Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Portfolio Exam - Part 1 | Methods 3 F2022, CogSci @ AU

Group 3 – Ditlev Kræn Andersen (699440), Bryan Roemelt (670705),

Patrik Molnar (694214), Sára Anna Szabó (690442), Manuela Skov Thomasen (650504)

12/10/2022

Knitted script for assignment 1 at page 17 to 64

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?

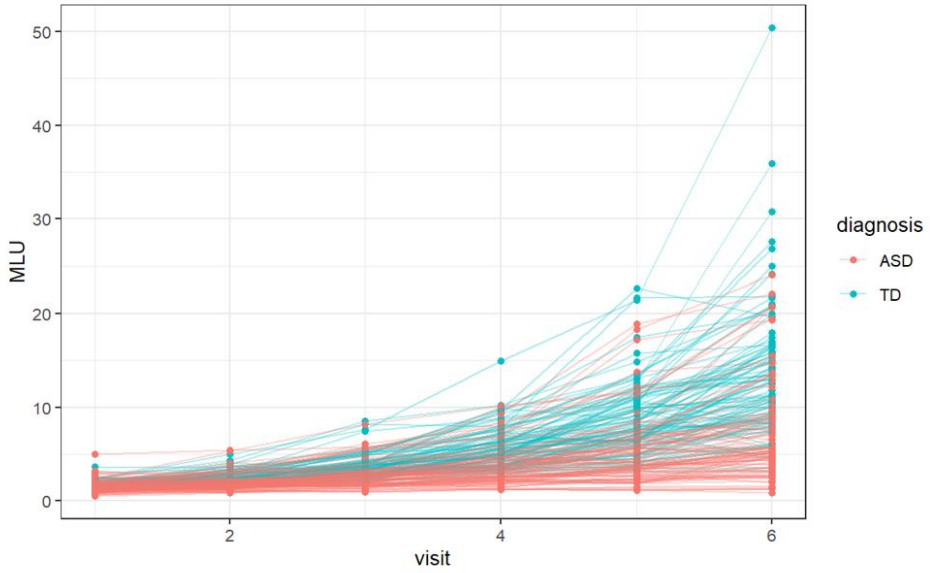
(BR, MST) Our simulations were designed to provide an informed understanding of the model we need to build for practical purposes, such as predicting the number of participants necessary to run a significant study and to suggest factors with significance in the linguistic development of autistic children.

(BR)Considering that our cleaned experimental data was sourced from the performance of 29 ASD and 32 TD participants, the intention of raising the simulated population sample to 200 was to test the diagnostic precision of the prior data for our estimations and modeling strategy going forward.

(MST, BR) We ran a simulation of 200 participants, 100 classified as typically developed children (TD) and 100 diagnosed with Autistic Spectrum Disorder (ASD), with six independent visits each. The intercept and slope for each participant was sampled from a normal distribution based on the prior data from the literature described in the assignment; these variables were then used to simulate the Mean Length of Utterance (MLU). Plot 1 shows the results of the simulation, where we can see the MLU increases across visits and, as anticipated, it looks like there is a bigger increase in MLU for the TD participants.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

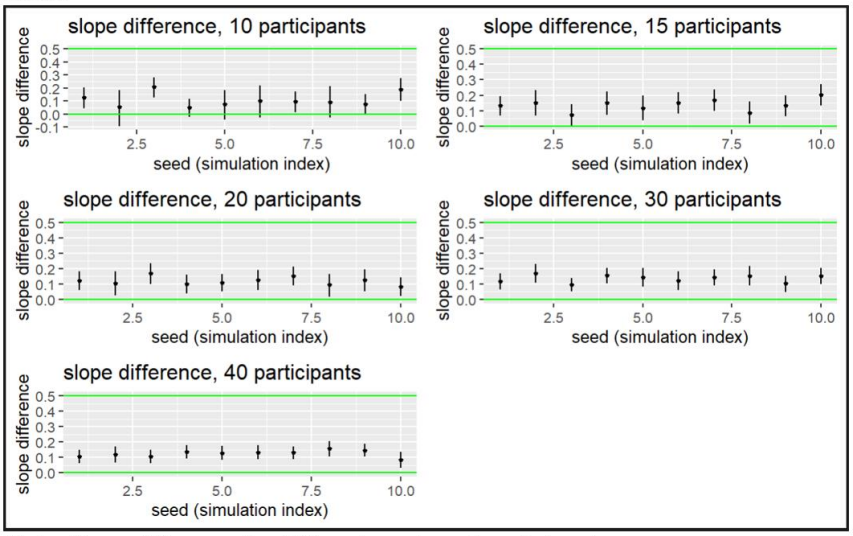
Plot 1



Note: plot 1 shows the development in mean length of utterance across visits, grouped by ASD and TD

(MST) The simulated data was used to test our model and observe how it operates on data. We found that the model learns from the defined priors. To determine an appropriate sample size, a power analysis was run with the objective of finding a sample size that would give us an effect size of 0.8 or above. The result of the power analysis is visualized in plot 2 and 3. As seen in tibble 1 we achieve an effect size of 0.9 at 15 participants.

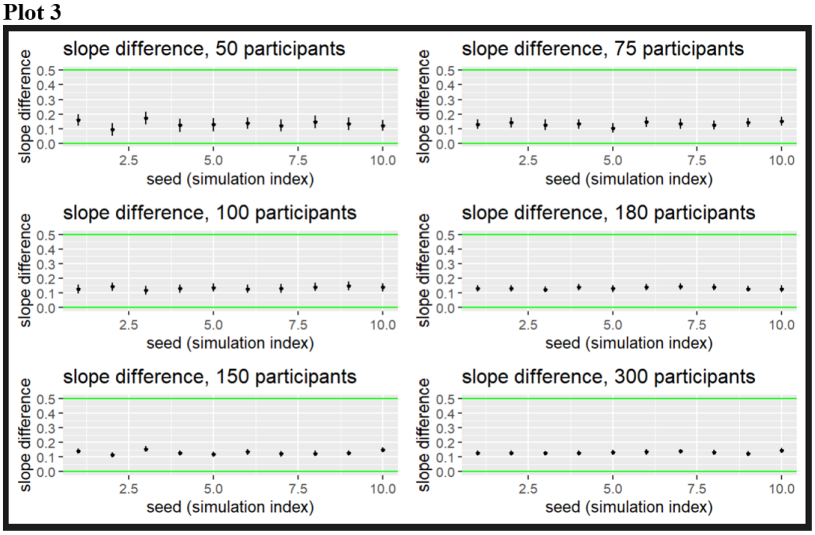
Plot 2



Note: Slope difference for different number of participants

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Plot 3



Note: slope differences with different number of participants

Tibble 1

A tibble: 11 × 2

power <dbl>	number_of_participants <dbl>
0.4	10
0.9	15
1.0	20
1.0	30
1.0	40
1.0	50
1.0	75
1.0	100
1.0	180
1.0	250
1.0	300

Note: outcome of power analysis

(MST) The standard error is an indication of the precision of the estimates, so the lower the standard error, the higher the precision. A small standard error means that there is a small variation in the sample statistics across many repeats of the experiment. The standard error depends on the variability in the data and the sample size and is calculated by:

$$standard\ error = \frac{standard\ deviation}{\sqrt{sample\ size}}$$

(MST) We have already looked at how an increase in sample size affects the effect size, and the bigger the sample size, the bigger the precision and effect size. Standard deviation describes the variability in the data, which is due to the variation between the participants, but can also be caused by the methodology or some lurking variable, something that affects the results that has not been considered. Therefore, it is important to investigate what can cause the variability in the data and consider how to avoid it.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Q2

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

(SS) During the data cleaning process, certain data points were removed, so not all participants provided an equal number of data points. This is in accordance with the realistic expectation of children not being equally communicative at all times or having too much noise in the data, etc. (For example, the child with ID 1 only produced one data point that could be included in the final dataset). The data was recorded over the span of 6 visits, where parent-child interactions were recorded and later analyzed.

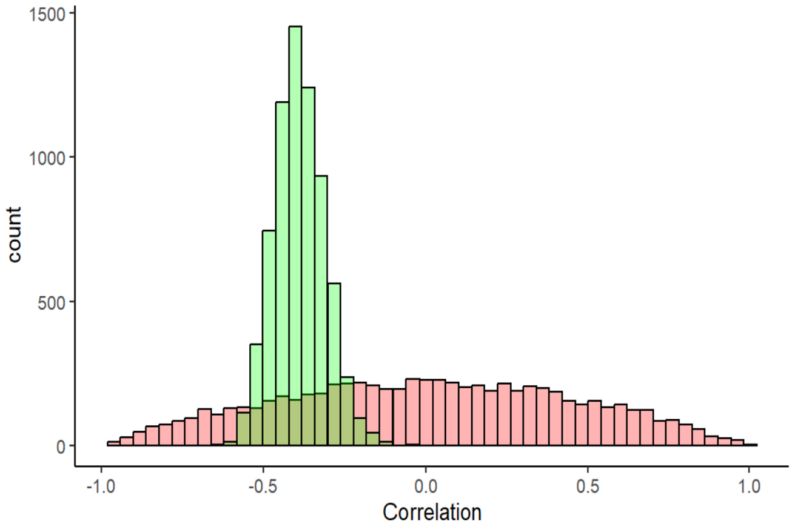
(SS) As a result of the cleaning, we obtained an R tibble containing 352 observations of 20 variables. Participants were anonymized via numeric IDs. 29 of the children were diagnosed with Autism Spectrum Disorder (ASD) and 32 were typically developed (TD). The variables describe the demographics of each participant, their diagnosis, measures of both verbal and non-verbal IQ, and their and their respective mothers' MLU throughout the 6 visits.

What did we learn from the simulated data?

(MST, SS) In contrast to the empirical data that contained incomplete data sets for some child participants, the simulated data contains the full 6 data points per participant. The simulated data showed us that MLU increases across visits and that the increase is bigger for TD children. Furthermore we saw that our model became more confident when running on the data and not just priors. Plot 4 is one of the plots made as part of the prior-posterior update check on the model. Here we see that the model becomes more confident as the standard deviation decreases for the posterior intercepts. Lastly the simulated data helped us determine a preferable sample size of 15 to get the wanted effect size of 0.8 or above.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

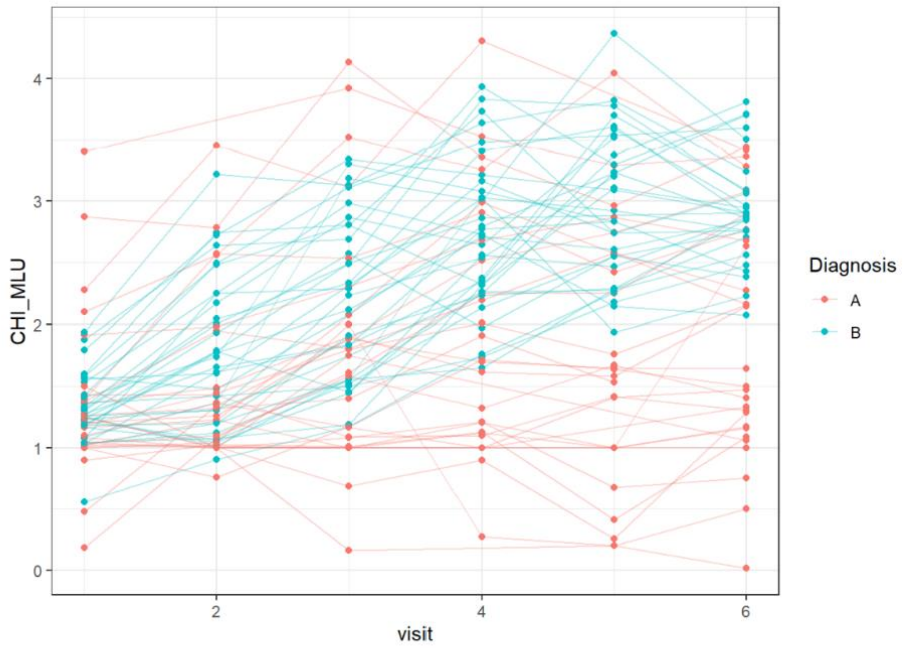
Plot 4



Note: prior-posterior update check on correlation between varying intercepts. Red is the prior and green is the posterior variable.

(MST, SS) Plot 5 is made on the empirical data and shows the MLU across visits, just like plot 1. With the empirical data we do not see as clear a difference between the ASD and TD group as we did with the simulated data. Plot 5 shows that some children still have an increase in MLU across visits but it is not all children and there is generally more variance in the empirical data.

Plot 5



Note:development in MLU across visits, grouped by diagnosis (A is children with autistic spectrum disorder and B is neurotypical children)

-Briefly describe your model(s) and model quality. Report the findings: how does development differ between autistic and neurotypical children

(SS) model was defined as the following:

$$CHI\_MLU \sim 0 + Diagnosis + Diagnosis{:}\,visit + (1 + visit|id))$$

(SS) Our population-level effects gave an intercept estimate of 0.44 for autistic children with an error of 0.04 and of 0.49. with an error of 0.02 for neurotypical children. The estimate for the difference between visits was -0.01 (error: 0.02) for autistic children and 0.07(error:0.02) for neurotypical children, meaning that the model estimated autistic children to get lower MLU across visits.
The group-level effect estimated a standard deviation of the intercept between participants to be 0.42 (error: 0.05) and of the slope 0.15 (error: 0.02) which tells us that there is a lot of individual variation, which was also what we could see in plot 5.

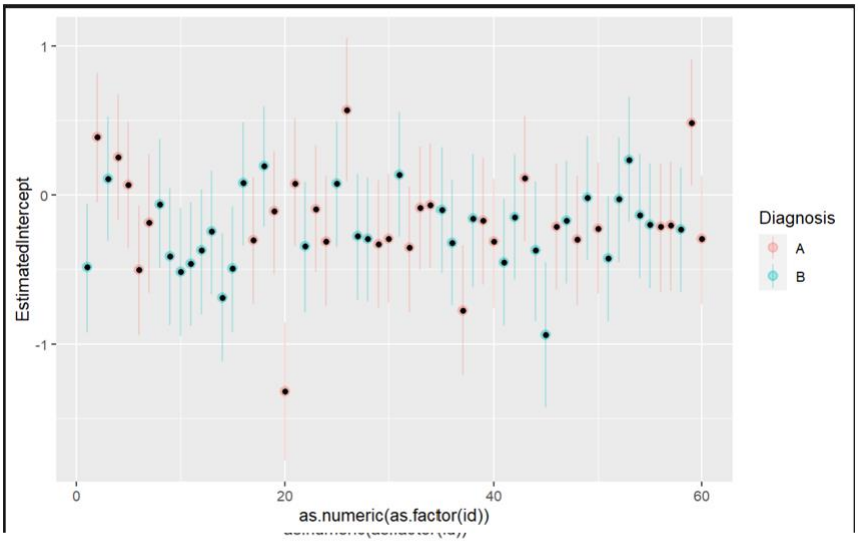
- *which additional factors should be included in the model?*

(DK, BR) Using the population level estimates for slopes and intercept for the two groups, we want to visualize the individual variability within the data. In plot 6 (intercept) and 7 (slope) we have visualized the individual estimates. A refers to Autistic children and B refers neurotypical children. We see that there is a lot of variability between participants, especially in terms of their intercepts. With plot 6. In plot 7, we can visually confirm what the model is also telling us, neurotypical children, tend to have a higher slope. We can also see that some autistic children have a high slope, telling us that within the autistic group there are outliers.

We fitted a model with random intercepts and slopes. By allowing for these random effects, we hoped to capture the individual variability between participants.

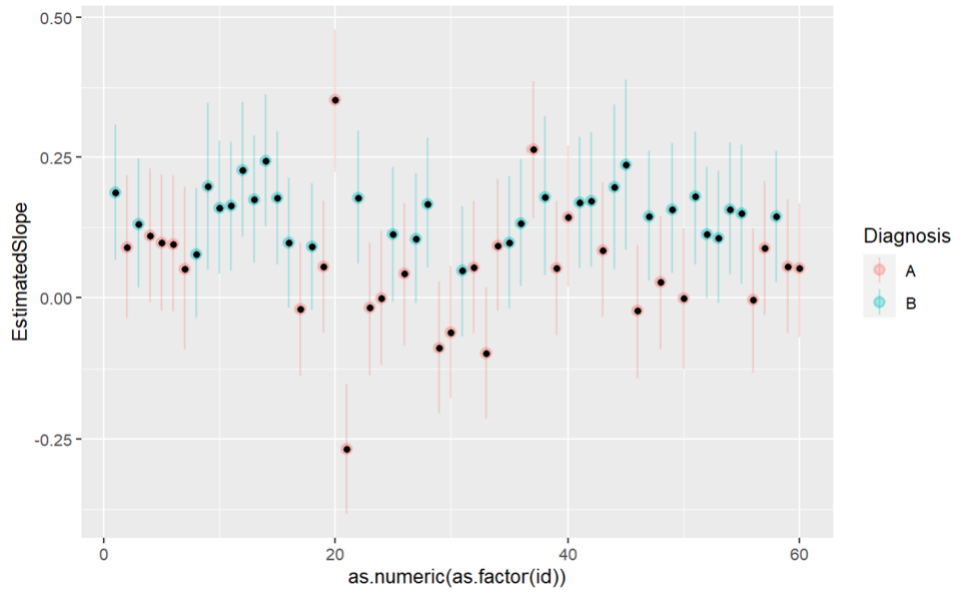
Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Plot 6



Note: visualization of estimates of intercepts

Plot 7



Note: visualization of estimates of slopes

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Portfolio Exam - Part 2 | Methods 3 F2022, CogSci @ AU

Group 3 – Ditlev Kræn Andersen (699440), Bryan Roemelt (670705),

Patrik Molnar (694214), Sára Anna Szabó (690442), Manuela Skov Thomasen (650504)

4/11/2022

Knitted script for assignment 2 at page 65 to 97

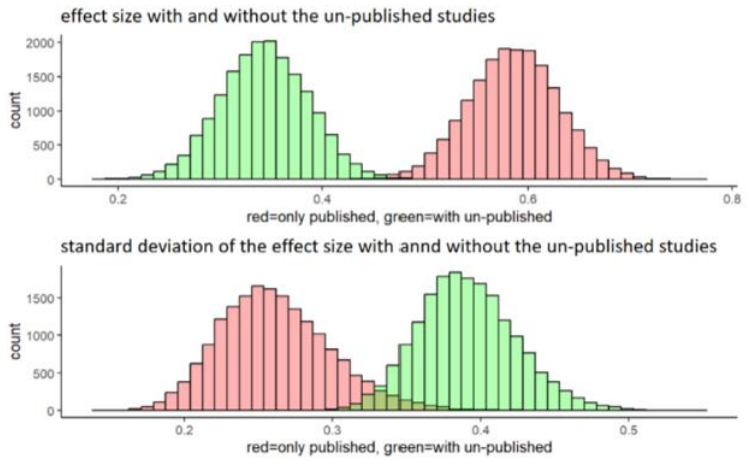
Q1

```
## Population-Level Effects:
##           Estimate Est.Error l-95% CI u-95% CI Rhat Bulk_ESS Tail_ESS
## Intercept     0.34      0.04     0.26     0.43 1.00     4104     7005
```

(DK) In our simulation of the publication bias, we find the estimates of the intercept on *population level* to be 0.34 when unpublished studies are included and 0.59 mẃhen they are excluded. The 95% confidence interval for estimation of the intercept for the dataset containing both published and unpublished studies is (0.26 : 0.43), whereas the confidence interval for the same estimate for the data set containing only the published studies is (0.50 : 0.67).

Furthermore, the sd (standard deviation) of the intercept for data including unpublished studies is 0.43, and 0.39 for data only including published.

Plot 8



(DK) Our simulation of the true underlying effect is centered around 0.4. This means that the publication bias results in a tendency to overestimate the underlying signal. We simulate the standard deviation of the effect size to be 0.4, and a measurement error of 0.8. Therefore we can show that the publication bias also results in an underestimation of variance/deviation.

(DK) We can see this effect in plot 8, the red (only published) has a higher estimate for the intercept (effect size) and lower standard deviation, whereas the green (all studies) has a lower estimated intercept and higher standard deviation.

(DK) This conclusion is based on our findings from the simulation. However, the simulation has its flaws. It would be better to run the simulation many times, record how much the publication bias influences the estimates, and then estimate the publication bias' effect from that. Furthermore, our process of determining the chance of publication is very simple; one could imagine that many other factors, than the effect size and standard error, plays a role. Such as reputation of the author(s), experimental design, and factors which are not as easily quantifiable. With all this in mind, we should not draw strong conclusions from our simulation, but rather keep it in mind as we evaluate the results of the meta analysis.

Q2

What is the current evidence for distinctive vocal patterns in schizophrenia?

(DK, SS) After filtering NAs, the following demographic values - grouped by patients with schizophrenia and healthy controls – the dataset contained 48 studies with the demographics described in tibble 2.

Tibble 2

	M sample size	M nr. Of males	M nr. Of females	Mean age	mean_sd_age
Schizophrenia	40.5	27.3	14.3	35.9	8.39
Healthy control	31.2	17.7	14.7	34.9	8.92

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

(SS) In this analysis we are using the function escalc(). The function calculates *Hedges' g,* the standardized mean difference between two groups. The value represents the effect size, and is similar to *Cohen's d*.

The formula for g is = (x1 – x2) / √((n1-1)*s12 + (n2-1)*s22) / (n1+n2-2)

$$g = (x_1 - x_2) : \sqrt{\frac{(n_1-1)*{s_1}^2+(n_2-1)*{s_2}^2}{n_1+n_2-2}}$$

and

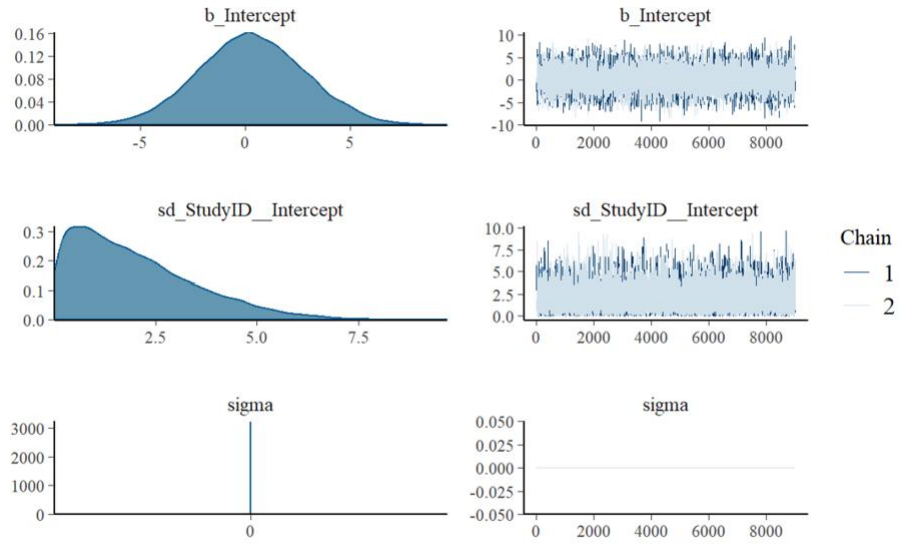
the formula for d is = (x1 – x2) / √(s12 + s22) / 2

$$d = (x_1 - x_2) : \sqrt{\frac{s_1^2+s_2^2}{2}}$$

(SS) Hedges' g takes the sample size of each group into account, and g = d when the two sample sizes are equal. Therefore we have chosen to use *Hedges' g* to calculate the effect size of the different studies.

(BR) Our model gives us an estimated intercept (effect size) of 0.3 with an estimated standard deviation of 2, which is due to the effect sizes of the studies varying a lot.

Plot 9

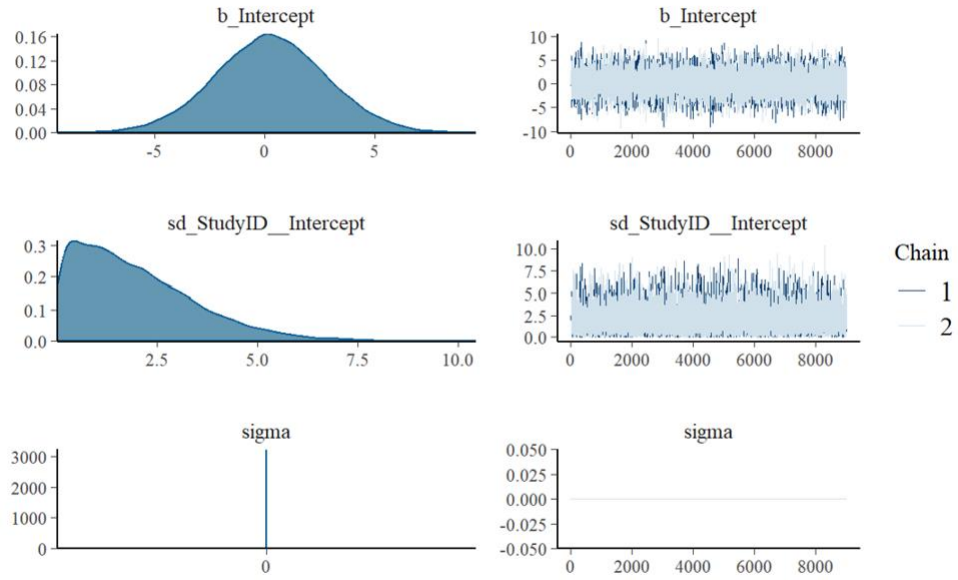


Note: model with priors and empirical data

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

(BR) As for influential studies, for example, the study by Cohen et al. (2014), with an effect size of -3.30 (sd=0.90, se=0.11, n=76) was taken into consideration, therefore we have run our model fit again. Here we have excluded the study by Cohen, to see how influential it is on the estimates.

Plot 10



Note: After removing Cohen et al. (2014)

(BR) When we compare the population-level estimate of the effect size, "Intercept" in the summary, between the estimates including the Cohen et al. study and the estimates excluding the study, we see, quite surprisingly, that there is very little difference. The estimated error remains the same, at 2.5, and the estimated effect size differs by 0.04 points from 0.29 to 0.33.

(BR, PM, MST) Publication bias is the tendency in scientific literature for studies with positive results to be published more often than those with negative results. For research investigating correlations between vocal patterns and schizophrenia diagnoses, publication bias may manifest, as a greater number of conclusions finding significance between these variables compared to studies finding potentially significant reasons to be critical of the hypothesis. This bias can lead to a skewed representation of the research data, resulting in an overestimation of the role of vocal patterns in the diagnosis of schizophrenia, and consequently, inaccurate or ineffective treatment of patients and utilization of healthcare resources.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Portfolio Exam - Part 3 | Methods 3 F2022, CogSci @ AU

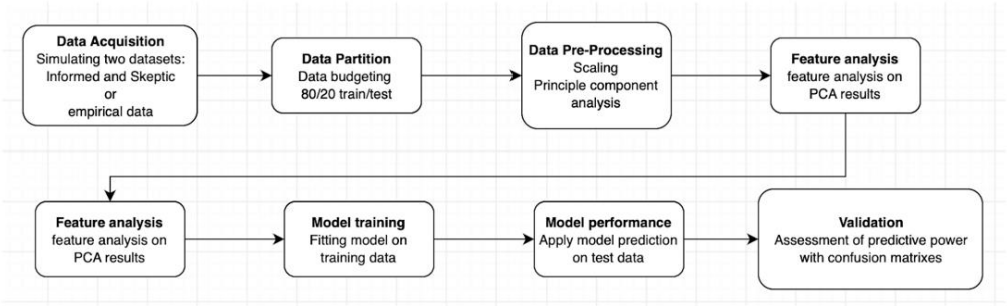
Group 3 – Ditlev Kræn Andersen (699440), Bryan Roemelt (670705),

Patrik Molnar (694214), Sára Anna Szabó (690442), Manuela Skov Thomasen (650504)

30/11/2022

Knitted script for assignment 3 at page 98 to 152

Diagram 1



Note: Diagram of machine learning pipeline (PM, MST)

(PM, MST) Diagram 1 shows the steps in the machine learning pipeline that is described further in the following section.

Data budgeting: Based on the standard practice and information presented in the lectures, we decided to split the data in 80/20 ratio.

Data preprocessing: We removed demographic data, except gender, from the data set. The data was split using initial_split() from rsample version 1.1.0. We specify we want a balanced ratio of men and women, using the argument strata = Gender. We scale the variables using the juice() and bake() functions from recipes version 1.0.3. By using the juice() and bake() functions we can define our recipe for scaling and make sure there is no leakage between the two datasets. After we split the data we run a principle component analysis using recipes version 1.0.3. We specify in the code that we want 5 principle components, from the results of the principal components analysis. A feature analysis is run to determine which variables to use when building our model.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

(BR) Model choice and training: Our model is built based on the results of the feature analysis. We expect to include a random intercept in the model to account for individual differences. The model is then trained on only the training data set and made predictions of diagnosis on each trial.

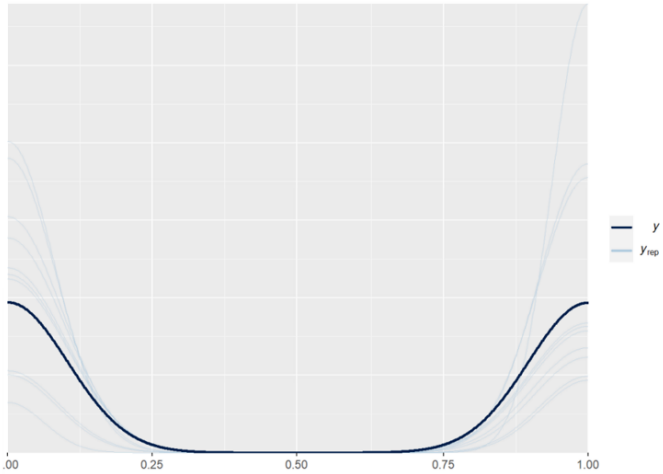
(BR) Assessment of performance: To assess the performance we ran the model on the test data and did a confusion matrix to determine how precise the model was in its predictions.

Briefly justify and describe your use of simulated data, and results from the pipeline on them.

(DK, PM) The data simulation was set to create a population of 100 pairs of schizophrenia and control and simulates 10 repeated measures with the informed and skeptical effect mean and standard deviation.

(DK, PM) The informed effect mean is set based on the literature, to a vector of 10 values. In 6 of them the values are ranging from -0.5 to 1.89, and in 4 of them, the values are set to 0 to account for random noise in measurement. While the Skeptic effect mean is set to 0 for all 10 trials. The individual variability from the population is also defined with Informed Standard Deviation set to 1 and Skeptical Standard Deviation is set to 0.5, while measurement error is set to 0.2.

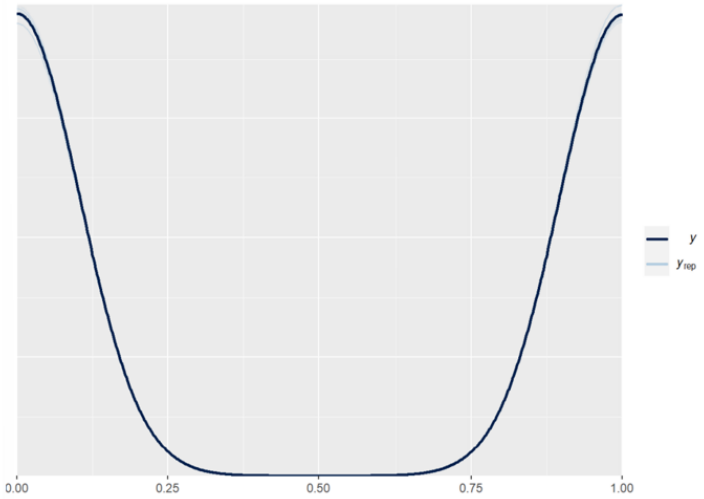
Plot 11



Note: pp-check on model run on only priors

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Plot 12



Note: pp-check on model fit on the informed simulated data

(MST, PM) Based on the plots 11 and 12, we can see that the model learnt when not only based on priors, but the simulated data.

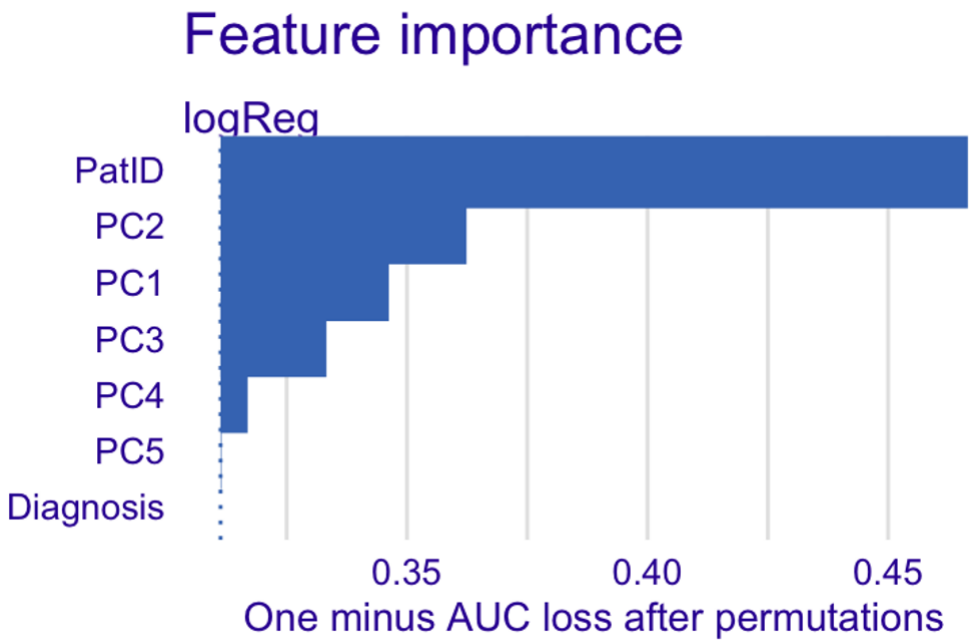
(DK, PM) The results from the confusion matrices on both training and test sets for both conditions showed that the models on simulated data are performing well in predicting the condition. The models have a high accuracy and wide confidence intervals, above 95%. The models also have a high sensitivity and Positive Predictive Value, above 0.95. This is consistent with the information from the data simulation. The results indicate that the models are accurately predicting the condition and have low error.

Describe results from applying the ML pipeline to the empirical data and what we can learn from them.

(DK, PM, MST) The data budgeting resulted in a training dataset of 1510 observations and a test data set of 379 observations. As part of the data preprocessing we ran the feature analysis on the results of principal component analysis. The results shown in plot 13 tell us that PC2, PC1, PC3 and PC4 are the most important features found in the principle component analysis. PatID is the most important feature.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

Plot 13



(DK, PM, MST) Based on the feature analysis we defined a model with the following formula:

$$Diagnosis \sim 1 + PC2 + PC1 + PC3 + PC4 + (1|PatID)$$

We allowed for random intercepts to account for individual variability.

(PM) The performance of the model is decent. The Group-level estimate values are within the expected range, considering the informed Standard Deviation and Skeptical Standard Deviation values of 1 and 0.5. The Estimate values for the Population-Level Effects are also close to the informed and skeptical effect mean values reported in the data simulation. In addition, the Rhat values for each parameter are close to 1, indicating that the model has converged. The Bulk_ESS and Tail_ESS are also high, indicating that the model is performing well.

Ditlev Kræn Andersen
202106495
Au699440
Study group 3

(DK) To the right we see the results from our ML algorithm's predictions on the test data, and the results on the training data to the right. We see the model predicts roughly at chance level, meaning the model either does not pick up on the signal in the data, or there simply is no signal in the data. In our simulated data set the model picked up on the signal given almost the same model-fit, fixed effects plus random slope. The only difference is the fact that we ran a principal component analysis. The PCA makes new variables, composed of the original data, this might be the cause of the poor performance of the model. If we wanted to test this, we could run a feature analysis on the "raw" data, and build our model from that, and see if the model then picks up on a signal. If it does not, we might be inclined to conclude that there is no signal for the model to pick up on.

```
Confusion Matrix and Statistics

          Reference
Prediction  CT SCZ
       CT  115  78
       SCZ 104  82

               Accuracy : 0.5198
```

```
Confusion Matrix and Statistics

          Reference
Prediction  CT SCZ
       CT  545 251
       SCZ 444 270

               Accuracy : 0.5397
```