Institute of Statistics and Mathematical Methods in Economics
TU Wien

# Gaussian Linear models on Antarctic Penguins

Barbaro Francesca
Carminati Roberto
Rondini Davide

24/01/2024
*Bayesian Statistic (WS 2023/24)*

Institute of Statistics and Mathematical Methods in Economics
TU Wien

# Contents

# 1 Introduction

In this project, our focus was on the *Antarctic Penguins* dataset, which contains information collected by Gorman et al. (2014). The dataset has four covariates providing insights into these fascinating polar inhabitants:

- **Species**: A categorical variable classifying penguins into distinct species, namely *Chinstrap*, *Gentoo*, and *Adélie*.
- **Sex**: A categorical variable indicating the gender of the penguins.
- **Bill Depth**: A numerical variable describing the depth of the penguins' bills.
- **Bill Length**: The target variable, representing the length of the bills for each penguin.

Our primary goal is to construct a Bayesian model capable of predicting the bill length of these birds. The selected models are Gaussian Linear Models, denoted by:

$$y \sim N(\gamma + \boldsymbol{\beta} \cdot X, \sigma^2) \tag{1}$$

In this equation, $y$ symbolizes the target variable under consideration, $\gamma$ represents the intercept, $\boldsymbol{\beta}$ is the vector of regression coefficients, and $X$ denotes the matrix of covariates.
We explored three distinct types of model: the *Pooled*, *Hierarchical*, and *Separate* models.

The chosen programming language for our analysis was R, complemented by Stan, a probabilistic programming framework. Our complete code is provided on GitHub (see Appendix A).

Delving deeper into the Bayesian framework, the aim is inferring the posterior distribution. We start with a *prior distribution* $p(\gamma, \boldsymbol{\beta})$, representing our beliefs about the parameters $(\gamma, \boldsymbol{\beta})$ before observing any data. Coupled with this is the *likelihood* $p(\mathbf{y}|(\gamma, \boldsymbol{\beta})$, expressing the probability of the observed data $\mathbf{y}$ given the parameters $(\gamma, \boldsymbol{\beta})$. The combination of the prior and likelihood yields the *posterior distribution* $p(\gamma, \boldsymbol{\beta}|\mathbf{y})$, a refined and updated understanding of the parameters considering the observed data.

$$p(\gamma, \boldsymbol{\beta}|\mathbf{y}) \propto p(\gamma, \boldsymbol{\beta}) p(\mathbf{y}|\gamma, \boldsymbol{\beta}) \tag{2}$$

The inference process, executed through algorithms like Markov Chain Monte Carlo (MCMC) implemented in Stan, allows us to explore and draw meaningful conclusions from this posterior distribution.

# 2 Data Exploration

We started out work conducting an exploratory analysis on the *Antarctic Penguins* dataset.

Firstly we analysed the data visualising the Scatterplots, in Figure ( 1), where the different colors shows the species and the sex of the penguins. It's possible to observe that the species are divided into three quite distinguishable clusters. On the other hand, the division between sexes is less pronounced, but it is still possible to notice that males generally have a longer and wider bill compared to females of the respective species. We expect that both this two distinctions will be relevant in the construction of our linear models.
In Figure ( 1), it is evident that without considering the distinction between species, there is no apparent positive correlation between the variables Bill Length and Bill Depth, as one might expect a priori. In fact, the calculated correlation is -0.22. However, upon considering the differentiation between species, a positive correlation becomes visibly apparent in the plot. Specifically, the calculated correlations for each species are as follows:

| Species | Correlation of Bill Length and Depth |
|---|---|
| Adélie | 0.386 |
| Chinstrap | 0.654 |
| Gentoo | 0.654 |

This table illustrates that when accounting for species distinctions, a positive correlation between Bill Length and Bill Depth becomes evident, with stronger correlations for the *Chinstrap* and *Gentoo* species.
This observation reinforces the significance of dividing the data by species, as it suggests considerable dissimilarity among the species.

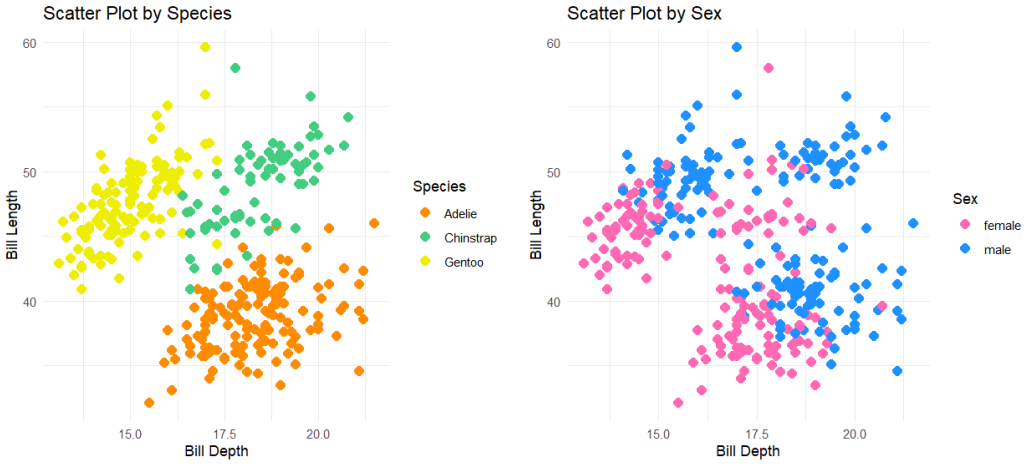



Figure 1: Scatterplot of the dataset were colors enphatise the different covariates

Afterward, we analyzed the distribution of the target variable Bill Length through Boxplots, in Figure ( 2), segmented by both species and sex. As anticipated earlier, we observe that males have a longer bill compared to females of the same species. Additionally, we notice a striking similarity between the *Gentoo* and *Chinstrap* species. This observation can be valuable in the creation of models, where we will attempt to account for this fact through hierarchy.

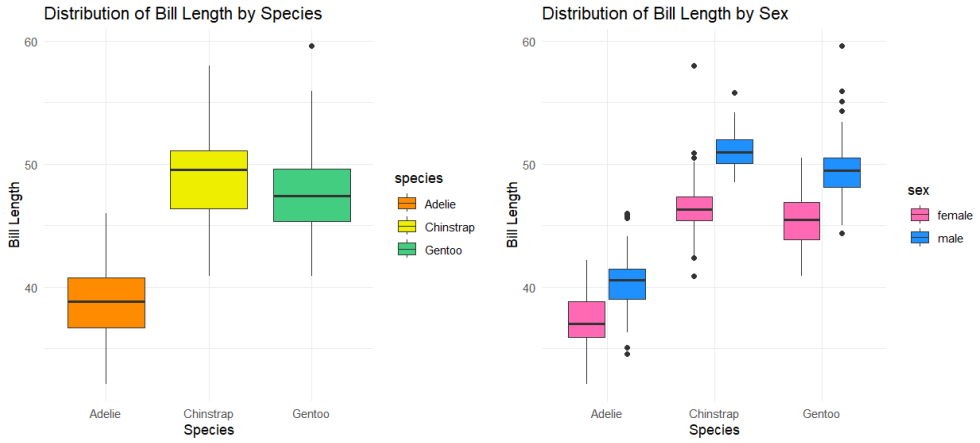



Figure 2: Boxplot of Bill Length by species and by sex

Afterward, we delved into the distribution of the target variable, plotting it in the graphs of figure ( 3). We observe that 'Bill Length' exhibits a multimodal distribution, with each mode corresponding to a distinct species. Upon further division of the variable, both by species, we notice that *Adélie* is distinct from the others, while the remaining species show considerable overlap, also displaying multimodal distributions—presumably one mode for each sex. The final plot illustrates the distribution by sex, where once again, we observe a clear distinction for the first species.

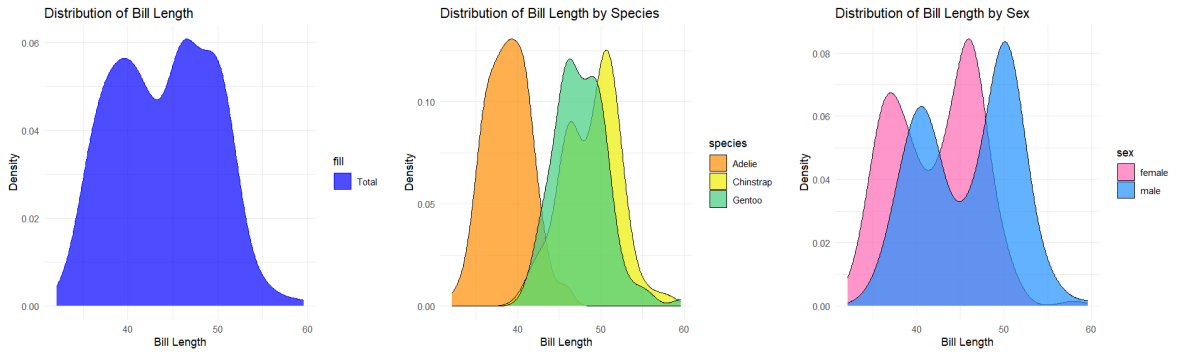



Figure 3: Distribution of Bill Length by species and by sex

As illustrated in Figure ( 4), we examined the normality of the target variable distribution. Utilizing a QQ-plot, we observed that the distributions, when divided by species, closely approximate normality.

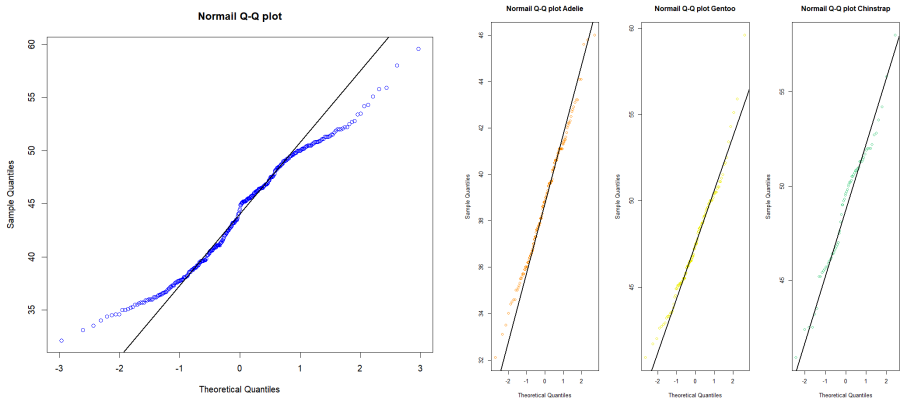



Figure 4: Normal QQplot of Bill Length by species

Concluding the exploratory analysis, our findings emphasize the importance of splitting the data by species and by sex to reveal nuanced patterns and relationships within the Antarctic penguin dataset.

# 3 Modelling

With the exploratory analysis providing valuable insights into the Antarctic Penguins dataset, our focus now shifts to constructing Bayesian models for predicting the bill length of these polar birds. The types of models we considered are the *Pooled*, *Separate*, and *Hierarchical* models:

- The *Pooled model* assumes a common set of parameters for the entire dataset, treating all observations as part of a unified population. By pooling information across all instances, this model provides a simplified and generalized representation of the underlying relationship between covariates and the target variable.

- the *Separate model* assigns unique parameters to different subsets or categories within the dataset. Each subset is treated independently, allowing for distinct relationships between covariates and the target variable within each category.

- The *Hierarchical model* balances the Pooled and Separate models by introducing a hierarchical structure. It assumes that parameters for each subset are drawn from a common distribution, facilitating the sharing of information across subsets while still allowing for individual characteristics.

### 3.1 Workflow

In this paragraph, we present in a general manner the models we have created, having a broad understanding of the data we need to analyze.

We began by crafting a ***Pooled*** model without distinguishing the data, treating them as a homogeneous group. As expected, this model did not perform well, given the distinct characteristics observed among species during the preliminary analysis.

As a second step, we formulated a ***Separate*** model. In this approach, we crafted three distinct models, assigns unique parameters for each species independently —contrary to the aggregated nature of the *Pooled* model. In this model we didn't considerate the variable Sex assuming the individuals are grouped only by their species.

Afterwards we specialised this model also including the variable Sex as a covariate. Since this variable has binary values the outcome of this model will consist in six different regression lines. This model will be thoroughly discussed in Section (3.2). The outcomes of this model not only validated our initial understanding of the data but also demonstrated its efficacy, highlighting the significant impact of species-specific differences on the observed patterns.

Next, we attempted to generalize these insights by constructing a ***Hierarchical*** model, as outlined in Section (3.3). In this model, while retaining the division by species, we assumed a shared distribution for the parameters across all species, since they are all Antarctic penguins. This approach enabled the sharing of information among species while preserving the flexibility to model differences across species.

In the exploratory data analysis, we observed a significant similarity between the *Gentoo* and *Chinstrap* species, while the *Adelie* species appeared distinctly separate. Consequently, we considered modeling this distinction by creating a hierarchical model solely for the *Gentoo* and *Chinstrap* species, treating the *Adelie* species separately.
To implement this approach, we introduced a hyperparameter to model the prior mean for the coefficients of *Gentoo* and *Chinstrap*, while maintaining a separate prior for the *Adelie* coefficients. Although this model demonstrated good performance, closely resembling the previously presented models, it was ultimately discarded in favor of parsimony with similar or slightly improved goodness of fit. This decision was motivated by a preference for more interpretable models.

From the previous analysis, we noted that the distinction between genders also contained valuable information. Consequently, we began incorporating this separation into our models. Initially, we constructed a Separate model for both species and gender, considering all the variables and their interactions. We then proceeded with variable selection, culminating in the final model outlined in Section (3.4). This model exhibited superior performance, emerging as the most effective among the considered alternatives.

As a final experiment, we implemented a hierarchical model based on the previous model. In this setup, we accounted for the distinction between species and gender, introducing a hierarchical structure to the relationship between species.

Now we are going to present in detail the mathematical formulation of our tree best performing models.

### 3.2 Model 1: Separate by Species

For reference, the subscripts $j$ and $i$ are employed, with $j$ denoting the various species and $i$ representing the individual penguins, and $y_i$ correspond to each value of the bill length. This model is defined as follows:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} + \boldsymbol{\beta} X$
- $p(\gamma_j) \sim \mathcal{N}(20, 15)$ : To determine the parameters for this prior, we conducted three linear models in the R programming language using the '$lm()$' function, dividing the dataset into three subsets

corresponding to the species. The analysis revealed that the average of the intercepts was approximately 20. Consequently, we chose to designate this value as the mean of the informative normal distribution. To maintain a degree of non-informativeness for the intercept, we set a relatively high variance.

- $p(\beta d_j) \sim \mathcal{N}(0.79, 2)$ : Using the same linear models, we determined the mean value for the regression coefficient associated with the bill depth variable, setting it as the mean of the prior. The variance is chosen to be around the observed variance of the coefficient, resulting in a reasonably informative prior.
- $p(\beta s_j) \sim \mathcal{N}(2.5, 2)$ : Similarly to the previous case, this prior also originates from the linear models, considering the sex variable.
- $p(\sigma) \sim Gamma(29.90633, 1)$ : This prior choice is commonly used for representing the prior knowledge about the variance of a normal distribution. We determined the variance of the bill length variable and utilized it as the mean for the prior distribution of the normal variance. The chosen variance for the prior is relatively small, imparting an informative nature to the prior.

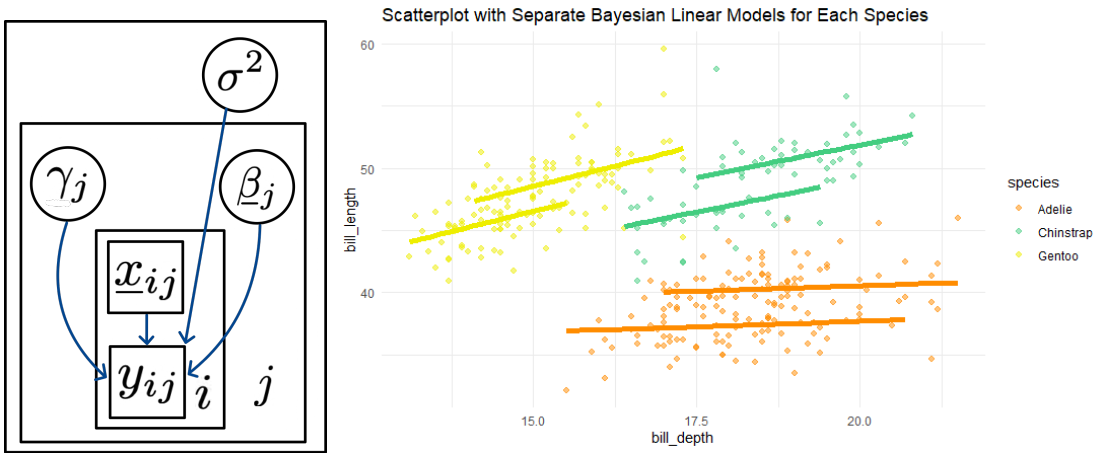



Figure 5: Structure and linear regression of the Separate model

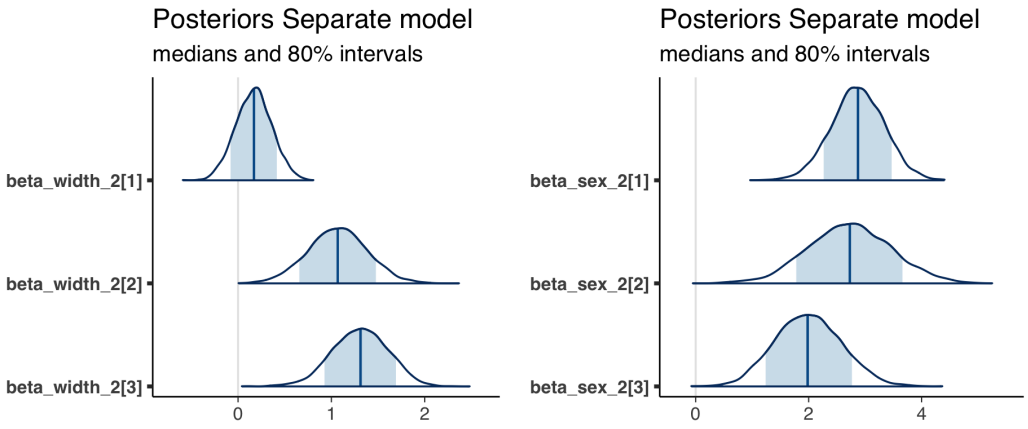



Figure 6: Posterior distributions of bill width and sex coefficients. The other posteriors are in the Appendix A.1

### 3.3 Model 2: Hierarchical

The second model is a Hierarchical model based on the previous separate one. We introduced a hierarchical with two hyper-parameters $\tau_d$ and $\tau_s$ for the respective parameters $\boldsymbol{\beta}_d$ and $\boldsymbol{\beta}_s$.

The model is defined as follows:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} + \boldsymbol{\beta} X$
- $p(\tau_d) \sim \mathcal{N}(0.79, 1)$ : We established a normal distribution as the prior for the hyperparameter of the regressor $\boldsymbol{\beta}_d$ associated with the covariate Bill Depth. The choice of mean aligns with Model (3.2), and we reduced the variance to 1.
- $p(\tau_s) \sim \mathcal{N}(2.5, 1)$ : Additionally, for the hyperparameter prior of the regressor $\boldsymbol{\beta}_s$ linked to the Sex covariate,we applied the same rationale as in the previous case to determine the values.
- $p(\beta d_j) \sim \mathcal{N}(\tau_d, 2)$ : The draws of the hyperparameter $\tau_d$ were employed as the mean for the prior of the regressor $\boldsymbol{\beta}_j$. We opted not to introduce hierarchy in the variance to avoid a substantial increase in the number of parameters, maintaining a simpler model.
- $p(\beta s_j) \sim \mathcal{N}(\tau_s, 2)$ : Similarly, the draws of the hyperparameter $\tau_s$ were utilized as the mean for the prior distribution of the regressor $\boldsymbol{\beta}_s$.
- $p(\gamma_j) \sim \mathcal{N}(20, 15)$ : The decision to refrain from making the intercept hierarchical was driven by the intent to provide even more flexibility for the models of different species to specialize. Indeed, upon observing the scatterplot in Figure (1) and experimenting, it becomes evident that the various species are positioned at different elevations. The prior for this parameter was selected analogously to that of Model (3.2).
- $p(\sigma) \sim Gamma(29.90633, 1)$ : We kept the same standard deviation of the Model 3.2.

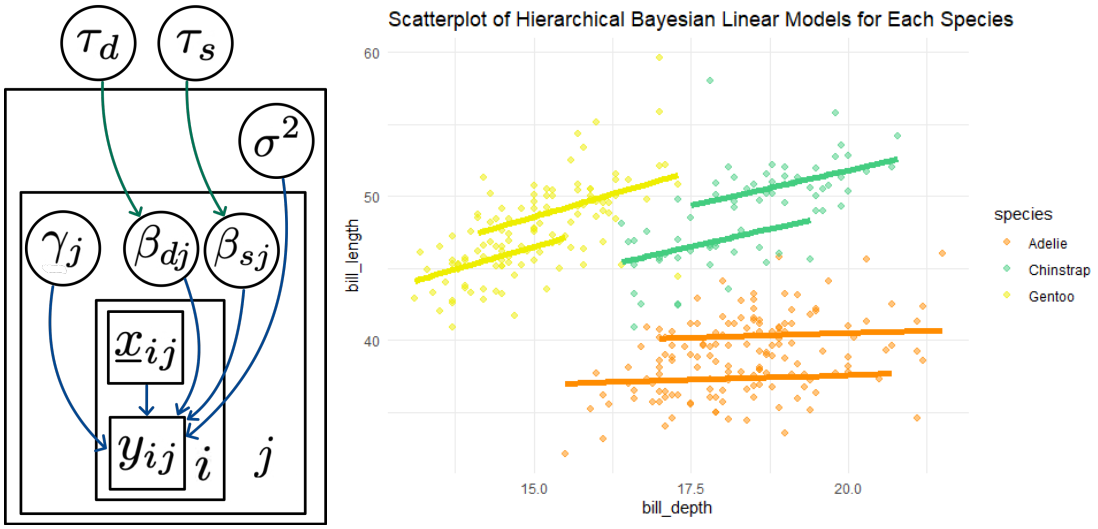



Figure 7: Structure and linear regression of the Hierarchical model

### 3.4 Model 3: Separate by Species and Sex

The third model we aim to present is one that distinguishes individuals both by species (index j) and by gender (index k).

### 3.4.1 Variable Selection

We started considering a complete model of the form:

$$Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2),\ where\ \boldsymbol{\mu} = \boldsymbol{\gamma} d + \boldsymbol{\gamma} s + \boldsymbol{\beta} spec \cdot X_{depth} + \boldsymbol{\beta} sex \cdot X_{depth} + interaction\ term \tag{3}$$

The parameters $\gamma_d$ and $\gamma_s$ represent the intercepts corresponding to species and sex, respectively. $\boldsymbol{\beta}_{\text{spec}}$ is the vector of regressors related to the covariate Bill Depth, separated by species, and $\boldsymbol{\beta}_{\text{sex}}$ is similarly divided by sex. For a detailed understanding of the interaction term, refer to Appendix (A.6).
Upon reviewing the model summary, we opted to exclude the *interaction term* since it added unnecessary complexity without contributing significant information.

The resulting model has the form:

$$Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2),\ where\ \boldsymbol{\mu} = \boldsymbol{\gamma} d + \boldsymbol{\gamma} s + \boldsymbol{\beta} spec \cdot X_{depth} + \boldsymbol{\beta} sex \cdot X_{depth} \tag{4}$$

By generating this model and examining the summary, we noticed that the values of the $\boldsymbol{\beta}_{\text{sex}}$ parameters were close to zero (around 0.02), thereby not adding significant information compared to the $\boldsymbol{\beta}_{\text{spec}}$ parameters, which were approximately 2.5. Consequently, we decided to eliminate these regressors and formulate the final model, which includes both intercepts, $\gamma_d$ and $\gamma_s$, and the regressor $\boldsymbol{\beta}_{\text{spec}}$.

The model takes the following form:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} d + \boldsymbol{\gamma} s + \boldsymbol{\beta} X$
- $p(\beta d_j) \sim \mathcal{N}(0.79, 2)$
- $p(\gamma_j) \sim \mathcal{N}(20, 15)$
- $p(\gamma_k) \sim \mathcal{N}(20, 15)$
- $p(\sigma) \sim Gamma(29.90633, 1)$

Priors were selected reasoning in the same way as for the previous models.

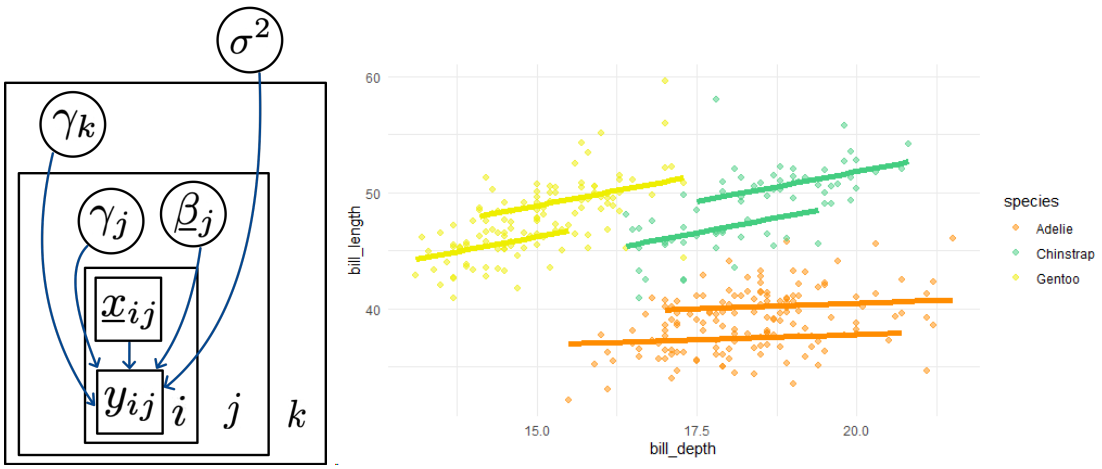



Figure 8: Structure and linear regression of the Separate model 3

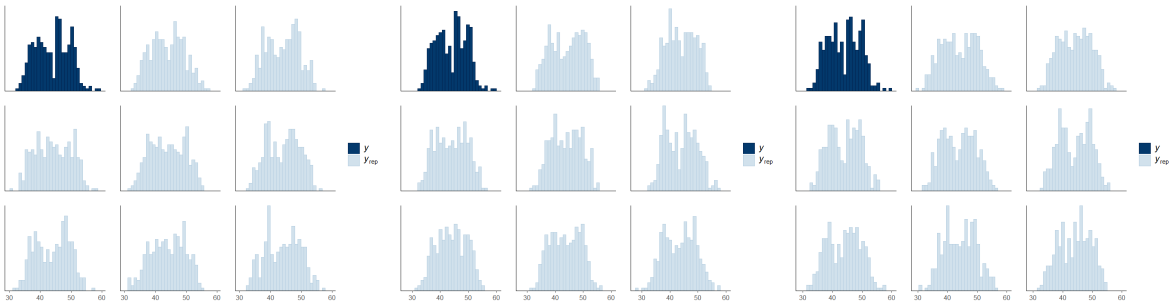
# 4 Model Checking

In this section we will discuss the goodness of fit and convergence of the three models described in the previous section.

### 4.1 Predictive Posterior Checking

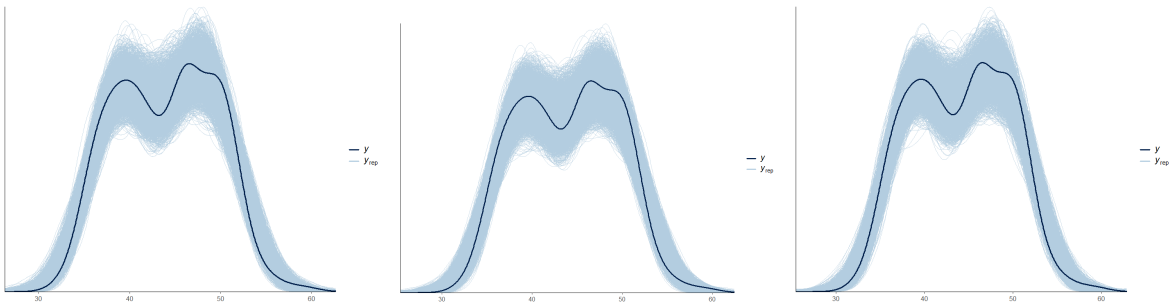
Posterior predictive checking involves simulating new data points from the posterior distribution and comparing these samples to the observed data.
In this section, we delve into the evaluation of our Bayesian models through posterior predictive checking. The primary goal is to assess the models' performance in replicating the observed data patterns. We employ the *bayesplot* package in R to generate informative visualizations of our three main models.

From the plots below, it is evident that the models behave in a very similar manner. Specifically, Figure (9) depicts a bar plot of the distribution of the target variable alongside replicated datasets. We observe that they are all quite similar and nicely simulate the true dataset. The plots in Figure (10) also represent the distribution of the target variable, and we can note that it is positioned nearly in the middle for all models. Lastly, Figure (11) illustrates the placement of the maxima and minima of the distribution. Despite the actual distribution not being precisely centered, the result is still deemed acceptable.



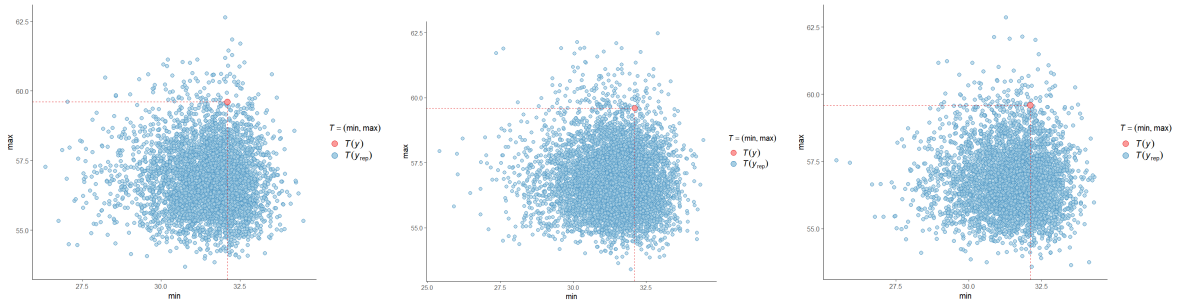
(a) Model 1 (b) Model 2 (c) Model 3

Figure 9: Barplot of the $y_{rep}$ dataset and the true one



(a) Model 1 (b) Model 2 (c) Model 3

Figure 10: Target variable distribution



(a) Model 1 (b) Model 2 (c) Model 3

Figure 11: Min-Max T statistic plot

Concluding this section, it's noteworthy to mention that additional plots were generated to further evaluate the models' performance. While these plots won't be reiterated here, it's essential to highlight that they consistently exhibited good behavior for the models. The comprehensive set of diagnostic plots collectively affirms the robustness and reliability of our Bayesian models.

### 4.2 Predictive Performance Assesment

The Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were evaluated for three models. The results are presented in the following table.

| Model | MAE | RMSE |
|---|---|---|
| Separate (Model 1) | 6.337994 | 7.799774 |
| Hierarchical (Model 2) | 6.337942 | 7.798255 |
| Separate (Model 3) | 6.334098 | 7.792790 |

Table 1: Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) for the three models.

**RMSE** (Root Mean Squared Error): Measures the average squared deviation between observed values ($y_i$) and predicted values ($\hat{y}_i$), providing a measure of the overall accuracy of the model. It's defined as: RMSE $= \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}$

**MAE** (Mean Absolute Error): Computes the average absolute deviation between observed and predicted values. This metric gives an estimate of the average deviation between the model predictions and the actual data. Its formula is: MAE $= \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$

The results are summarized in the table above, indicating small variations in MAE and RMSE across the models. These metrics provide insights into the models' predictive accuracy, with lower values suggesting better performance.
It can be observed that the third model exhibits slightly better performance compared to the other two; however, the minimal difference suggests that all models perform very similarly.

### 4.3 $\hat{R}$ Convergence Diagnostic

The Potential Scale Reductor Factor ($\hat{R}$) is a metric used to compare mean and variance of the chains used by the Hamiltonian Markov Chain employed by Stan.

$$\hat{R} = \sqrt{\frac{\hat{\text{var}}^{+}(\theta|\mathbf{y})}{W}} \tag{5}$$

M is the number of chains - M = 4 in our work
N i the number of draws - N = 2000

W is the within chains variance, defined as $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$.
B is the between chains variance and the total variance $var(\theta|\mathbf{y})$ is estimated as a weighted mean of W and B as: $\hat{\text{var}}(\theta|\mathbf{y}) = \frac{N-1}{N}W + \frac{1}{N}B$
The estimated variance $\hat{\text{var}}(\theta|\mathbf{y})$ overestimate the marginal posterior variance of the considered parameter and W underestimate the marginal posterior variance, so the metric $\hat{R}$ should be close to 1 and less than 1.01 to symbolise convergence of the chains.

We checked the $\hat{R}$ values for the parameters in our models by examining the summary, in the appendix.(A.2) With a sampling of N = 2000 draws, all $\hat{R}$ values are below 1.01. This indicates that the chains have reached convergence.

### 4.4 HMC Convergence Diagnostic

Hamiltonian Monte Carlo (HMC) is an advanced Markov Chain Monte Carlo (MCMC) algorithm used for sampling from complex probability distributions. Unlike traditional MCMC methods, HMC employs concepts from Hamiltonian dynamics and physics to propose more efficient and effective moves through the parameter space. It introduces the notion of a "momentum" variable that assists in exploring the target distribution, leading to faster convergence and improved sampling efficiency. We utilized HMC as the sampling algorithm for the Bayesian inference.

The Separate model and the Hierarchical model for species exhibited smooth convergence without encountering issues such as divergences or maximum treedepth problems. The sampling process for these models was successful, providing reliable results. We show in the following plots the parameter space exploration and the convergence of 1 chain of the Hamiltonian Monte Carlo for the Separate Model for species. (Figures (12), (13), (14)) The convergence plot for the others 2 models can be found in the Appendix (A.5).

During the fitting of the Separate model for both species and sex, we encountered a warning about convergence, specifically highlighting concerns with the maximum treedepth in the HMC algorithm. To address this, we adjusted the maximum treedepth to a higher value of 15, ensuring that the algorithm could effectively explore the parameter space without stopping prematurely.

The need for this adjustment was identified as being associated with our choice of a prior distribution for the intercepts of species and sexes, which had an high variance ($N(20, 80^2)$). This prior specification led to a more challenging parameter space for the sampler to navigate, requiring a higher maximum treedepth for a convergence and parameter estimation.

In Figures (12), (13), (14), are plotted the convergence diagnostics for Model 1 and in Appendix A.5 for the other two models. Overall, all the models show nice convergence for all the parameters. In Figure (12) and (14), we can see that the explorations of the parameters are stable and the parameter space nicely is visited. Figure (13) represents the values of the parameters at each iterations, and after approximately 100 iterations the parameters reach the convergence values.

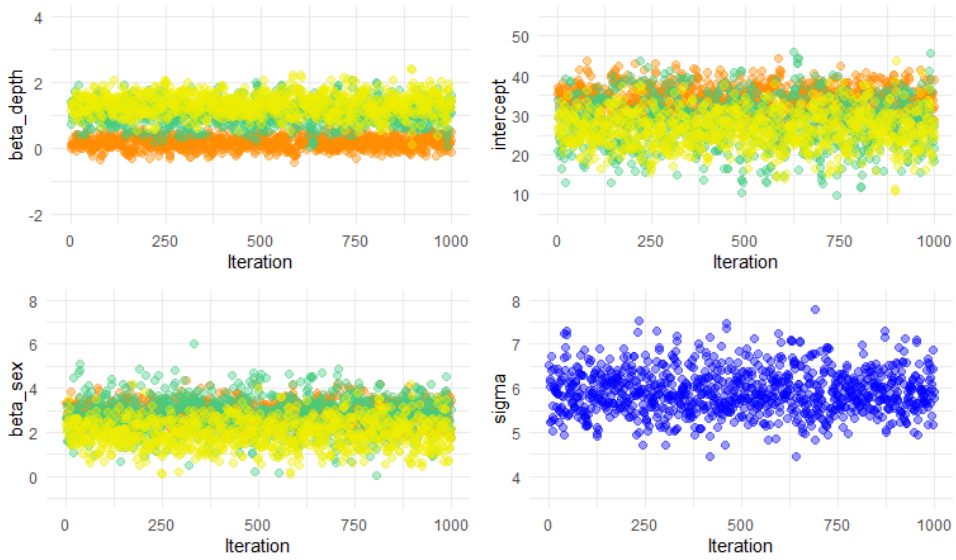



Figure 12: Uni-dimensional parameters space exploration in the Separate model for Species. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

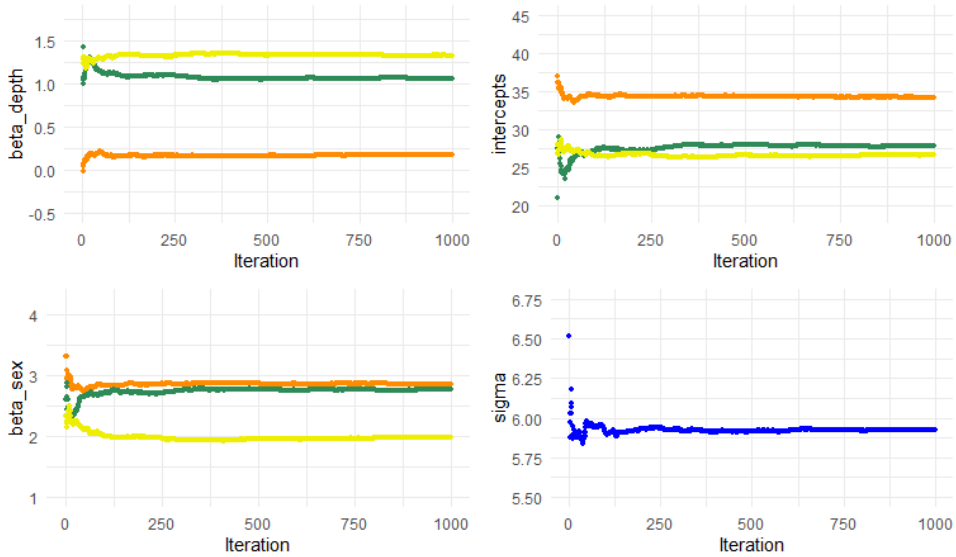



Figure 13: Convergences of means of the parameters in the Separate model for Species. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.


Figure 14: Two-dimensional parameter space exploration by the parameter Beta_sex and Beta_depth. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

### 4.5 ESS Diagnostic

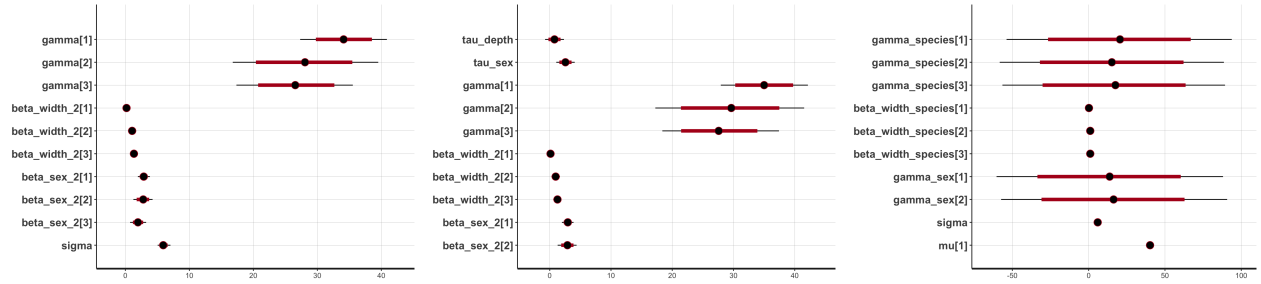
Effective Sample Size (ESS) is a measure used in Markov Chain Monte Carlo (MCMC) methods to quantify the efficiency of the samples generated from the Markov chain, since it's an estimate of the sample size required to achieve the same level of precision if that sample was a simple random sample. If all ESS values are high, it indicates that the MCMC sampling has effectively explored the parameter space, and the estimates are likely reliable.
If specific parameters have low ESS values, it might indicate poor convergence for those parameters.

We checked in the summary (A.2) the Effective Sample Size (ESS) for every parameter in Separate Model 1 (3.2), it exceeds 2000, indicating high reliability in the parameter estimates.
In the case of Hierarchical Model 2 (3.3), the ESS ranges between 2000 and 4000 for all parameters, demonstrating no significant issues across any parameter.
While the ESS in Separate Model 3 (3.4) is slightly lower than the other two models, it remains at a reliable level. The intercepts exhibit the lowest ESS (around 900); however, examination of the trace plot suggests well-mixed chains for these parameters.

## 5 Model Comparison

This section aims to provide a comparison of our three primary models. By evaluating and contrasting their performance, interpretability, and predictive capabilities. Through an analysis, we seek to identify the model that best aligns with the underlying patterns in the Antarctic penguin dataset and offers the most reliable insights into the relationship between covariates and the target variable.

Looking at the estimate values in Table (2) and (3), we can notice differences between the models. The coefficients comparable across the three models include the regression coefficients, denoted as $\beta_{\text{depth\_species}}$, for the Bill Depth variable, and the parameter $\sigma$. These values exhibit notable similarities across all three cases, with the most significant differences appearing in the last model. In this instance, the $\beta_{\text{depth}}$ for species 2 closely resembles that of species 3.




(a) Model 1 (b) Model 2 (c) Model 3

Figure 15: Estimated parameters of the three models

| Model 1 | | | Model 2 | | |
|---|---|---|---|---|---|
| Parameter | Mean | SD | Parameter | Mean | SD |
| $\gamma[1]$ | 34.1343472 | 3.42995480 | $\tau_{\text{depth}}$ | 0.7751507 | 0.77247831 |
| $\gamma[2]$ | 28.0562958 | 5.85154269 | $\tau_{\text{sex}}$ | 2.5652159 | 0.77739975 |
| $\gamma[3]$ | 26.6227971 | 4.59180723 | $\gamma[1]$ | 35.0173789 | 3.64944823 |
| $\beta_{\text{depth}}[1]$ | 0.1775301 | 0.19341901 | $\gamma[2]$ | 29.4899910 | 6.28640163 |
| $\beta_{\text{depth}}[2]$ | 1.0531996 | 0.33125545 | $\gamma[3]$ | 27.6837291 | 4.82807798 |
| $\beta_{\text{depth}}[3]$ | 1.3292044 | 0.32074966 | $\beta_{\text{depth}}[1]$ | 0.1272387 | 0.20615693 |
| $\beta_{\text{sex}}[1]$ | 2.8658170 | 0.48209318 | $\beta_{\text{depth}}[2]$ | 0.9720250 | 0.35547625 |
| $\beta_{\text{sex}}[2]$ | 2.7570436 | 0.76695340 | $\beta_{\text{depth}}[3]$ | 1.2549541 | 0.33789620 |
| $\beta_{\text{sex}}[3]$ | 1.9715207 | 0.62960150 | $\beta_{\text{sex}}[1]$ | 2.9378238 | 0.49076855 |
| $\sigma$ | 5.9460650 | 0.50869100 | $\beta_{\text{sex}}[2]$ | 2.8741374 | 0.78822092 |
| - | - | - | $\beta_{\text{sex}}[3]$ | 2.0801424 | 0.65145594 |
| - | - | - | $\sigma$ | 5.9343382 | 0.50040057 |

Table 2: Coefficients of Model 1 and Model 2

| Model 3 | | |
|---|---|---|
| Parameter | Mean | SD |
| $\gamma_{\text{species}}[1]$ | 20.5261494 | 36.65651582 |
| $\gamma_{\text{species}}[2]$ | 15.0412296 | 36.84815307 |
| $\gamma_{\text{species}}[3]$ | 17.2076495 | 36.67267837 |
| $\beta_{\text{depth\_species}}[1]$ | 0.1921333 | 0.18478301 |
| $\beta_{\text{depth\_species}}[2]$ | 1.0325848 | 0.29789899 |
| $\beta_{\text{depth\_species}}[3]$ | 1.0366700 | 0.26695028 |
| $\gamma_{\text{sex}}[1]$ | 13.4421372 | 36.64462186 |
| $\gamma_{\text{sex}}[2]$ | 16.1147205 | 36.66150748 |
| $\sigma$ | 5.9534098 | 0.51571670 |

Table 3: Coefficients of Model 3

### 5.1 Leave-One-Out Cross-Validation for Model Selection

To assess and compare the prediction performances of our models, we employ the Expected Log Predictive Density (ELPD) computed through Leave-One-Out Cross-Validation (LOO). This method involves iteratively leaving out each data point and evaluating the model's predictive accuracy. By applying LOO to our three main models, we aim to gain insights into their relative performances. Additionally, we extend our comparison to include four alternative models outlined in Section (3.1): Pooled[3], Separate for Species 1 and Hierarchical for Species 2 and 3[1], Separate with Only Intercepts[4], and Separate only by species without Sex Variable[2].

| | elpd_diff | se_diff |
|---|---|---|
| **Model 3** | 0.0 | 0.0 |
| **Model 1** | -0.3 | 1.2 |
| Hierar.+Sep.[1] | -0.5 | 1.3 |
| **Model 2** | -0.6 | 1.0 |
| No sex[2] | -26.1 | 6.7 |
| Pooled[3] | -59.6 | 9.3 |
| Intercepts[4] | -89.9 | 11.6 |

The table presents the differences in Expected Log Predictive Density (elpd_diff) and their standard errors (se_diff) for the models using Leave-One-Out Cross-Validation. Notably, the top four models (Model 3, Model 1, Hierar.+Sep.[1], Model 2) exhibit similar outcomes, emphasizing the significance of incorporating information like the sex and depth variables.
The third-place model, despite competitive results, is not emphasized in the report due to its close alignment with Model 2.
As highlighted, the last three models are not able to capture the intricacies of the data and effectively modeling the target variable, Bill Length. This comparative analysis emphasizes that our selected models, in contrast, captures significant patterns and demonstrates robust performances.

The LOO function in R provides a valuable diagnostic tool known as the Pareto K diagnostic, which is instrumental in assessing the reliability of Leave-One-Out Cross-Validation results. In the Appendix (A.3), the R output for the main three models is available, including the Pareto K values. The Pareto K diagnostic computes the relative importance of each observation, highlighting potential points of concern.

It is worth noting that all points in the models under consideration were deemed satisfactory, with their Pareto K values consistently below 0.5. In conclusion, the Pareto K diagnostic results reinforce our confidence in the reliability of the models.

# 6 Sensitivity Analysis

In this part, we assessed the impact of prior selection on the results of three models, ranging from parameter estimates to model stability. Before starting with the description of the procedures and the results, we must say that we tried to change for every model the distribution from normal to student-t, but there were almost no difference in the outcomes, probably because the relatively high sample size allows us to use normal distributions. For this reason, we focused more on changing the parameters of the distributions rather than altering the distributions themselves.

## 6.1 Separate Model 1

Our first separate model has quite informative priors (see (3.2) for the priors). The first test we wanted to try is the difference in results if we set the priors very uninformative. So, we set $\gamma_i$, $\beta_{dj}$, $\beta_{sj}$ and $\sigma$ all $\mathcal{N}(0, 1e+14)$. Such high variances makes the model striking uninformative. This model has converged nicely, having all $\hat{R}$ lower than 1.1 (see Appendix (A.2) for model checking).

The results indicate a similarity between the outcomes, suggesting that the model is highly robust and influenced primarily by the data. This implies that the impact of the priors is marginal.

Assessing the LOO diagnostic for the predictive performance of the models, we observe similarities in the results, but with a slight improvement in the uninformative model. However, given the minimal nature of this improvement, it is challenging to definitively determine which model is superior in this sense. In Figure (16) we can see that the drawings from the models are remarkably alike.

| | Uninformative | | Informative | |
|---|---|---|---|---|
| Parameter | Mean | SD | Mean | SD |
| $\gamma[1]$ | 35.3973559 | 3.37974206 | 34.2786602 | 3.47498591 |
| $\gamma[2]$ | 29.4388530 | 6.35827167 | 27.8277349 | 5.81403349 |
| $\gamma[3]$ | 27.1661635 | 4.57842030 | 26.8314839 | 4.52603612 |
| $\beta_{\text{depth}}[1]$ | 0.1055596 | 0.19123717 | 0.1693109 | 0.19605960 |
| $\beta_{\text{depth}}[2]$ | 0.9740629 | 0.36104189 | 1.0660480 | 0.32900600 |
| $\beta_{\text{depth}}[3]$ | 1.2920154 | 0.32081048 | 1.3146261 | 0.31655533 |
| $\beta_{\text{sex}}[1]$ | 2.9854825 | 0.46941127 | 2.8752563 | 0.47993043 |
| $\beta_{\text{sex}}[2]$ | 2.9042879 | 0.83050298 | 2.7377248 | 0.76026536 |
| $\beta_{\text{sex}}[3]$ | 1.9960787 | 0.64145404 | 1.9855729 | 0.62077269 |
| $\sigma$ | 5.1038434 | 0.40792468 | 5.9234718 | 0.50914733 |

Table 4: Summary statistics for Uninformative and Informative separate models

| | Uninformative | | Informative | |
|---|---|---|---|---|
| Measure | Estimate | SE | Estimate | SE |
| elpd_loo | -748.6 | 19.2 | -749.5 | 16.4 |
| p_loo | 12.1 | 2.2 | 9.5 | 1.8 |
| looic | 1497.2 | 38.4 | 1498.9 | 32.9 |

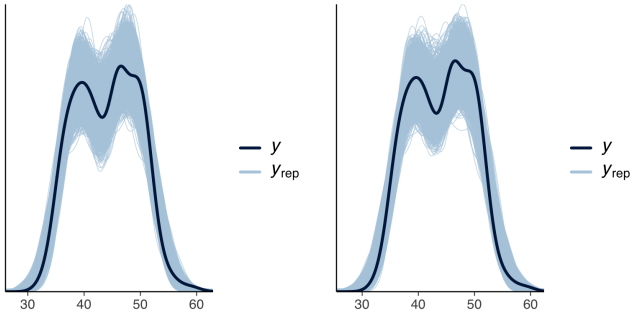



Figure 16: Drawings from the uninformative model on the left and from the informative model on the right

Since with uninformative priors is hard to see any difference, we also tried setting priors with "wrong" means, in the sense that the mean of the distribution is way different from the estimates of the data. We used the following priors:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} + \boldsymbol{\beta} X$
- $p(\gamma_j) \sim \mathcal{N}(0, 12)$ : The intercept are far from 0, so this prior may be considered wrong to use. In order to not set it too bad, the variance is chosen such that the estimated values are still in the tails of the distribution.
- $p(\beta d_j) \sim \mathcal{N}(10, 4)$ : Applying the same reasoning, we set this prior for the regression coefficient of bill depth.
- $p(\beta s_j) \sim \mathcal{N}(11, 4)$ : *Same as before.*
- $p(\sigma) \sim Gamma(6, 0.1)$ : *Same as before.*

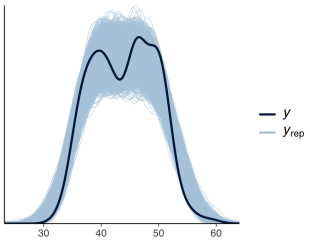



Figure 17: Drawings from the 'wrong' model

The estimate parameters are now remarkably different from the original separate model. Despite this, we can see that they are relatively not too far from the estimated parameters, meaning that the data has an higher impact compared to the priors, which might be a sign of robustness of the model.

In this case, as expected, the estimated ELPD LOO is $-773.4$, which is lower than the original separate model, even if not that terrible.

| | Wrong | | Selected Model | |
|---|---|---|---|---|
| Parameter | Mean | SD | Mean | SD |
| $\gamma[1]$ | 16.8544283 | 2.44020964 | 34.2786602 | 3.47498591 |
| $\gamma[2]$ | 6.8615444 | 3.09114106 | 27.8277349 | 5.81403349 |
| $\gamma[3]$ | 9.8922211 | 2.78314668 | 26.8314839 | 4.52603612 |
| $\beta_{\text{width\_2}}[1]$ | 1.1408371 | 0.13845559 | 0.1693109 | 0.19605960 |
| $\beta_{\text{width\_2}}[2]$ | 2.2289663 | 0.17583539 | 1.0660480 | 0.32900600 |
| $\beta_{\text{width\_2}}[3]$ | 2.4811646 | 0.19498985 | 1.3146261 | 0.31655533 |
| $\beta_{\text{sex\_2}}[1]$ | 1.9632054 | 0.44546616 | 2.8752563 | 0.47993043 |
| $\beta_{\text{sex\_2}}[2]$ | 1.7240289 | 0.65527713 | 2.7377248 | 0.76026536 |
| $\beta_{\text{sex\_2}}[3]$ | 0.8292231 | 0.52194463 | 1.9855729 | 0.62077269 |
| $\sigma$ | 5.9972517 | 0.09791851 | 5.9234718 | 0.50914733 |

### 6.2 Hierarchical Model 2

We use the same approach as before on the hierarchical model.

As uninformative priors we used $\mathcal{N}(0, 1e+14)$ for the parameters $\tau_d$ $\tau_s$ $\gamma_i$ and $\sigma$, meanwhile $\beta_{dj}$ and $\beta_{sj}$ are determined hierarchically using $\tau_d$ and $\tau_s$ as means and $1e+14$ as variance.

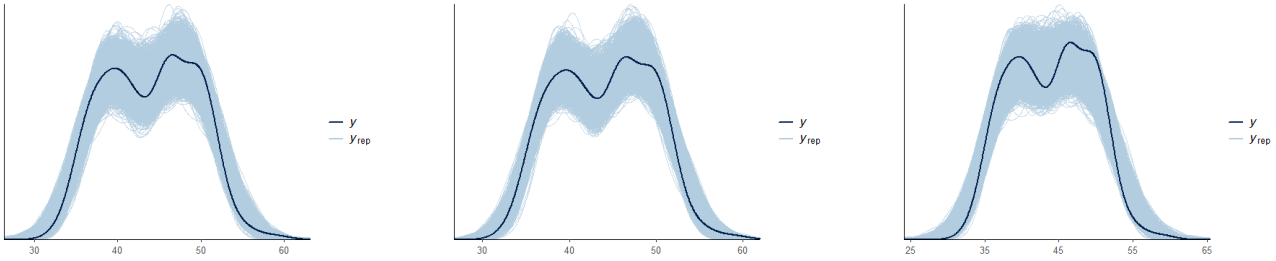
As "wrong" priors we used:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} + \boldsymbol{\beta} X$
- $\tau d \sim \mathcal{N}(10, 4)$
- $\tau_s \sim \mathcal{N}(10, 4)$
- $p(\gamma_j) \sim \mathcal{N}(0, 12)$
- $p(\beta d_j) \sim \mathcal{N}(\tau_d, 4)$
- $p(\beta s_j) \sim \mathcal{N}(\tau_s, 4)$
- $p(\sigma) \sim Gamma(6, 0.1)$

| Parameter | **Selected Model** | | **Uninformative** | | **Wrong** | |
|---|---|---|---|---|---|---|
| | Mean | SE_Mean | Mean | SE_Mean | Mean | SE_Mean |
| $\tau_{\text{depth}}$ | 0.7915403 | 0.0093431021 | 91857.227229 | 130628.04383 | 4.012518 | 0.013578 |
| $\tau_{\text{sex}}$ | 2.5333379 | 0.0098782696 | 81206.131959 | 130461.36333 | 3.266633 | 0.013082 |
| $\gamma[1]$ | 35.0903047 | 0.0640304799 | 35.276938 | 0.04557 | 16.297996 | 0.048124 |
| $\gamma[2]$ | 29.3910263 | 0.1298006340 | 29.583950 | 0.08121 | 6.109085 | 0.058162 |
| $\gamma[3]$ | 27.5168203 | 0.0865451577 | 27.095400 | 0.06260 | 9.253754 | 0.051088 |
| $\beta_{\text{depth\_2}}[1]$ | 0.1235972 | 0.0036491113 | 0.112441 | 0.00259 | 1.180872 | 0.002777 |
| $\beta_{\text{depth\_2}}[2]$ | 0.9776790 | 0.0074154315 | 0.966019 | 0.00463 | 2.289718 | 0.003400 |
| $\beta_{\text{depth\_2}}[3]$ | 1.2663930 | 0.0061154528 | 1.297450 | 0.00440 | 2.538968 | 0.003674 |
| $\beta_{\text{sex\_2}}[1]$ | 2.9369377 | 0.0073885806 | 2.963884 | 0.00597 | 1.600970 | 0.007194 |
| $\beta_{\text{sex\_2}}[2]$ | 2.8572171 | 0.0147529799 | 2.914874 | 0.01015 | 0.997978 | 0.010366 |
| $\beta_{\text{sex\_2}}[3]$ | 2.0628052 | 0.0108633619 | 1.982915 | 0.00812 | 0.383350 | 0.008103 |
| $\sigma$ | 5.9472802 | 0.0068006859 | 5.087638 | 0.00458 | 6.133047 | 0.007043 |

From the coefficients values, we observe that the use of Uninformative priors complicates convergence, although the resulting model closely resembles the one we selected. When the "wrong" priors are used in this model, they have a significant influence, leading to substantially different coefficients and indicating reduced robustness. Also the LOO of the model with the uninformative priors is very similar to the one of the selected model, meanwhile the LOO of the model with the wrong priors is higher. The Figure 18 shows that the "wrong" priors make the model follow the tail worse than with the other 2 models.

| | **Selected Model** | | **Uninformative** | | **Wrong** | |
|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE |
| elpd_loo | -749.8 | 16.4 | -748.2 | 19.2 | -774.4 | 16.9 |
| p_loo | 9.8 | 1.8 | 11.7 | 2.2 | 8.6 | 1.5 |
| looic | 1499.7 | 32.8 | 1496.4 | 38.4 | 1548.7 | 33.9 |




(a) Selected priors (b) Uninformative priors (c) "Wrong" priors

Figure 18: Density draws

### 6.3 Separate Model 3

As uninformative priors we used $\mathcal{N}(0, 1e+14)$ for all the parameters: $\gamma_{species}$, $\gamma_{sex}$, $\beta_{depth_species}$ and $\sigma$.

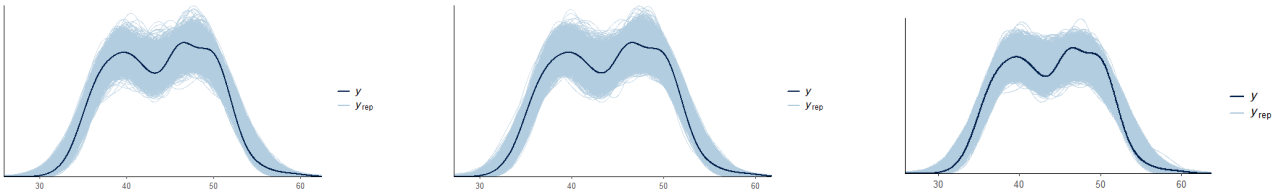
As "wrong" priors we used:

- $Y \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2)$, where $\boldsymbol{\mu} = \boldsymbol{\gamma} + \boldsymbol{\beta} X$
- $p(\gamma_j) \sim \mathcal{N}(0, 12)$
- $p(\gamma_j) \sim \mathcal{N}(0, 12)$

- $p(\beta_{depthspecies}) \sim \mathcal{N}(\tau_{sex}, 4)$
- $p(\sigma) \sim Gamma(6, 0.1)$

| Parameter | **Selected Model** | | **Uninformative** | | **Wrong** | |
|---|---|---|---|---|---|---|
| | Mean | SE_Mean | Mean | SE_Mean | Mean | SE_Mean |
| $\gamma_{\text{species}[1]}$ | 20.9860353 | 1.478071559 | 666.08624 | 535.09083 | 10.706844 | 0.042517 |
| $\gamma_{\text{species}[2]}$ | 15.4043824 | 1.482269843 | 660.25068 | 534.89628 | 3.738666 | 0.047621 |
| $\gamma_{\text{species}[3]}$ | 17.6428818 | 1.468737844 | 663.09478 | 535.34163 | 6.979063 | 0.042924 |
| $\beta_{\text{width\_species}[1]}$ | 0.1849692 | 0.003541498 | 0.17277 | 0.01515 | 0.949533 | 0.002521 |
| $\beta_{\text{width\_species}[2]}$ | 1.0304351 | 0.005076587 | 1.03196 | 0.02497 | 1.867793 | 0.003224 |
| $\beta_{\text{width\_species}[3]}$ | 1.0296529 | 0.005023941 | 0.99142 | 0.01492 | 1.992262 | 0.003421 |
| $\gamma_{\text{sex}[1]}$ | 13.1073366 | 1.466682387 | -631.78388 | 535.20891 | 9.968224 | 0.036776 |
| $\gamma_{\text{sex}[2]}$ | 15.7899687 | 1.466259894 | -629.07696 | 535.21331 | 11.363785 | 0.039298 |
| $\sigma$ | 5.9284883 | 0.010100467 | 5.08468 | 0.03087 | 5.661602 | 0.006816 |

| | **Selected Model** | | **Uninformative** | | **Wrong** | |
|---|---|---|---|---|---|---|
| | Estimate | SE | Estimate | SE | Estimate | SE |
| elpd_loo | -748.8 | 16.3 | -746.9 | 19.0 | -760.3 | 17.3 |
| p_loo | 8.4 | 1.6 | 9.5 | 1.8 | 7.7 | 1.5 |
| looic | 1497.7 | 32.6 | 1493.7 | 38.0 | 1520.7 | 34.5 |




(a) Selected priors (b) Uninformative priors (c) "Wrong" priors

Figure 19: Density draws

The results indicate that the model with the "wrong" prior yields different coefficients compared to the considered model. However, despite the discrepancies in coefficients, the leave-one-out (LOO) cross-validation score is lower for the selected model.

The model with the not-so-helpful starting information has problems converging, even when adjusting the maximum treedepth of the Hamiltonian Monte Carlo. This takes a long time to compute and shows strange patterns in LOO cross-validation. Despite its seemingly good performance in LOO, the coefficients lack meaningful interpretation.

In comparison, the Hierarchical model for Sex and Species demonstrates less robustness than the Hierarchical model solely for Species.

# 7 Discussion

In the discussion, it is essential to acknowledge the limitations and potential areas for improvement in our modeling approach. In the first place one could improve our chosen priors, which are basic and not deeply informed by prior and scientific knowledge. Additionally, the parameters within these distributions are heavily data-driven, reflecting a potential limitation in leveraging more nuanced prior information. This is evident in the deviation observed in Figure 11, where the placement of the min-max point is not perfectly centered, prompting further investigation into this discrepancy.

Furthermore, our models exhibited overall satisfactory performance in model checking methods. However, it is plausible that more sophisticated techniques could uncover potential pitfalls not captured

by our current approaches. Exploring advanced model checking techniques could lead to a more comprehensive understanding of model behavior and guide improvements, for example comparing the predictive performances of the models with unseen data.

In summary, while our Bayesian models have provided valuable insights into the Antarctic penguins dataset, there is room for enhancement by refining the choice of priors and delving deeper into advanced model checking methodologies.

# 8 Conclusion

In conclusion, our exploration of Bayesian models for predicting the bill length of Antarctic penguins has provided valuable insights and considerations. The main models, despite their differences in complexity, demonstrated overall robust behavior across various model checking techniques.

While Model 3 exhibited slightly better results in LOO checking, the significance of this improvement is tempered by the potential for overfitting and the inherent variance in LOO estimates.
Considering simplicity, the two separate models (Models 1 and 3) emerge as strong contenders due to their minimal parameter counts (10 and 9, respectively). Simplicity is a crucial consideration, and these models strike a balance between performance and complexity.

Notably, Model 1 showcased robustness in sensitivity analysis (6.1). This characteristic establishes Model 1 as a trustworthy choice, especially in situations where stability is important.

Conversely, the Hierarchical Model 2, with its elevated parameter count (12), seems less justified for our specific problem's simplicity. The additional parameters do not seem to contribute substantially to the model's performance, making it a less favorable choice when prioritizing parsimony.

In summary, while each model has its merits, Models 1 and 2, with their simplicity and robust performance, stand out as preferable choices for predicting penguin Bill Lengths in our Bayesian framework.

# 9 AI Disclosure

The application of AI has been employed in these cases:

- To help us plot the figures of the Data Exploration 2. In particular for Figure (1), (2), (3).
- To enhance the accuracy of English sentences.

# 10 References

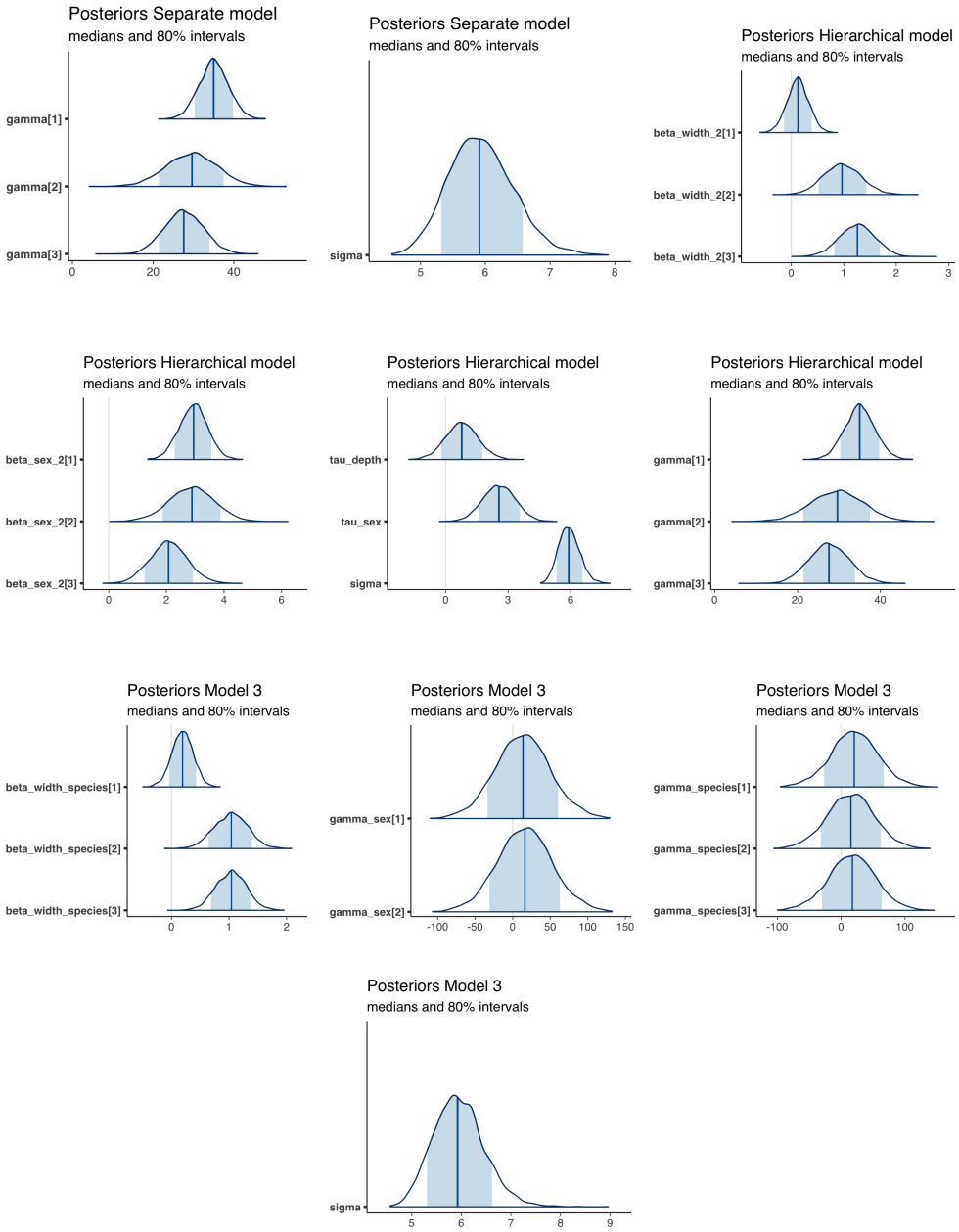

1. Avalos Pacheco, Alejandra. Slide of Bayes-Statistik, Technische Universität Wien, 2023W
2. `https://mc-stan.org/`
3. Gorman, K. B., Williams, T. D., and Fraser, W. R. (2014). Ecological sexual dimorphism and environmental variability within a community of antarctic penguins (genus pygoscelis). PloS one, 9(3):e90081–e90081.
4. Lee, W. Y., Jung, J.-W., Han, Y.-D., Chung, H., and Kim, J.-H. (2015). A new sex determination method using morphological traits in adult chinstrap and gentoo penguins on king george island, antarctica. Animal cells and systems, 19(2):156–159.

# A Appendix

You can find the project on GitHub at the following link:
`https://github.com/rondvde/Bayesian-Project-Erasmus-peeps.git`.

## A.1 Posterior distributions

Plot of the posterior ditribution of Model1, Model 2, and Model 3 (besides posteriors of regression coefficients of Model 1, which are already plotted in Section 3.2.




## A.2 Summary of the three models

In this section, summaries of the three models are showed, without the 333 $\mu$, $log\_lik$ and the resampled data.

| Parameter | Mean | Se_Mean | SD | 2.5% | 25% | 50% | 75% | 97.5% | $n_{eff}$ | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma[1]$ | 34.134 | 0.070 | 3.430 | 27.315 | 31.795 | 34.109 | 36.481 | 40.869 | 2419.150 | 1.001 |
| $\gamma[2]$ | 28.056 | 0.131 | 5.852 | 16.780 | 24.217 | 28.069 | 31.998 | 39.520 | 1996.848 | 1.001 |
| $\gamma[3]$ | 26.623 | 0.106 | 4.592 | 17.357 | 23.612 | 26.550 | 29.649 | 35.544 | 1872.798 | 1.001 |
| $\beta_{depth}[1]$ | 0.178 | 0.004 | 0.193 | -0.202 | 0.044 | 0.177 | 0.308 | 0.564 | 2379.563 | 1.001 |
| $\beta_{depth}[2]$ | 1.053 | 0.007 | 0.331 | 0.411 | 0.835 | 1.052 | 1.269 | 1.695 | 1953.341 | 1.001 |
| $\beta_{depth}[3]$ | 1.329 | 0.007 | 0.321 | 0.708 | 1.117 | 1.334 | 1.538 | 1.965 | 1843.361 | 1.001 |
| $\beta_{sex}[1]$ | 2.866 | 0.009 | 0.482 | 1.944 | 2.532 | 2.866 | 3.187 | 3.828 | 3076.832 | 1.000 |
| $\beta_{sex}[2]$ | 2.757 | 0.016 | 0.767 | 1.252 | 2.231 | 2.788 | 3.262 | 4.246 | 2413.125 | 1.001 |
| $\beta_{sex}[3]$ | 1.972 | 0.013 | 0.630 | 0.737 | 1.548 | 1.959 | 2.389 | 3.244 | 2180.903 | 1.001 |
| $\sigma$ | 5.946 | 0.009 | 0.509 | 5.043 | 5.600 | 5.915 | 6.263 | 7.039 | 3315.533 | 1.001 |

Table 5: Summary Statistics Model 1

| Parameter | Mean | Se_Mean | SD | 2.5% | 25% | 50% | 75% | 97.5% | $n_{eff}$ | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_{depth}$ | 0.775 | 0.010 | 0.772 | -0.758 | 0.255 | 0.773 | 1.291 | 2.317 | 5863.918 | 1.000 |
| $\tau_{sex}$ | 2.565 | 0.011 | 0.777 | 1.064 | 2.036 | 2.561 | 3.092 | 4.078 | 5186.152 | 1.000 |
| $\gamma[1]$ | 35.017 | 0.068 | 3.649 | 27.929 | 32.641 | 35.005 | 37.409 | 42.147 | 2875.428 | 1.000 |
| $\gamma[2]$ | 29.490 | 0.120 | 6.286 | 17.240 | 25.220 | 29.634 | 33.839 | 41.540 | 2721.829 | 1.001 |
| $\gamma[3]$ | 27.684 | 0.089 | 4.828 | 18.383 | 24.413 | 27.583 | 30.959 | 37.426 | 2924.914 | 1.001 |
| $\beta_{width\_2}[1]$ | 0.127 | 0.004 | 0.206 | -0.279 | -0.007 | 0.128 | 0.261 | 0.532 | 2816.234 | 1.000 |
| $\beta_{width\_2}[2]$ | 0.972 | 0.007 | 0.355 | 0.286 | 0.725 | 0.967 | 1.215 | 1.664 | 2670.841 | 1.002 |
| $\beta_{width\_2}[3]$ | 1.255 | 0.006 | 0.338 | 0.576 | 1.024 | 1.261 | 1.483 | 1.899 | 2879.319 | 1.001 |
| $\beta_{sex\_2}[1]$ | 2.938 | 0.008 | 0.491 | 1.983 | 2.609 | 2.947 | 3.266 | 3.886 | 3885.554 | 1.000 |
| $\beta_{sex\_2}[2]$ | 2.874 | 0.014 | 0.788 | 1.291 | 2.348 | 2.889 | 3.406 | 4.385 | 3273.271 | 1.001 |
| $\beta_{sex\_2}[3]$ | 2.080 | 0.011 | 0.651 | 0.823 | 1.650 | 2.077 | 2.512 | 3.383 | 3475.626 | 1.001 |
| $\sigma$ | 5.934 | 0.007 | 0.500 | 5.023 | 5.580 | 5.909 | 6.260 | 6.975 | 5390.032 | 1.001 |

Table 6: Summary Statistics Model 2

| Parameter | Mean | Se_Mean | SD | 2.5% | 25% | 50% | 75% | 97.5% | $n_{eff}$ | Rhat |
|---|---|---|---|---|---|---|---|---|---|---|
| $\gamma_{species}[1]$ | 20.526 | 1.290 | 36.657 | -53.787 | -3.701 | 20.580 | 44.525 | 93.766 | 807.925 | 1.002 |
| $\gamma_{species}[2]$ | 15.041 | 1.305 | 36.848 | -58.283 | -9.453 | 15.221 | 39.714 | 88.615 | 797.836 | 1.002 |
| $\gamma_{species}[3]$ | 17.208 | 1.289 | 36.673 | -56.470 | -7.034 | 17.576 | 41.934 | 89.357 | 809.810 | 1.002 |
| $\beta_{width\_species}[1]$ | 0.192 | 0.003 | 0.185 | -0.177 | 0.069 | 0.193 | 0.318 | 0.550 | 2984.896 | 1.000 |
| $\beta_{width\_species}[2]$ | 1.033 | 0.005 | 0.298 | 0.453 | 0.831 | 1.040 | 1.233 | 1.599 | 2946.168 | 1.001 |
| $\beta_{width\_species}[3]$ | 1.037 | 0.005 | 0.267 | 0.518 | 0.853 | 1.041 | 1.221 | 1.560 | 2587.678 | 1.000 |
| $\gamma_{sex}[1]$ | 13.442 | 1.291 | 36.645 | -60.348 | -10.817 | 13.715 | 37.707 | 88.019 | 805.428 | 1.002 |
| $\gamma_{sex}[2]$ | 16.115 | 1.291 | 36.662 | -57.337 | -8.041 | 16.295 | 40.267 | 90.741 | 806.264 | 1.002 |
| $\sigma$ | 5.953 | 0.011 | 0.516 | 5.022 | 5.599 | 5.925 | 6.268 | 7.048 | 2181.186 | 1.000 |

Table 7: Summary Statistics Model 3

### A.3 LOO summaries

Outputs of *loo()* function on the three models.

**Model 1:**

```
loo3 <- loo(fit3, moment_match = TRUE) #Model 1
loo3

Computed from 4000 by 333 log-likelihood matrix

         Estimate   SE
elpd_loo   -749.5 16.3
p_loo         9.5  1.8
looic      1499.1 32.7
------
Monte Carlo SE of elpd_loo is 0.1.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

**Model 2:**

```
loo2 <- loo(fit2, moment_match = TRUE) #Model 2
loo2

Computed from 6000 by 333 log-likelihood matrix

         Estimate   SE
elpd_loo   -749.8 16.5
p_loo         9.9  1.8
looic      1499.6 32.9
------
Monte Carlo SE of elpd_loo is 0.1.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

**Model 3:**

```
loo8 <- loo(fit8, moment_match = TRUE) #Model 3
loo8

Computed from 4000 by 333 log-likelihood matrix

         Estimate   SE
elpd_loo   -749.2 16.3
p_loo         8.7  1.7
looic      1498.4 32.5
------
Monte Carlo SE of elpd_loo is 0.1.

All Pareto k estimates are good (k < 0.5).
See help('pareto-k-diagnostic') for details.
```

## A.4 Stan Code

### A.4.1 Model 1: Separate Model

```
data {
```

```
  int<lower=0> N;// number of observations
  int<lower=0> N_new[3];
  vector[N] bill_length;   // response variable
  vector[N] bill_depth;// covariate
  vector[N] sex;
  int<lower=1> N_species;   // number of unique species
  int<lower=1, upper=N_species> species[N];  // species indicator
}

parameters {
  vector[N_species] gamma;// species-specific intercepts
  vector[N_species] beta_width_2; //spesicif intercept for bill width
  vector[N_species] beta_sex_2;
  real<lower=0> sigma;      // standard deviation of the residuals
}
transformed parameters{
  vector[N] mu;  // linear predictor
  for (i in 1:N){
    mu[i]= gamma[species[i]]+ beta_width_2[species[i]] * bill_depth[i]+
    beta_sex_2[species[i]] * sex[i];
  }
}
model {

  // Priors
  sigma ~ gamma(29.90633, 1);

  for (j in 1:N_species){
    gamma[j] ~ normal(20, 15);
    beta_width_2[j] ~ normal(0.79, 2);
    beta_sex_2[j] ~ normal(2.5, 2);
  }

  bill_length ~ normal(mu, sqrt(sigma));
}

generated quantities {
  vector[N] log_lik;
  real ynew[N];

  for (i in 1:N) {
    log_lik[i] = normal_lpdf(bill_length[i] | mu[i], sqrt(sigma));
  }
  for (n in 1:N){
    ynew[n] = normal_rng(mu[n], sqrt(sigma));
  }
}
```

### A.4.2 Model 2: Hierarchical Model

```
data {
  int<lower=0> N;// number of observations
```

```
  int<lower=0> N_new[3];
  vector[N] bill_length;   // response variable
  vector[N] bill_width;// covariate
  vector[N] sex;
  int<lower=1> N_species;   // number of unique species
  int<lower=1, upper=N_species> species[N];  // species indicator
}

parameters {
  real tau_depth;
  real tau_sex;
  vector[N_species] gamma;// species-specific intercepts
  vector[N_species] beta_width_2; //spesicif intercept for bill width
  vector[N_species] beta_sex_2;
  real<lower=0> sigma;      // standard deviation of the residuals
}

transformed parameters{
  vector[N] mu;  // linear predictor
  for (i in 1:N){
    mu[i]= gamma[species[i]]+ beta_width_2[species[i]] * bill_width[i]+
    beta_sex_2[species[i]] * sex[i];
  }
}

model {
  // Priors
  sigma ~ gamma(29.90633, 1);
  tau_depth ~ normal(0.79, 1);
  tau_sex ~ normal(2.5, 1);

  for (j in 1:N_species){
    gamma[j] ~ normal(20, 80);
    beta_width_2[j] ~ normal(tau_depth, 2);
    beta_sex_2[j] ~ normal(tau_sex, 2);
  }
  bill_length ~ normal(mu, sqrt(sigma));
}

generated quantities {
  vector[N] log_lik;  // vector to store log likelihood values for each species
  vector[N] ynew;

  for (i in 1:N) {
    log_lik[i] = normal_lpdf(bill_length[i] | mu[i], sqrt(sigma));
  }

  for (n in 1:N){
    ynew[n] = normal_rng(mu[n], sqrt(sigma));
  }
}
```

### A.4.3 Model 3: Separate for Species and Sex with Feature Selection

```
data {
  int<lower=0> N;                    // number of observations
  int<lower=0> N_new[3];
  vector[N] bill_length;              // response variable
  vector[N] bill_width;               // covariate
  vector[N] sex;                      // sex indicator
  int<lower=1> N_species;             // number of unique species
  int<lower=1, upper=N_species> species[N];  // species indicator
  int<lower=1> N_sex;                 // number of unique sexes
  int<lower=1, upper=N_sex> sex_id[N];      // sex indicator
}

parameters {
  vector[N_species] gamma_species;       // species-specific intercepts
  vector[N_species] beta_width_species;  // species-specific intercept for bill width


  vector[N_sex] gamma_sex;               // sex-specific intercepts
  real<lower=0> sigma;                   // standard deviation of the residuals
}

transformed parameters {
  vector[N] mu;  // linear predictor
  for (i in 1:N) {
    mu[i] = gamma_species[species[i]] + beta_width_species[species[i]] * bill_width[i] +
    gamma_sex[sex_id[i]];
  }
}

model {
  // Priors
  sigma ~ gamma(29.90633, 1);

  // Species priors
  for (j in 1:N_species) {
    gamma_species[j] ~ normal(20, 80);
    beta_width_species[j] ~ normal(0.79, 2);
  }

  // Sex priors
  for (k in 1:N_sex) {
    gamma_sex[k] ~ normal(20, 80);
  }

  // Likelihood
  bill_length ~ normal(mu, sqrt(sigma));
}

generated quantities {
  vector[N] log_lik;
```

```
  real ynew[N];

  for (i in 1:N) {
    log_lik[i] = normal_lpdf(bill_length[i] | mu[i], sqrt(sigma));
  }

  for (n in 1:N) {
    ynew[n] = normal_rng(mu[n], sqrt(sigma));
  }
}
```

## A.5 Hamiltonian Monte Carlo Plots

### A.5.1 Hierarchical Model

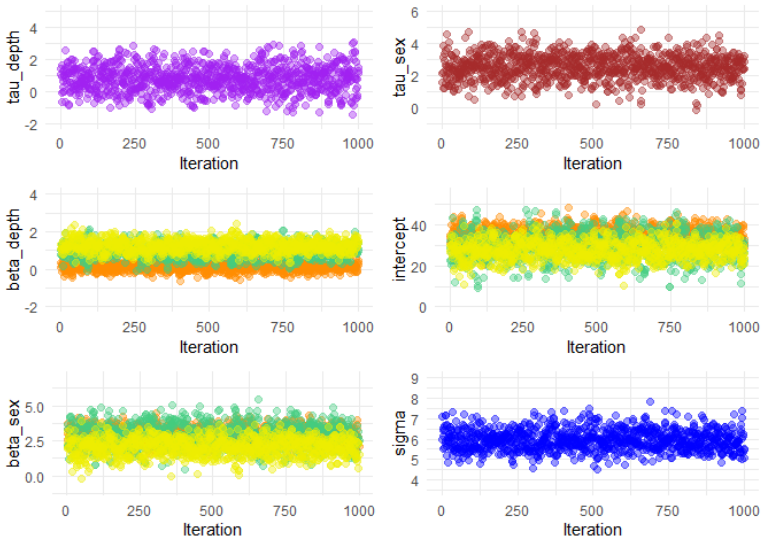



Figure 20: Uni-dimensional parameters space exploration in the Hierarchical model for Species. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

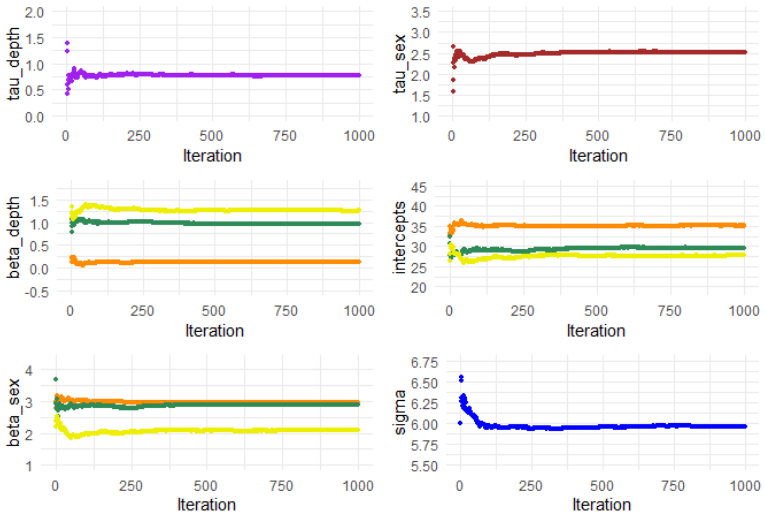



Figure 21: Convergences of means of the parameters in the hierarchical model for Species. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

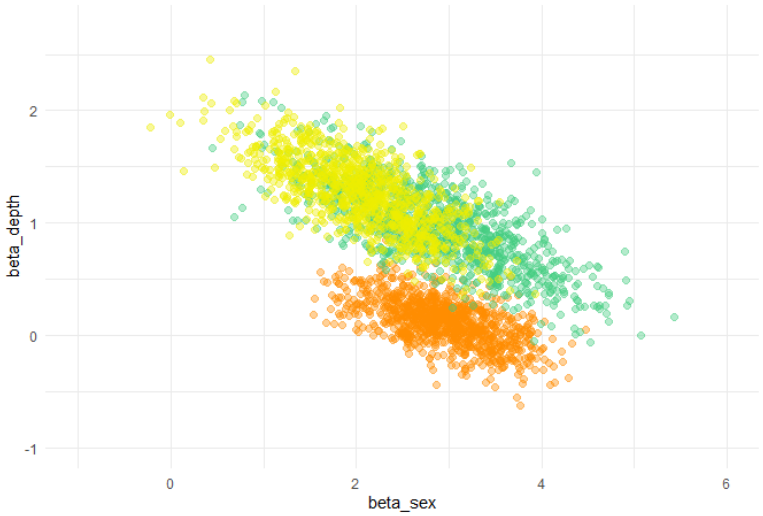



Figure 22: Two-dimensional parameter space exploration by the parameter Beta_sex and Beta_depth. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

### A.5.2 Separate Model for Species and Sex

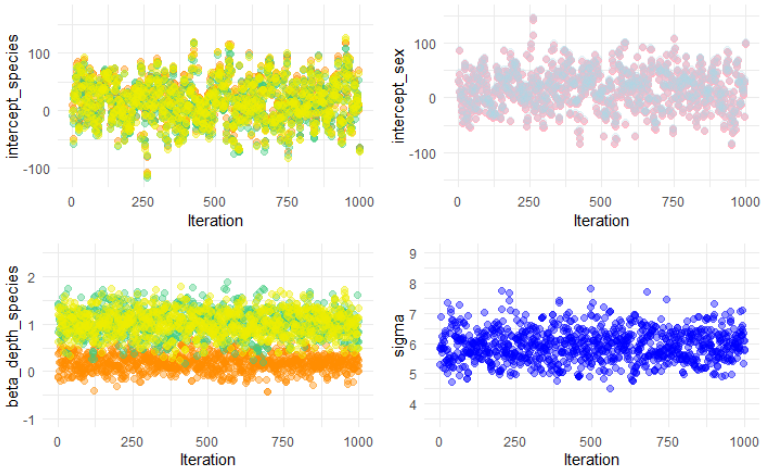



Figure 23: Uni-dimensional parameters space exploration in the Separate model for Species and Sex. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

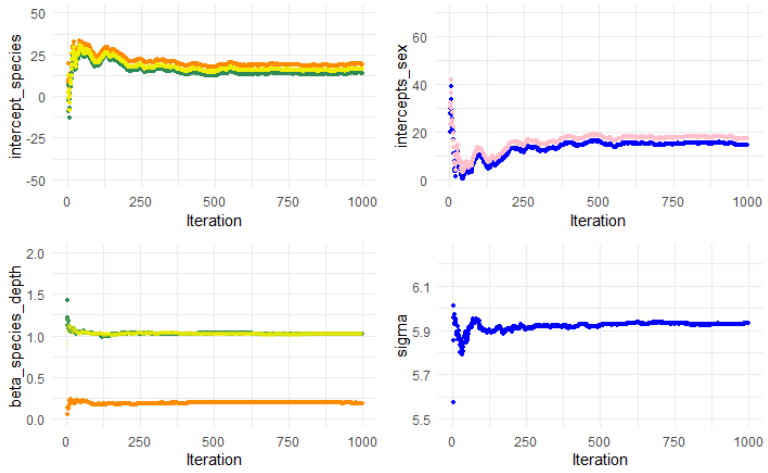



Figure 24: Convergences of means of the parameters in the Separate model for Species and Sex. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.


Figure 25: Two-dimensional parameter space exploration by the parameter Beta_sex and Beta_depth. Species: *Adélie* in orange, *Gentoo* in yellow, *Chinstrap* in green.

## A.6 Separate Model for Species and Sex before variable selection

```
// The input data is a vector 'y' of length 'N'.
data {
  int<lower=0> N;                    // number of observations
  int<lower=0> N_new[3];
  vector[N] bill_length;              // response variable
  vector[N] bill_width;               // covariate
  vector[N] sex;                      // sex indicator
  int<lower=1> N_species;             // number of unique species
  int<lower=1, upper=N_species> species[N];  // species indicator
  int<lower=1> N_sex;                 // number of unique sexes
```

```
  int<lower=1, upper=N_sex> sex_id[N];      // sex indicator
}

parameters {
  vector[N_species] gamma_species;       // species-specific intercepts
  vector[N_species] beta_width_species;  // species-specific intercept for bill width
  vector[N_species] beta_sex_species;    // species-specific intercept for sex

  vector[N_sex] gamma_sex;               // sex-specific intercepts
  vector[N_sex] beta_width_sex;          // sex-specific intercept for bill width
  vector[N_sex] beta_species_sex;        // interaction term between species and sex
  real<lower=0> sigma;                   // standard deviation of the residuals
}

transformed parameters {
  vector[N] mu;  // linear predictor
  for (i in 1:N) {
    mu[i] = gamma_species[species[i]] + beta_width_species[species[i]] * bill_width[i] +
            gamma_sex[sex_id[i]] + beta_width_sex[sex_id[i]] * bill_width[i] +
            beta_sex_species[species[i]] * sex[i] + beta_species_sex[sex_id[i]] * species[i];
  }
}

model {
  // Priors
  sigma ~ gamma(29.90633, 1);

  // Species priors
  for (j in 1:N_species) {
    gamma_species[j] ~ normal(20, 80);
    beta_width_species[j] ~ normal(0.79, 2);
    beta_sex_species[j] ~ normal(2.5, 2);
  }

  // Sex priors
  for (k in 1:N_sex) {
    gamma_sex[k] ~ normal(20, 80);
    beta_width_sex[k] ~ normal(0, 10);
    beta_species_sex[k] ~ normal(0, 10);
  }

  // Likelihood
  bill_length ~ normal(mu, sqrt(sigma));
}

generated quantities {
  vector[N] log_lik;
  real ynew[N];
  for (i in 1:N) {
    log_lik[i] = normal_lpdf(bill_length[i] | mu[i], sqrt(sigma));
  }
```

```
  for (n in 1:N) {
    ynew[n] = normal_rng(mu[n], sqrt(sigma));
  }
}
```