Data Analysis Project

for Bayesian Statistics

Osmanagic Selma (12129484), Jakub Kotala (12305828)
GitHub Repo

January 23, 2024

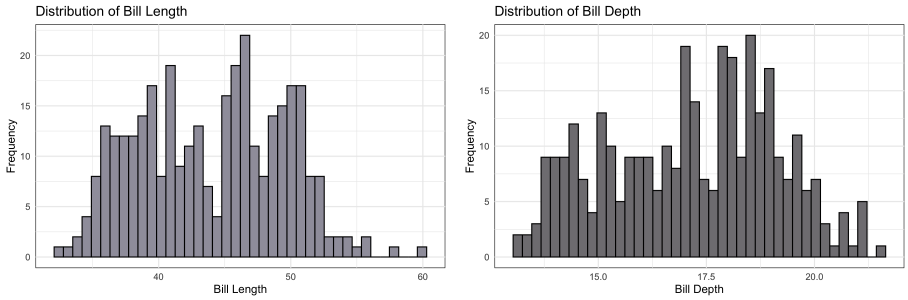
Contents

1 Introduction

This project aims to conduct Bayesian data analysis on the famous penguin dataset, which centers around three penguin species: Chinstrap, Gentoo, and Adélie. Our primary objective is to develop a Bayesian model that can accurately predict the bill length of these three species of penguins.

2 Data exploration

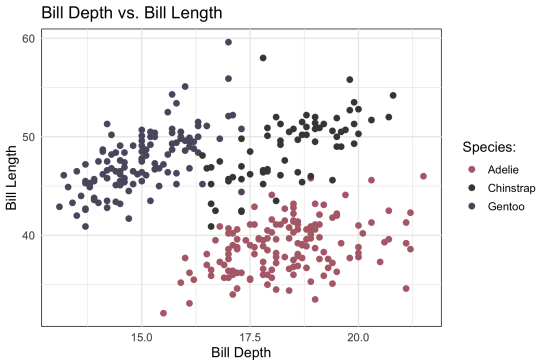
The dataset consists of 333 samples and 4 variables, of which two are categorical (*species* and *sex*) and two are numerical (*bill length* and *bill depth*). Distributions of said numerical variables are visible on the figures bellow: 1a and 1b.



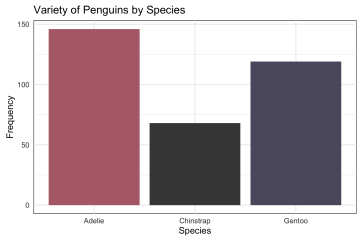
(a) Distribution of bill length

(b) Distribution of bill depth

The categorical variable *species* consists of three categories (Adelie, Chinstrap and Gentoo), which can be observed in the figure 2b. These three categories are not equally represented in the dataset, which makes the dataset imbalanced. The scatter plot 2a shows three clearly distinguishable clusters. It can be concluded that the Adelie penguins have deeper and shorter bills, Chinstrap deeper and longer bills and Gentoo shallower and longer bills.



(a) Scatter plot of bill depth vs. bill length



(b) Variety of species

3 Modelling

During the modelling phase, two basic models were implemented; one falls into the category of pooled models, and the other one is hierarchical. As response, we opted for the *bill length*, and as a predictor for the pooled model *bill depth* variable was chosen. In addition to the *bill depth*, the hierarchical model made use of the *species* variable, as well. In the further sections the models were presented in a detailed manner.

3.1 Model 1 - Pooled Model

In our first model, the likelihood is modeled using a normal distribution, with the expected value in the form of linear equation. The standard deviation is modeled simply by the parameter σ. This model doesn´t have a hiearchical structure and doesn´t distinguish between penguins from individual groups based on variable *species* or *sex*.

$$Expected\ \ Value = \alpha + \beta \cdot bill\_depth[n] \tag{1}$$

$$Standard\ \ Deviation = \sigma \tag{2}$$

This simple model has three parameters: α (intercept), β (slope) and σ (standard deviation). Priors for α and β were selected using classical regression model. We performed linear regression and got summary of this model. From summary we conclude that estimated value of intercept equals approximately 55 and its estimated standard deviation is approximately 2.5. We used this values for setting prior for α in our Bayesian pooled model. The prior for β was selected identically.

Parameter	Model 1	Model 2
α	$\mathcal{N}(10, 1)$	$\mathcal{N}(55, 2.5)$
β	$\mathcal{N}(2, 1)$	$\mathcal{N}(-0.63, 0.15)$
σ	$\mathcal{C}(0, 1)$	$\mathcal{C}(0, 5)$

Table 1: Priors for the pooled model

3.2 Model 2 - Hierarchical Model

In our hierarchical model, the likelihood is modeled using a normal distribution, with the expected value in the form of linear equation. The standard deviation is modeled simply by the parameter σ.

$$Expected\ \ Value = \alpha + \beta \cdot bill\_depth[n] + \gamma[species[n]] \tag{3}$$

$$Standard\ \ Deviation = \sigma \tag{4}$$

The α parameter presented in the equation 3 represents the intercept, while β represents the slope term associated with the *bill depth*. The hierarchical term γ is dependant on the *species* to which the instance belongs. Since this part is *species*-specific, our model gets a hierarchical structure.

As for the priors, γ is normally distributed with the mean of 0 and the standard deviation, that is itself drawn from a Cauchy distribution with the scale of 5. α and β are also drawn from a normal distribution, while σ is drawn from a Cauchy distribution with the scale of 5 as well. We opted for the Cauchy priors as they are often used for scale parameters, because of its heavy tails. As we don't expect the standard deviation to be extremely large, we opted for a smaller scale value (in this case, 5).

$$\begin{aligned} \alpha &\sim \mathcal{N}(\mu_1, \sigma_1) \\ \beta &\sim \mathcal{N}(\mu_2, \sigma_2) \\ \gamma &\sim \mathcal{N}(\mu_3, \sigma_3) \\ \sigma_{\text{species}} &\sim \mathcal{C}(x_1, \lambda_1) \\ \sigma &\sim \mathcal{C}(x_2, \lambda_2) \end{aligned}$$

In our experiments, we put to use several versions of weakly informative priors, to minimize the impact of priors to the posterior. Some parameters have specific priors, that we derived from the linear model. The same model was run with the different values for means and standard deviations. In addition to these, we implemented the same model with flat priors, out of sheer curiosity. The values and priors we used can be observed in the table 2.

Parameter	Model 1	Model 2	Model 3
α	$\mathcal{N}(0, 10)$	$\mathcal{N}(54.89, 2.56)$	$\mathcal{N}(43, 5.5)$
β	$\mathcal{N}(0, 10)$	$\mathcal{N}(-0.63, 0.15)$	$\mathcal{N}(17.2, 1.97)$
γ	$\mathcal{N}(0, 10)$	$\mathcal{N}(0, 10)$	$\mathcal{N}(0, 7)$
σ_{species}	$\mathcal{C}(0, 5)$	$\mathcal{C}(0, 5)$	$\mathcal{C}(0, 3)$
σ	$\mathcal{C}(0, 5)$	$\mathcal{C}(0, 5)$	$\mathcal{C}(0, 3)$

Table 2: Priors for the hierarchical model

4 Model checking

For both models, sampling was performed with 4 chains and 4000 iterations per chain. The burn-in period accounts for 50% of the iterations, resulting in a total of 16000 iterations. After discarding the burn-in samples, we retained the remaining 8000 samples for analysis. To assess the convergence of the samples, we computed statistics for each model. Of particular interest were the $\hat{R}$ values, where $\hat{R} < 1.1$ serves as an indicator of convergence. Additionally, we examined density plots to assess the fit of the density function created from the sampled data (includes first 50 instances). The assure that the convergence was done properly, we also assessed the trace plots, which were added to the appendix, but not the report itself.

4.1 Model 1 - Pooled Model

The table 3 with mean and $\hat{R}$ values can be observed bellow. Since $\hat{R}$ values for all parameters and models are lower than 1.1, it can be concluded that the model converged properly for every parameter.

Parameters	Mean M1	Rhat M1	Mean M2	Rhat M2
alpha	13.543134	1.000312	54.9086316	1.002225
beta	1.742197	1.000132	-0.6359274	1.002061
sigma	7.118982	1.001125	5.3418854	1.000264

Figure 3: Table with the means and $\hat{R}$ values for the pooled model

The density plot bellow 4 shows that the density plot of pooled model V1 looks completely different compared to the density plot of the pooled model V2, which can be ascribed to the use of different priors.

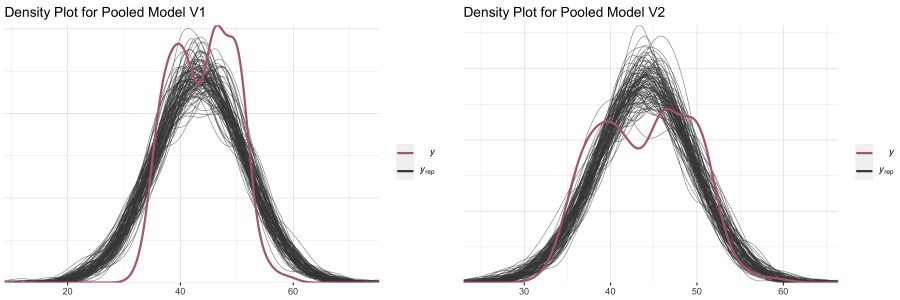


Figure 4: Density plots for 2 versions of the pooled model

4.2 Model 2 - Hierarchical Model

Below, you'll find Table 5, displaying mean and $\hat{R}$ values. The values of $\hat{R}$ for all parameters and models consistently fall below 1.1, indicating convergence of all models for all parameters.

Parameters	Mean M1	Rhat M1	Mean M2	Rhat M2	Mean M3	Rhat M3	Mean FP	Rhat FP
alpha	17.969862	1.0027858	50.7389504	1.003147	29.067953	1.002485	28.890741	1.0013633
beta	1.442136	0.9997070	0.3703851	1.001561	1.334179	1.000047	1.334661	1.0005715
gamma[1]	-5.593426	1.0034092	-18.6960670	1.003574	-14.707576	1.003129	-14.538616	1.0022660
gamma[2]	4.286108	1.0035129	-8.7071019	1.003370	-4.804527	1.003108	-4.638526	1.0024178
gamma[3]	7.958293	1.0035697	-8.7118898	1.003602	-1.504848	1.003240	-1.338000	1.0022703
sigma_species	8.202911	1.0001460	14.7944215	1.002311	10.837993	1.000941	10.587190	1.0016946
sigma	2.533087	0.9998823	2.7838132	1.000849	2.532361	1.000195	2.531504	0.9999211

Figure 5: Table with the means and $\hat{R}$ values for the hierarchical model

The density plots generated from the sampled data exhibit a consistent pattern, closely aligning with the expected true distribution. All plots look visually very similar, suggesting that different priors didn't have a notable effect on the posterior.

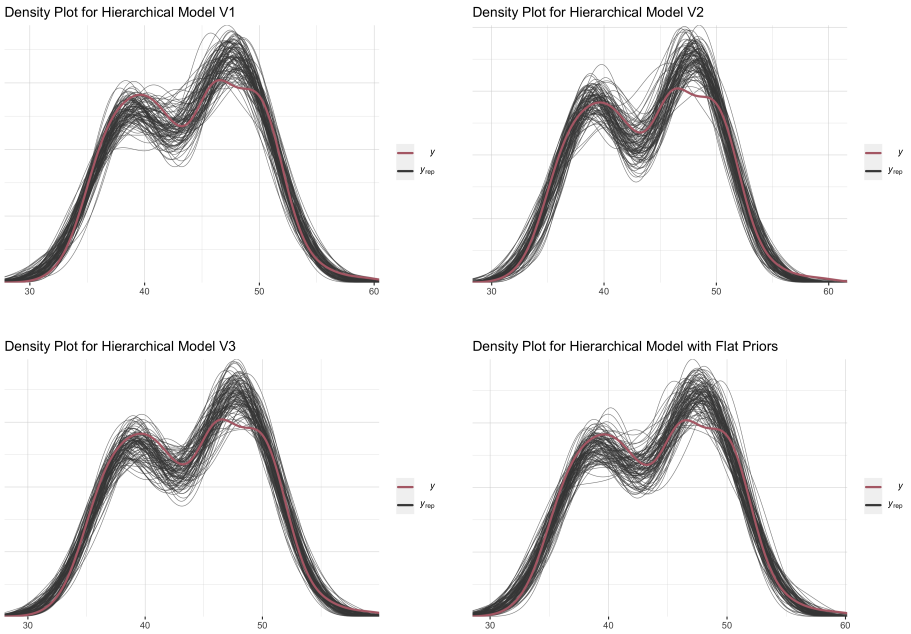


Figure 6: Density plots for 4 versions of the hierarchal model

5 Model comparison

For the purpose of model comparison, as suggested to us, the functions of *loo* package were put to use. To examine the quality of models and their predictive accuracy, leave-one-out cross-validation was performed, and the ELPD (Expexted Log Pointwise Predictive Density) as well as the SE (Standard Error) were computed. Additionally, the models were compared among each other using the *loo_compare* function. As an indicator of the reliability of leave-one-out cross-validation samples, the Pareto k-values were checked. For all models, they were sufficiently small, which lead us to believe that the tails of importance sampling distribution were not heavy and that, consequently, the comparison bellow can be trusted.

5.1 Model 1

Model V2 outperformed model V1 significantly, which is expected. The reason behind such discrepancy is the inclusion of the priors determined based on the linear regression parameters for the model V2. The parameters of the model V1 were assigned somewhat arbitrarily.

model	elpd_diff	se_diff
pooled model V2	0.0	0.0
pooled model V1	-95.8	11.4

Table 3: ELPD Differences and Standard Errors for pooled models

5.2 Model 2

Observing Table 4, it is evident that models V1, V3, and FP demonstrate similar performance, whereas the performance of V2 is substantially worse. Model V2 was based on the parameters that were obtained from the linear regression model. Since the linear model does not include any differentiation of the parameters based on the *species* variable, it is expected that these priors are not suitable for the hierarchical model.

model	elpd_diff	se_diff
hierar. model V1	0.0	0.0
hierar. model V3	0.0	1.2
hierar. model FP	-0.1	1.2
hierar. model V2	-31.3	8.8

Table 4: ELPD Differences and Standard Errors for hierarchical models

5.3 Comparison of all the version of model 1 and model 2

The performance for the hierarchical models was substantially better compared to the pooled models, which can be concluded by observing the ELPD values from the table 5. This is expected behavior, as the hierarchical model is more complex and takes as an predictor another variable.

model	elpd_diff	se_diff
hierar. model V1	0.0	0.0
hierar. model V3	0.0	1.2
hierar. model FP	-0.1	1.2
hierar. model V2	-31.3	8.8
pooled model V2	-246.8	15.6
pooled model V1	-342.6	14.4

Table 5: ELPD Differences and Standard Errors for pooled and hierarchical models

6 Sensitivity analysis for pooled and hierarchical models

As we are interested in the robustness of our models to change in priors, we conducted a small sensitivity analysis. Completely different priors were chosen for the aforementioned models, and the convergence as well as the ELPD values were checked again. The new set of priors for both the pooled model and the hierarchical model can be observed in the table 7a and 7b respectively.

Parameter	Priors
α	$\mathcal{U}(-10, 10)$
β	$\mathcal{C}(0, 5)$
σ	$\mathcal{N}(0, 4)$

(a) Pooled model

Parameter	Priors
α	$U(-10, 10)$
β	$U(-10, 10)$
γ	$T(3, 0, 1)$
σ_{species}	$LN(0, 1)$
σ	$LN(0, 1)$

(b) Hierarchical model

Interestingly, the plots obtained from new priors are similar to the previous plots.

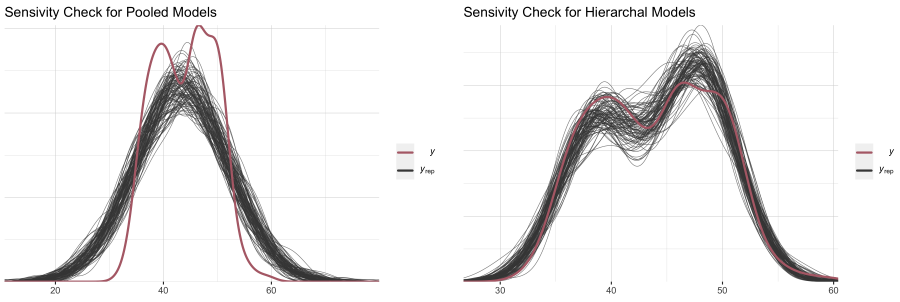


Figure 8: Density plots for sensitivity analysis

In the table bellow 6, it is interesting to spot the trend of hierarchical models performing better persisting, even though the models obtained from the sensitivity analysis are taken into account. The hierar. sensitive model still performed better than its counterpart hierar. model V2.

model	elpd_diff	se_diff
hierar. model V1	0.0	0.0
hierar. model V3	-0.1	1.1
hierar. model FP	-0.3	1.1
hierar. model sens	-0.8	1.8
hierar. model V2	-31.3	8.7
pooled model V2	-246.8	15.6
pooled model V1	-343.0	14.4
pooled model sens	-356.2	14.6

Table 6: Comparison of all the models based on ELPD and SE

7 Discussion of issues and potential improvements

Throughout the project, we did not encounter any significant issues. However, we acknowledge the possibility of making mistakes.

If time permitted, we would have gladly implemented another hierarchical model that would not only be dependant on the *species* variable, but on the *sex* variable as well. This could serve as an idea for the possible future work.

8 Conclusion

In our prior courses, we learned the basics of data analysis using the traditional frequentist approach. However, in this course, we had the opportunity to delve into a field completely new to us. Throughout the project, we applied our newly obtained theoretical knowledge to real-world penguin data. Among many insights we gained while working on this project, the one that stood out the most for us are the hierarchical models and their superiority to other models we implemented. During this interesting project we collaborated and supported each other through the whole process, and delivered a roughly equal amount of work.

9 AI disclosure

We used AI, more precisely chatbot Chat GPT, in few occasions during our collaboration on this project. AI proved to be extremely helpful when it came to helping us with latex code. During our work, there were moments, when we were not sure about theoretical or/and programming concepts. However, during these moments, AI didn't provide us with anything useful very often, so we relied on our knowledge and the code/slides provided in TUWEL, as well as the official documents of the libraries we used (i.e. loo, knitr, rstan, ...)

10 Appendix

10.1 Trace Plots of Pooled Models

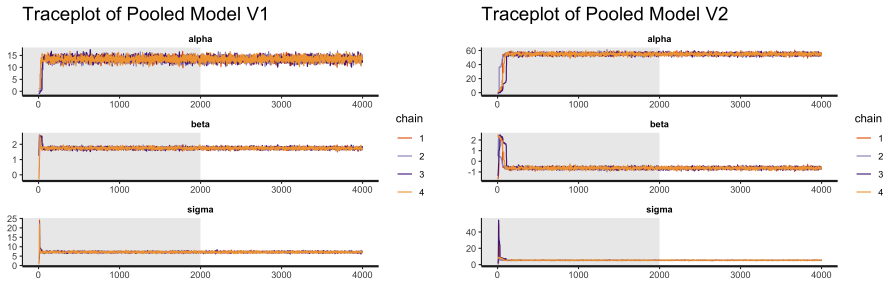


Figure 9: Trace plots of all variations of pooled models

10.2 Trace Plots of Hierarchical Models

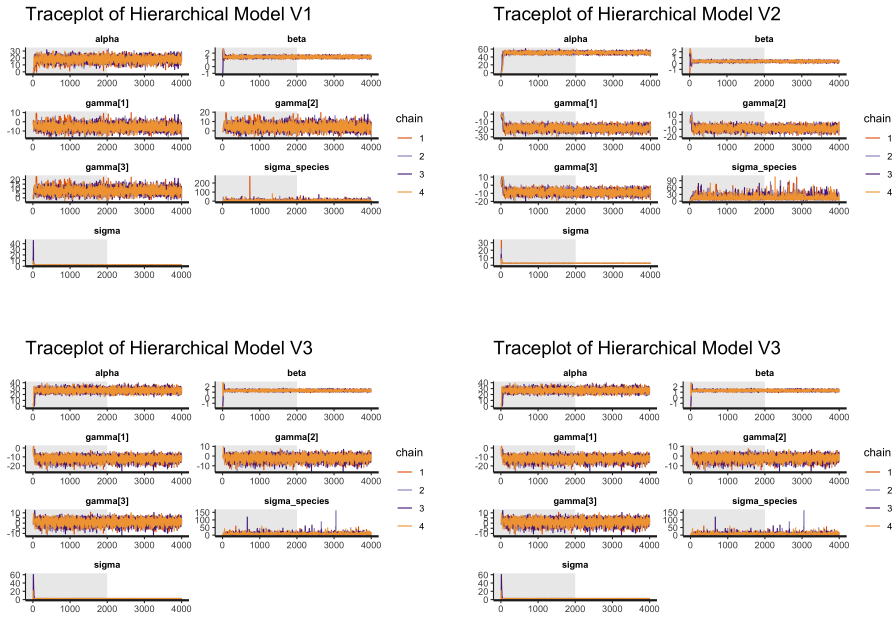


Figure 10: Trace plots of all variations of hierarchical models

10.3 Stan Plots of Pooled Models

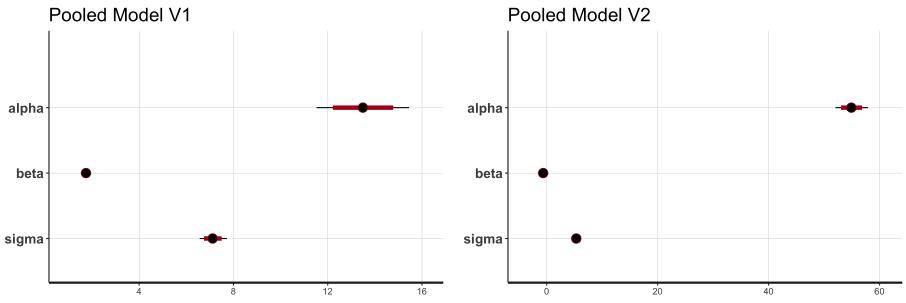


Figure 11: Stan plots of all variations of pooled models

10.4 Stan Plots of Hierarchical Models

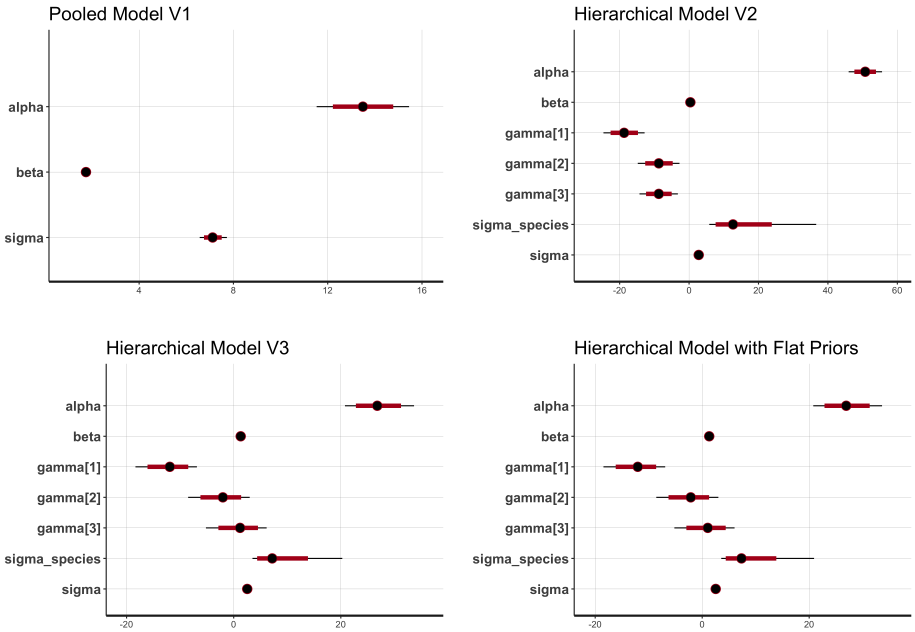


Figure 12: Stan plots of all variations of hierarcical models

10.5 PPC Histograms with Means of Pooled Models

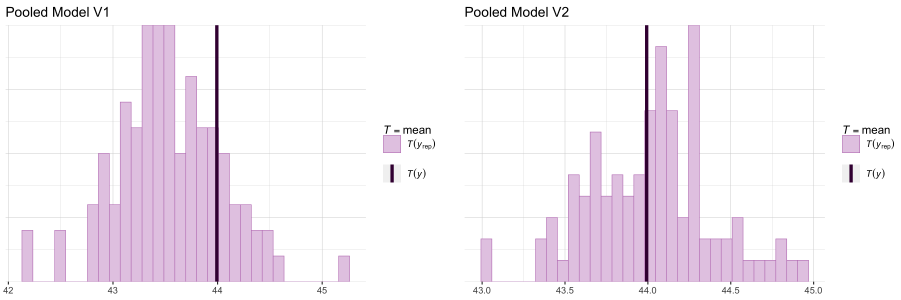


Figure 13: Histograms with means for all variations of pooled models

10.6 PPC Histograms with Means of Hierarchical Models

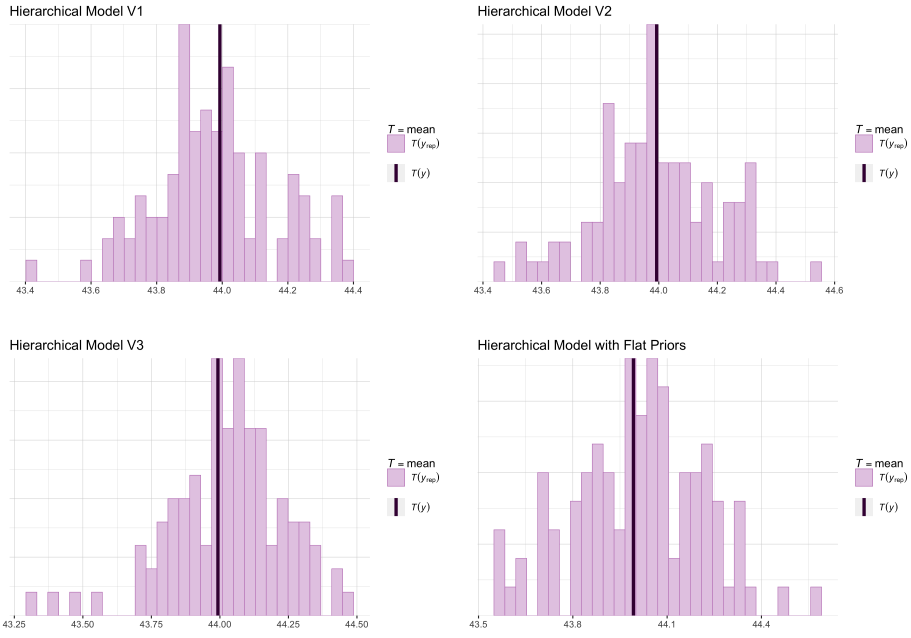


Figure 14: Histograms with means for all vairations of hierarchical models

10.7 ECDF of the Observed data vs. ECDF of simulated data for all pooled models

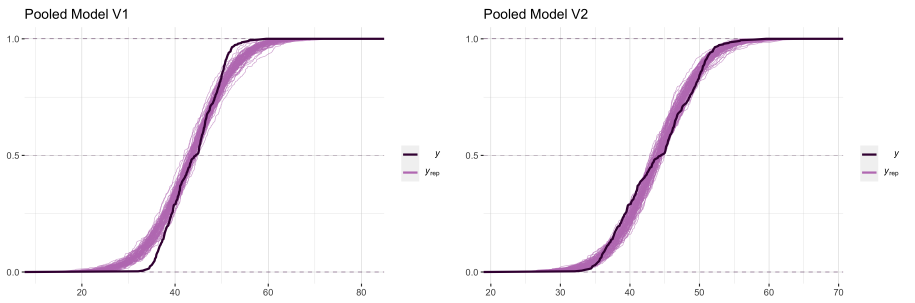


Figure 15: ECDF plots for pooled models

10.8 ECDF of the Observed data vs. ECDF of simulated data for all hierarchical models

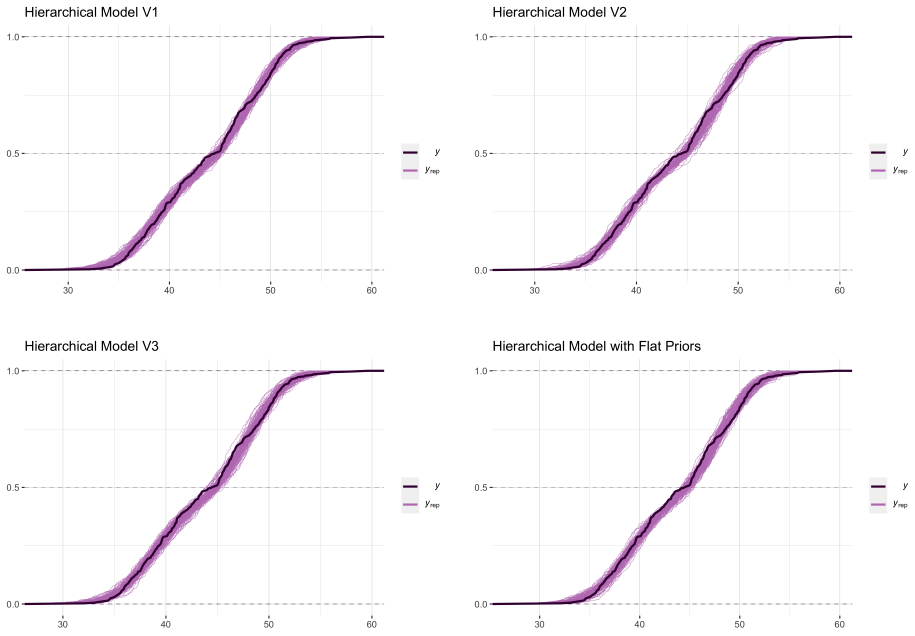


Figure 16: ECDF plots for hierarchical models