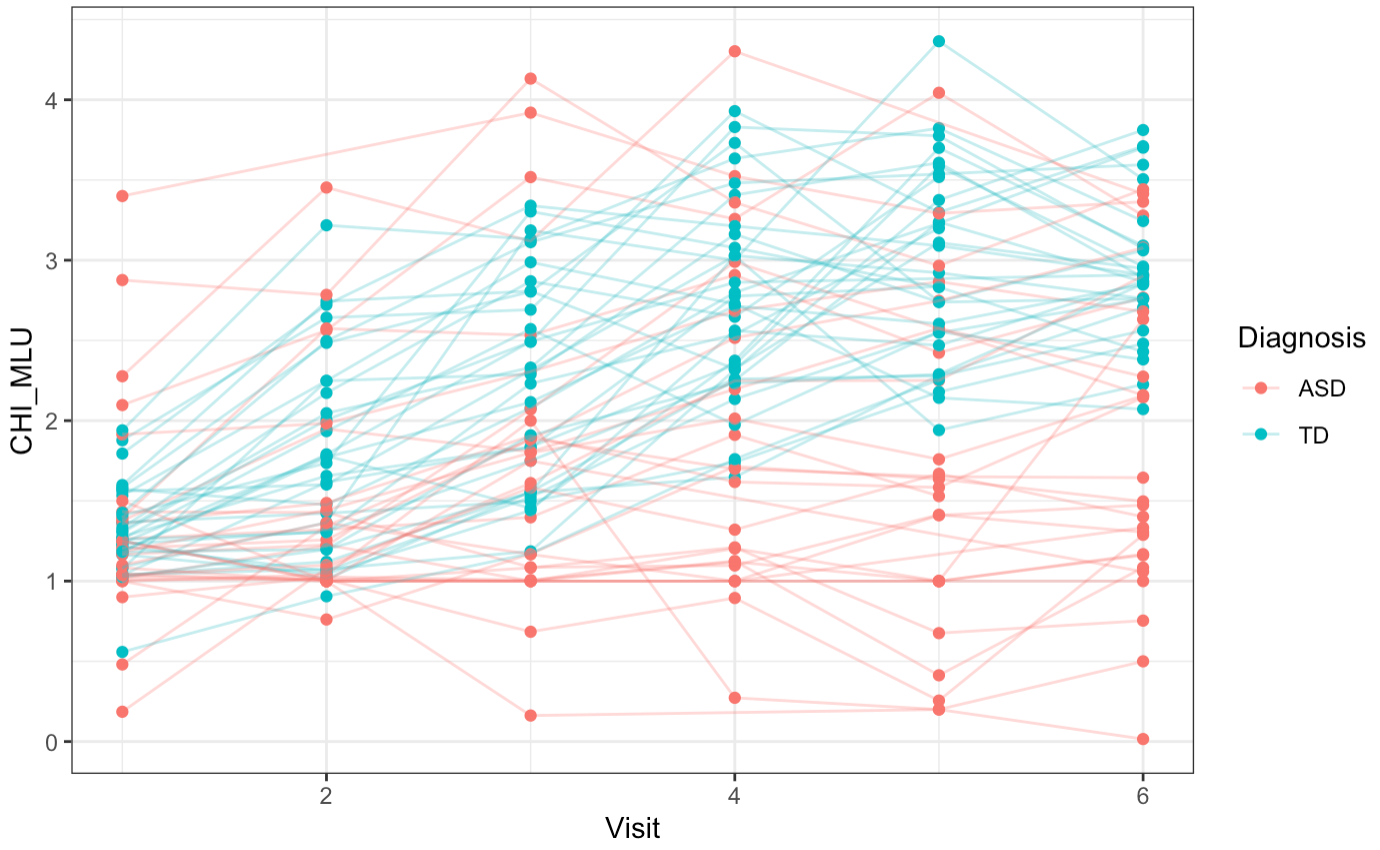
|  |  |  |  |
| --- | --- | --- | --- |
|  | n = 30 | n = 50 | n = 100 |
| Average width | 0.25 | 0.2 | 0.14 |
| Power | 1 | 1 | 1 |

**Q2 - Briefly describe the empirical data and how they compare to what you learned from the simulation (what can you learn from them?).**

Our real data consists of a data frame with 22 variables and 372 observations and a sample size of 66, with an equal distribution of typically developed children (TD) and children with autism (ASD). This is a totally acceptable sample size, as we learned from our Power-analysis with the simulated data that a sample size of 30 per condition is sufficient to get reliable results.

With a much larger range of variables on the actual data, there is a few more parameters that could possibly be included in our model. These contain for example age, gender, ethnicity and the mothers MLU. The age ranges from 18 months to 62 months, so quite a difference.

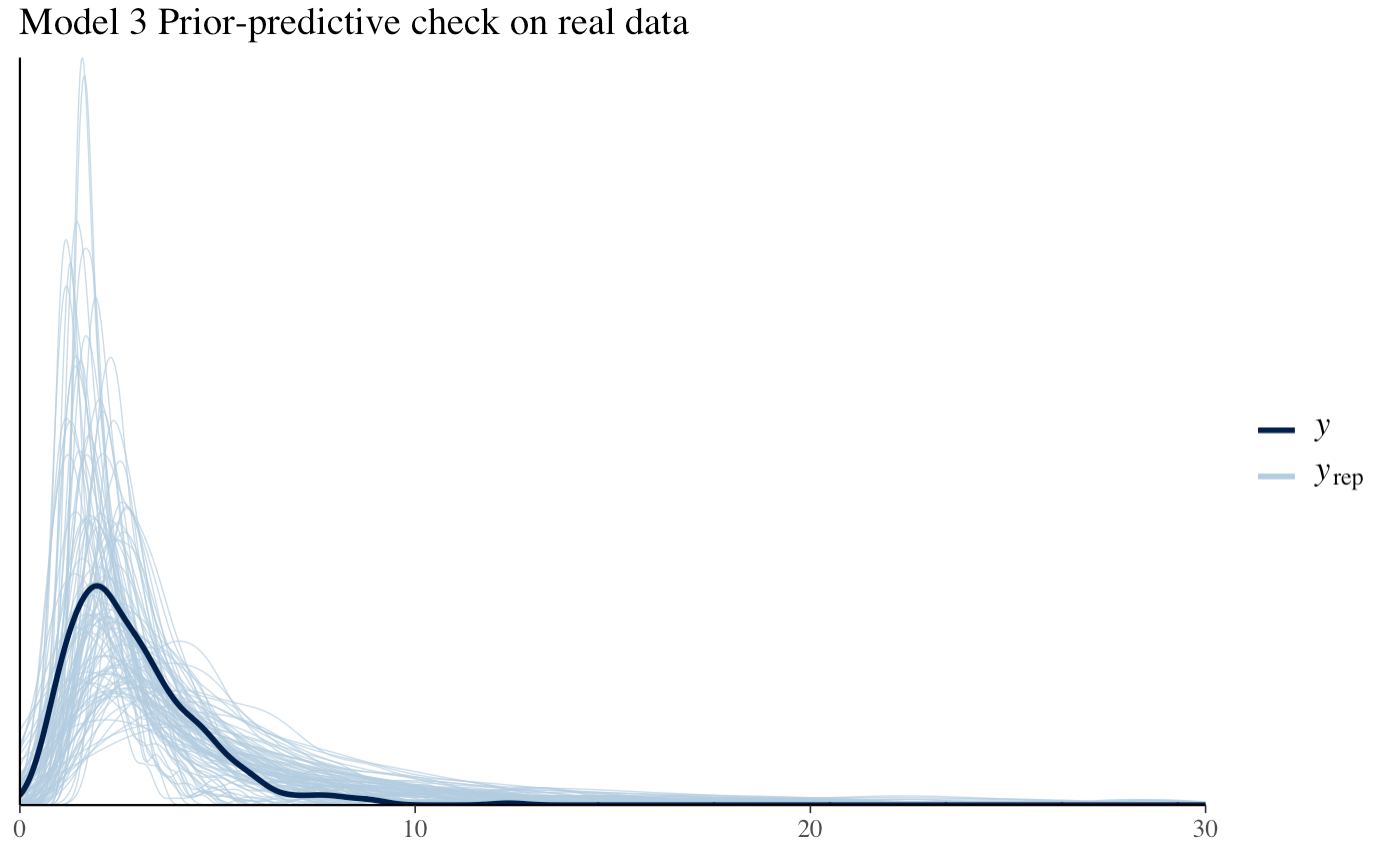
We filter out the children who didn’t complete all 6 visits, and plot the data:

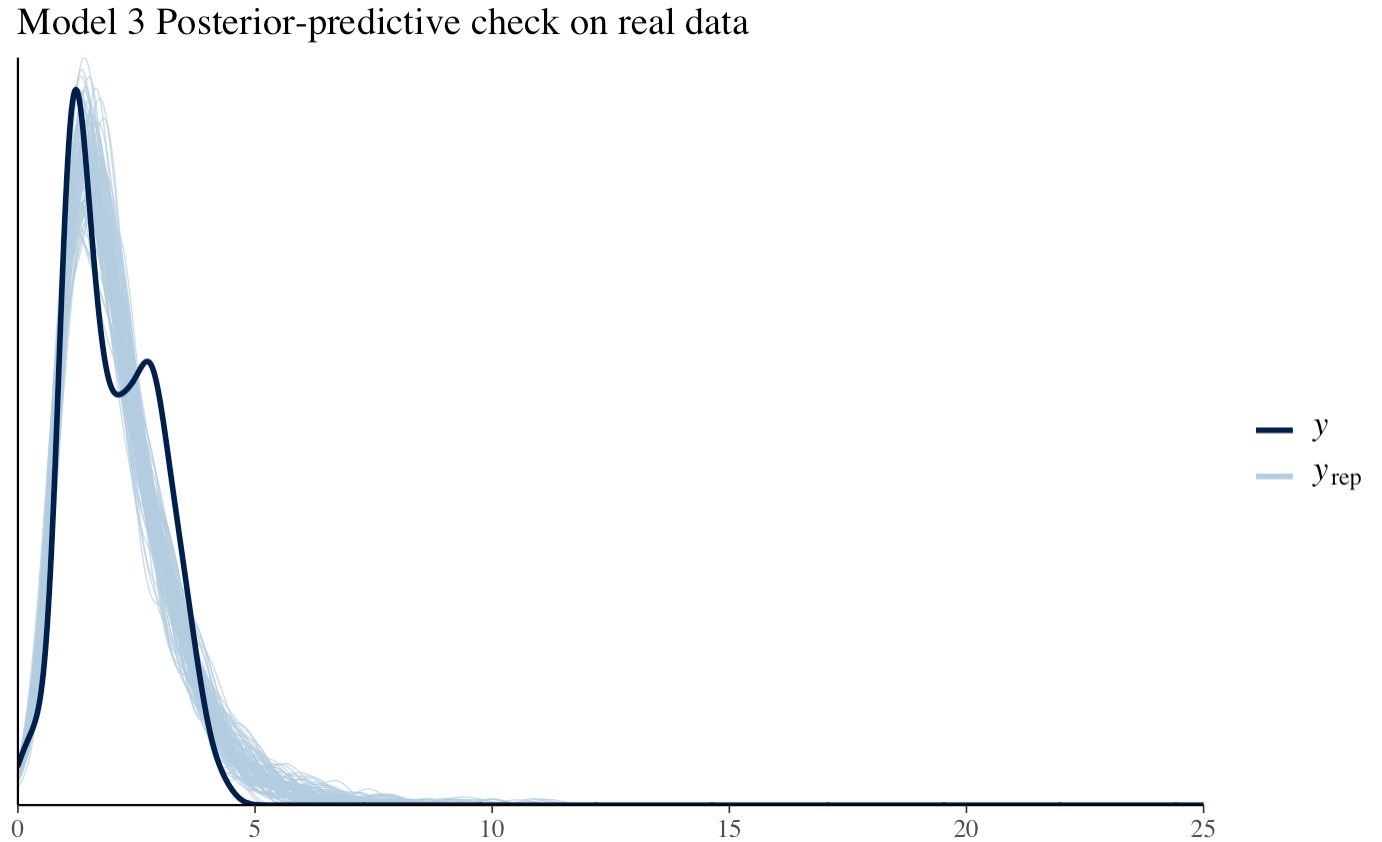


When we plot our real data, of course we see a less clean pattern as the simulated, however there is still an obvious difference between the two conditions. As we saw in the simulated data and predicted from looking at relevant literature, there’s a far greater variability within the ASD group who are represented both at the very bottom and top of the plot i.e. having best and worst lexical ability, whereas the typically developed kids follow a more similar straight development.

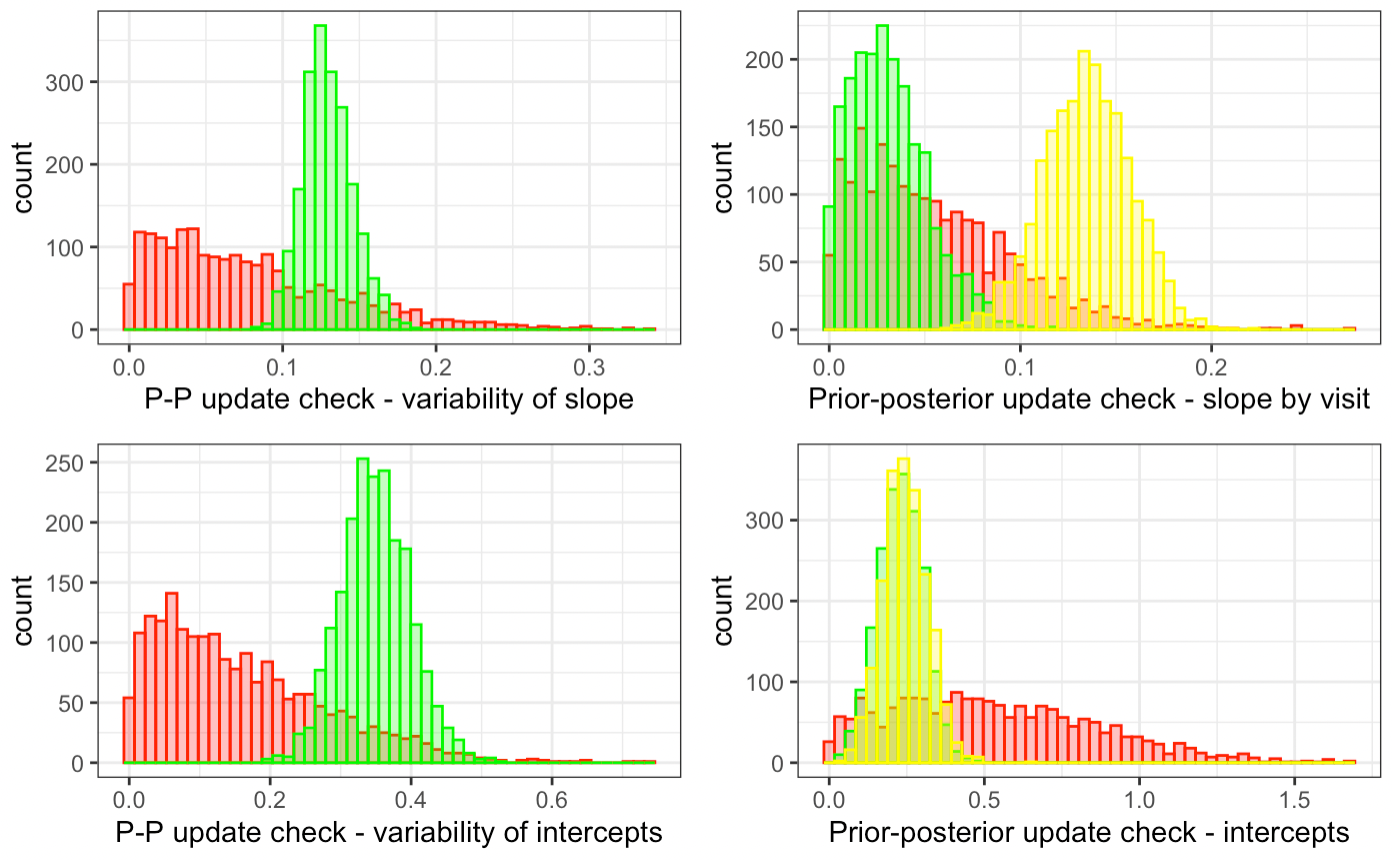
To further investigate our real data, we define our formula. Since, the real data looks similar to our simulation, we use the same model for the simulated data with random slopes for each visit and adds a cross-level interaction between Diagnosis and Visit: MLU\_f3 <-bf(MLU ~ 0 + Diag + Diag:Visit + (1 + Visit|ID)).

We set up the model with the same weekly skeptical priors as in the simulation and fit the actual data.





Both the prior predictive check and the posterior predictive check looks good. The priors are within a sensible range and the posterior follows the data pretty well.



Prior – posterior update check on the real data - Prior in red, ASD in green, TD in yellow

The prior posterior update check for slope by visit looks alright ASD and TD overlaps a little, but it still suggests different slopes for the two conditions. The plot for the intercepts has almost totally overlapping distributions, so almost identical intercepts.

Summary:

|  |
| --- |
| Family: lognormal  Links: mu = identity; sigma = identity |
| Formula: CHI\_MLU ~ 0 + Diagnosis + Diagnosis:Visit + (1 + Visit | Child.ID)  Data: df (Number of observations: 349)  Draws: 2 chains, each with iter = 2000; warmup = 1000; thin = 1;  total post-warmup draws = 2000 |
| Group-Level Effects:  ~Child.ID (Number of levels: 61)  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  sd(Intercept) 0.35 0.05 0.26 0.45 1.00 797 1191  sd(Visit) 0.13 0.02 0.10 0.16 1.01 214 483  cor(Intercept,Visit) -0.48 0.14 -0.71 -0.19 1.01 156 523 |
| Population-Level Effects:  Estimate Est.Error l-95% CI u-95% CI Rhat Bulk\_ESS Tail\_ESS  DiagnosisASD 0.23 0.07 0.09 0.37 1.00 787 912  DiagnosisTD 0.25 0.07 0.11 0.39 1.00 1022 943  DiagnosisASD:Visit 0.03 0.02 0.00 0.08 1.01 718 891  DiagnosisTD:Visit 0.13 0.02 0.09 0.18 1.00 515 754 |
| Family Specific Parameters:  Estimate. Est.Error l-95% CI. u-95% CI Rhat Bulk\_ESS Tail\_ESS  sigma 0.27 0.01 0.24 0.29 1.00 1029 1494 |
| 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). |

The summary of our model provides some fine values. On neither the population or individual level do Rhat exceed 1.1, so the model has converged and there’s no need to run more iterations or redefine our priors. Bulk\_ESS and Tail\_ESS are a bit lower than in the simulation with a Bulk\_ESS for the correlation of intercept and visit down to 156. The rest of the values exceed 200 both for the group level and population level indicating pretty reliable estimates of the mode of the distribution.

The average difference between the slope for ASD and TD (effect = `b\_DiagTD:Visit` - `b\_DiagASD:Visit`) is at 0.1 which is pretty decent, indicating an increase of roughly 10%.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **EffectMean**  <dbl> | **EffectSE**  <dbl> | **EffectLowCI**  <dbl> | **EffectHighCI**  <dbl> | **EffectWidth**  <dbl> | **ER**  <dbl> | **Cred**  <dbl> |
| 0.1036472 | 0.03069453 | 0.04226 | 0.1640059 | 0.1217459 | 665.6667 | 0.9985 |

Some additional predictors could be interesting to include in another model. As the child’s lexical ability is also largely shaped by its environment, I would presume that this would help the model if included. Thus, I would argue that Socialization could be a good predictor to include, since being around other people and being exposed to daily talk really help practice speaking and acquire new skills and therefore has a big influence on your development.

In the same spirit mother’s MLU could also be an additional predictor, since children are eminent at mimicking and picking up new skills from the environment and with the assumption that the children spent much if not most of the time with their mother, her language complexity level would influence the development of the child.