Aalto university

CS-E5710 - Bayesian Data Analysis D

BDA Project
Bayes analysis on acupuncture for Chronic headache

Daud Abucar, 1021842, daud.abucar@aalto.fi

Imose Iduozee, 1035618, imose.iduozee@aalto.fi

Edris Hakimi, 929327, edris.hakimi@aalto.fi

Contents

1 Introduction

Newborn mortality is a large concern in public health, serving as the leading reason of all child deaths among children under the age of five. Millions of newborn deaths occur yearly, and while this number has decreased since 1990, there are still challenges that remain, especially in low-income countries where quality healthcare is not as accessible [4]. Identifying and understanding factors that increase newborn mortality is crucial as it enhances prevention strategies to combat this problem.

1.1 Problem

Many factors contribute to newborn mortality. Changes in variables such as body weight and platelet count are common indicators used to predict newborn survival. For example, if body weight falls below a certain threshold, mortality becomes highly probable. In this project, we aim to examine whether these threshold values differ by gender. The variable sex is binary, where 0 represents male infants and 1 represents female infants. It is well-known that biological differences exist between males and females. Our analysis begins by creating two models: one without considering sex as a factor, and another with sex as a group effect.

1.2 Outline

In Section 1, we introduce the topic of newborn mortality and explain the motivation behind our project. Section 2 will present our dataset and explore the variables involved in the analysis. In Section 3, we will describe the Bayesian analysis methods used and assess the performance of our models. Section 4 will provide the conclusions drawn from our findings, while Section 5 will reflect on the lessons learned throughout the project.

2 Data

The dataset used in this project is the Very Low Birth Weight Infants Dataset, which is publicly available [3]. It consists of data from 671 infants with very low birth weights, collected at Duke University Medical Center by Dr. Michael O'Shea between 1981 and 1987. The dataset examines the relationship between intraventricular hemorrhage (IVH) and factors such as birth weight, gestational age, pneumothorax, delivery method, and whether the infant was born at Duke or transferred there later.

For our analysis, we downloaded the dataset in SAV format, imported it into R, and cleaned it to handle missing data. After cleaning, we used a final subset of 174 rows, as many records had incomplete information.

Variable	**Description**
PltCt	Platelet amount
BWt	Birth weight of infant
Pneumo	Pneumothorax occurred
Birth	Date of birth (admission)
Exit	Date of death
HospStay	Length of hospital stay
LowPh	Lowest pH in first 4 days of life
Race	Race of infant
Gest	Gestational age
InOut	Infant born in Duke hospital or transferred to Duke hospital after birth
Twn	Multiple gestation
Lol	Duration of labor
Meth	Mother treated with beta-methasone
Toc	Tocolysis - mother treated with beta-adrenergic drug
Delivery	Delivery method
Apg1	Apgar at one minute
Vent	Ventilation assistance used
Pda	Patent ductus arteriosus detected
Cld	On supplementary oxygen at 30 days
Pvh	Has periventricular hemorrhage occurred
Ivh	Has intraventricular hemorrhage occurred
Ipe	Has intraparenchymal echodensity occurred
Year	Year of birth (+fraction)
Sex	Sex of infant
Death	Did infant die

Table 1: Variable Descriptions for the Dataset

It is important to note that this dataset specifically focuses on infants with existing health concerns related to intraventricular hemorrhage (IVH), with the primary goal of understanding the correlation between this condition and sex. As such, the dataset cannot be used to infer the probability of IVH in the general infant population. Additionally, given the data's collection period (1981-1987), it may be considered outdated for modern research, as healthcare practices have significantly advanced since then. However, due to the large and diverse sample size, we assume that the data is independent and identically distributed (i.i.d.), with minimal influence from genetic factors or localized conditions. It is also worth noting that no prior Bayesian analyses have been found on this specific dataset.

2.1 Data Analysis

Our data contains the 25 features. First task is to select the best features. for this we plotted correlation matrix and pruned the features with high multicollinearity with VIF. Note that the variable *sex* is coded as 0 for male infants and 1 for female infants.

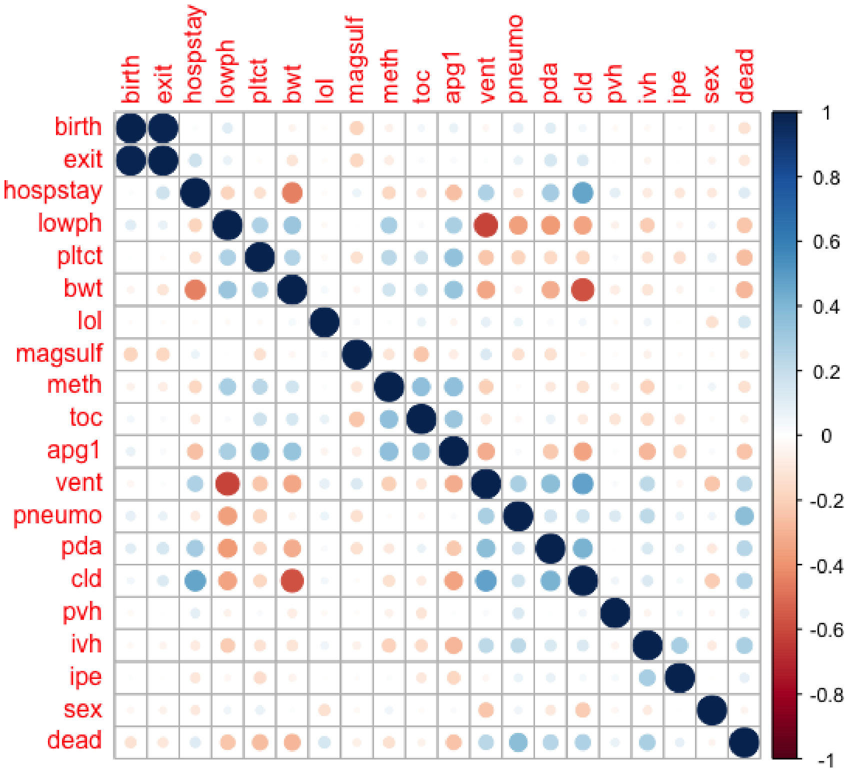

Figure 1: Correlation of all the variables

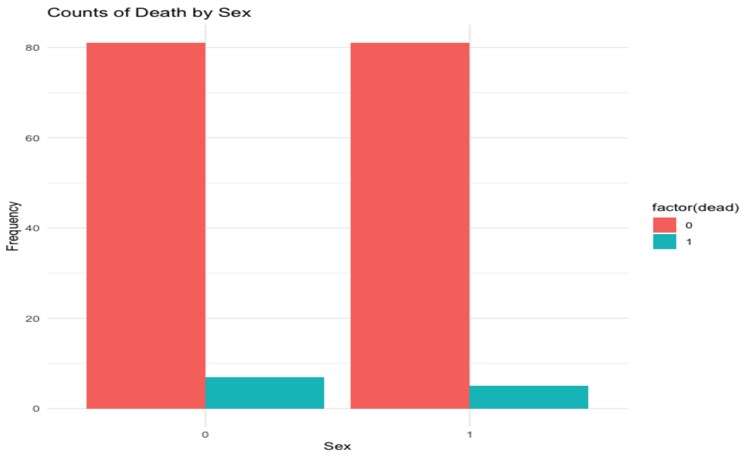

Figure 2: Counts of dead groubed by sex

After feature selection we chose *PltCt*, *BWt*, *IVH* and *Pneumo* to predict whether newborn dies after birth. Also, the histogram of these features are plotted below.

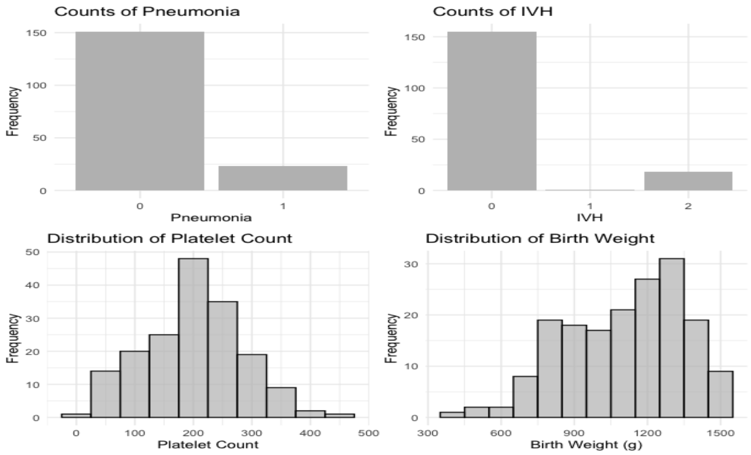

Figure 3: Distributions of the selected features

In Figure 1, the correlations between each variable are displayed. Figure 2 shows the distribution of the outcome variable, dead, and Figure 3 illustrates the distribution of the selected features.

The features were chosen based on two criteria: first, that there is no significant multicollinearity (VIF $< \sqrt{5} \approx 2$), and second, that the absolute value of the Pearson correlation is greater than 0.25.

3 Model

As discussed earlier two model are implemented: one with sex as a predictor and group-level effect, and one without sex. In general, mathematically notated, the non-hierarchical model:

$$dead \sim Bernoulli(p_i),$$

$$logit(p_i) \sim PltCt + BWt + IVH + Pneumo,$$

and hierarchical model:

$$logit(p_i) \sim (PltCt + BWt + IVH + Pneumo|sex)$$

In the first model, the information about sex is entirely omitted, which results in a pooled model where the effect of sex is assumed to be the same for all observations. The second model (hierarchical model) includes sex as a group-level effect. This allows for different baseline probabilities of the outcome dead for each group (male and female).

The Bernoulli distribution is used for the outcome variable, as it is a binary variable: 1 for deceased and 0 for alive. The logit function is chosen as the link function because the predictors can take values outside the [0, 1] range. The logit function scales the linear combination of predictors back to the [0, 1] interval, making it suitable for modeling probabilities.

3.1 Priors

3.1.1 The Effect of Platelet Count on Mortality

Platelets are small cells in the blood that help with clotting, which stops bleeding. When a newborn has a low platelet count called thrombocytopenia, or if their platelet count drops significantly by 30% or more, their risk of dying increases. This is especially true for preterm babies. Low platelet levels are often linked to serious conditions like sepsis (a dangerous blood infection) and necrotizing enterocolitis (a severe intestinal problem) [1].

How the Prior Was Chosen:

Average Effect Size ($\mu = -0.3$): The value -0.3 means that low platelet counts moderately increase the risk of death. This number comes from research studies that looked at how often platelet drops led to bad outcomes. Studies show that the odds of death can increase by 30%–50% in such cases. The prior reflects this finding.

Uncertainty ($\sigma = 0.15$): The uncertainty or variability is set to 0.15 to account for differences between hospitals or NICUs (Neonatal Intensive Care Units). For example, in some settings, babies might get faster or better treatments, which could reduce the risk. In others, the risk might be higher due to delays or fewer resources. The prior accounts for these differences.

Prior:

$$N(\mu = -0.3, \sigma = 0.15)$$

3.1.2 The Effect of Birth Weight

Birth weight is one of the strongest predictors of neonatal survival. Lower birth weights are consistently linked to higher risks of death, but instead of using fixed categories (e.g., < 2500 g), we model weight as a continuous variable.

How the Prior Was Chosen:

Studies show that as birth weight decreases, the mortality risk increases significantly. For each 500 g decrease below the average weight (around 3000 g for term infants), the risk can rise dramatically [2].

To Reflect This:

- $\mu = -0.5$: Represents the strong average effect of low birth weight on mortality.
- $\sigma = 0.25$: Accounts for variability due to medical interventions and individual differences among newborns.

Prior:

$$N(\mu = -0.5, \sigma = 0.25)$$

This prior integrates the effect of weight as a continuous variable, meaning the model adjusts for the full range of possible weights.

3.1.3 The Effect of Pneumonia

Pneumonia remains the leading infectious cause of death in children under 5, responsible for over 740,000 deaths in 2019 [5]. Pneumonia is especially dangerous for babies with low birth weights or compromised immune systems.

How the Prior Was Chosen:

The average effect ($\alpha = 2, \beta = 8$) reflects the strong impact pneumonia has on mortality but also the rarerity of the condition.

Prior:

$$\text{Beta}(\mu = 2, \sigma = 8)$$

3.1.4 The Effect of IVH

Based on the link and related research, severe IVH significantly increases neonatal mortality rates, particularly in very low birth weight infants [**?**]. A reasonable weakly informative prior for IVH's effect on mortality could be:

Prior:

$$N(\mu = 0, \sigma = 1)$$

3.1.5 Priors Table

Predictor	**Prior**	**Explanation**
Platelet Count	$N(\mu = -0.3, \sigma = 0.15)$	Moderate risk increase with platelet drops [1].
Birth Weight	$N(\mu = -0.5, \sigma = 0.25)$	Weight modeled as continuous; lower weight increases risk [2].
Pneumonia	$\text{Beta}(\mu = 2, \sigma = 8)$	Pneumonia is a leading cause of child mortality [5].
IVH	$N(\mu = 0, \sigma = 1)$	[?]

Table 2: Prior Distributions and Explanations

3.2 BRMS

The model for hierarchical and non-hierarchical is specified in the code below. The priors are implemented as instructed above. After running the code, the model succesfully converged.

```
priors = c(
  prior(normal(-0.3, 0.15), coef = "pltct"),
  prior(normal(-0.5, 0.25), coef = "bwt"),
  prior(beta(2, 8), coef = "pneumo"),
  prior(normal(0, 1), coef = "ivh")
)

# Fit the model
fit_dead <- brm(
  dead ~ pltct + bwt + pneumo + ivh,
  data = cath,
  family = bernoulli(),
  prior = priors,
  chains = 4,
  iter = 2000
)

priors_hierarchical <- c(
  prior(normal(0, 0.01), coef = "pltct"),
  prior(normal(-0.1, 0.05), coef = "bwt"),
  prior(beta(2, 8), coef = "pneumo"),
  prior(normal(0, 1), coef = "ivh"),
  prior(exponential(2), class = "sd", group = "sex")
)

fit_sig_hier <- brm(
  dead ~ pltct + bwt + pneumo + ivh + (1 | sex),
  data = cath,
  family = bernoulli(),
  prior = priors_hierarchical,
  chains = 4,
  iter = 2000,
  control = list(adapt_delta = 0.95)
)
```

3.3 Convergence

Table 3: Non-Hierarchical Regression Coefficients

Predictor	**Estimate**	**Est.Error**	**l-95% CI**	**u-95% CI**	**Rhat**	**Bulk_ESS**	**Tail_ESS**
Intercept	2.42	1.47	-0.36	5.38	1.00	3435	3059
pltct	-0.01	0.00	-0.02	-0.00	1.00	2733	2423
bwt	-0.00	0.00	-0.01	-0.00	1.00	3450	2816
pneumo	0.26	0.14	0.05	0.56	1.00	2432	1907
ivh	0.74	0.36	0.03	1.41	1.00	3006	2732

Table 4: Hierarchical Model Results

Multilevel Hyperparameters	**Estimate**	**Est.Error**	**l-95% CI**	**u-95% CI**	**Rhat**	**Bulk_ESS**	**Tail_ESS**
sd(Intercept)	0.38	0.38	0.01	1.39	1.00	1457	1261

Regression Coefficients	**Estimate**	**Est.Error**	**l-95% CI**	**u-95% CI**	**Rhat**	**Bulk_ESS**	**Tail_ESS**
Intercept	2.39	1.53	-0.48	5.41	1.00	4137	3237
pltct	-0.01	0.00	-0.02	-0.00	1.00	3733	2746
bwt	-0.00	0.00	-0.01	-0.00	1.00	4044	2990
pneumo	0.27	0.14	0.04	0.57	1.00	2022	2384
ivh	0.76	0.36	0.05	1.48	1.00	2883	2423

Both models converged with no diverging transitions. The effective sample size was high for all parameter values, which indicates that the sampling process was efficient and provided reliable estimates. Additionally, the Rhat values were 1 for all population-level effects, which confirms that the chains have converged properly.

3.4 Posterior predictive

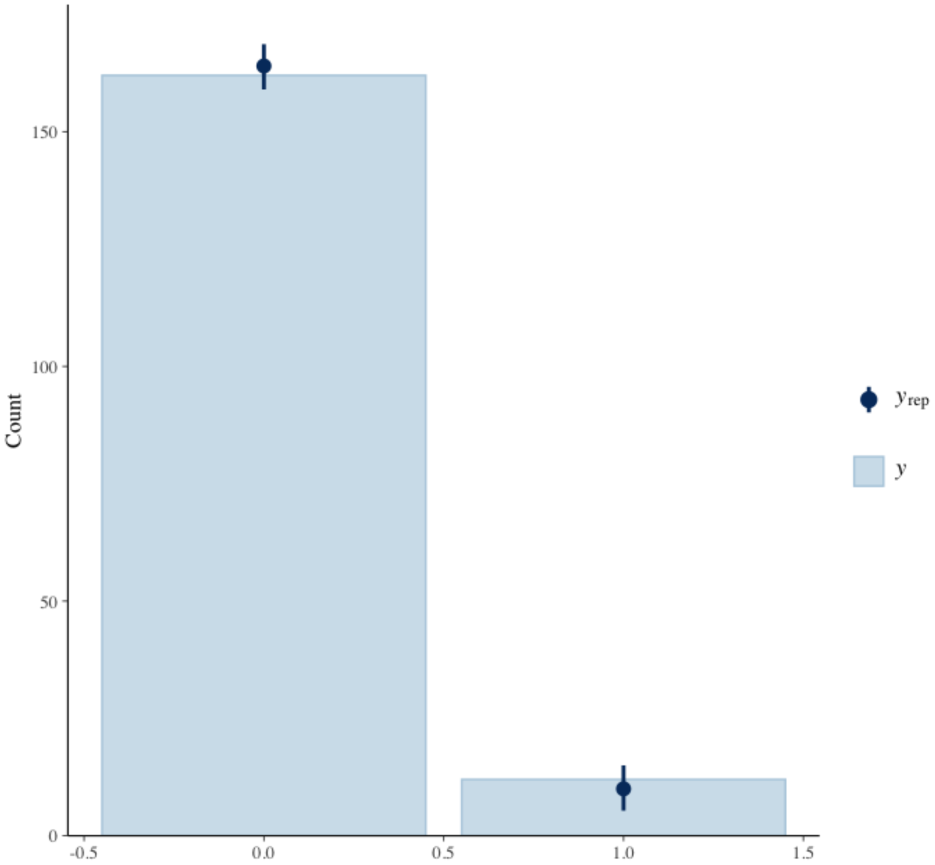

Figure 4: Posterior predictive check of the non-hierarchical model

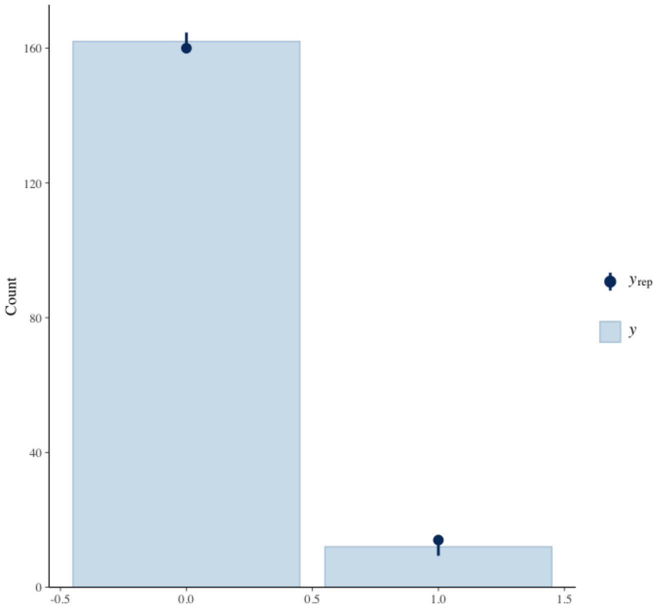

Figure 5: Posterior predictive check of the hierarchical model

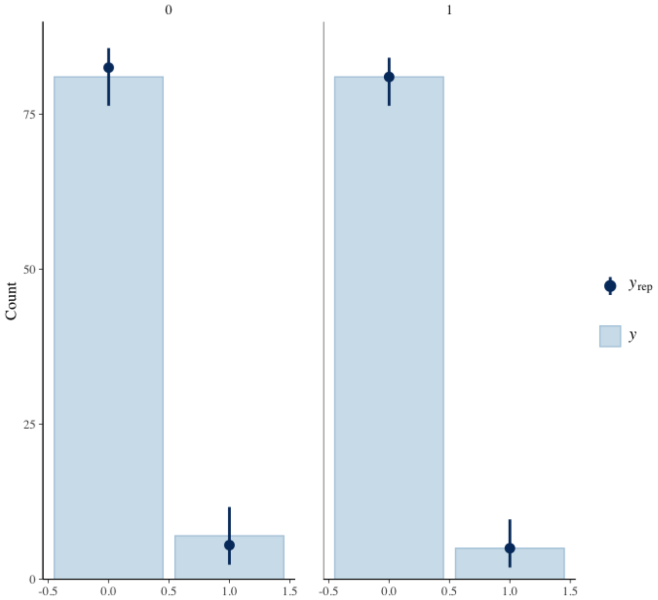

Figure 6: Posterior predictive check of the hierarchical model grouped by sex

In the non-hierarchical model (figure 4), the predictive values slightly overestimated the number of survivors (dead = 0), with a relatively large variance. Vice-verse, the number of deceased infants (dead = 1) was also predicted with nearly the same variance. This suggests some degree of imprecision in the model's predictions, particularly in estimating the number of survivors.

The hierarchical model (figure 5) predicts closer to the real value. However, when performing a posterior predictive check with groubed effect (figure 6), we observed that the predictive data closely matched the observed data for females. Unfortunately, for males, the model overestimated the number of survivors and underestimated the number of deceased infants.

3.4.1 Posterior predictive Analysis

Despite these discrepancies, the classification accuracy for both models (non-hierarchical and hierarchical) was approximately 0.93, indicating good overall predictive performance.

Table 5: Leave-One-Out Cross-Validation (LOO) Non-hierarchical

Statistic	**Estimate**	**SE**
elpd_loo	-34.4	7.0
p_loo	4.5	1.3
looic	68.7	14.0

Table 6: Leave-One-Out Cross-Validation (LOO) Hierarchical

Statistic	**Estimate**	**SE**
elpd_loo	-34.6	7.1
p_loo	4.9	1.3
looic	69.2	14.2

The PSIS diagnostic plot shows that the Pareto K values are all below 0.7, confirming that there are no problematic outliers in the models.

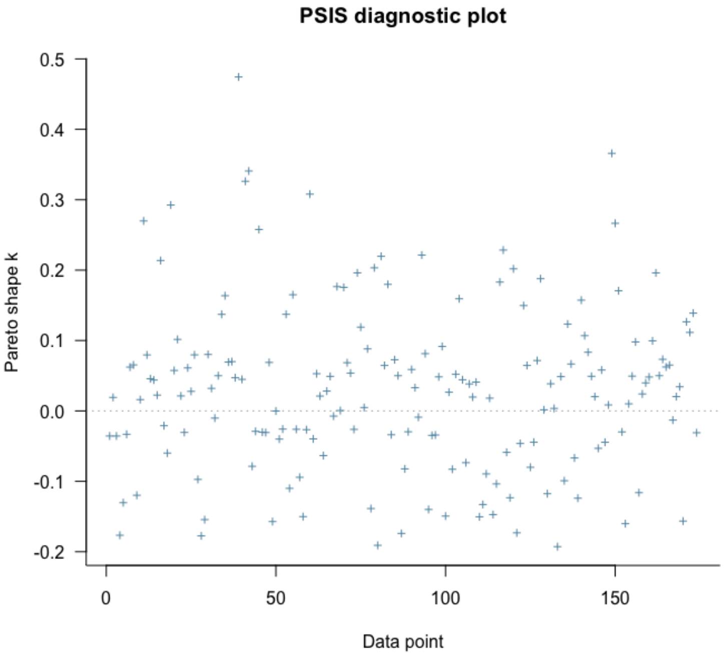

Figure 7: PSIS diagnostic plot(non-hierarchical)

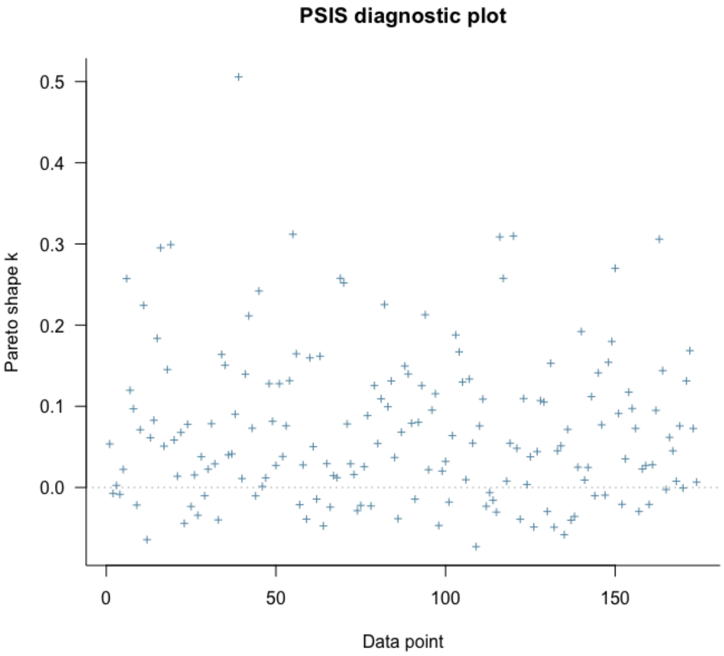

Figure 8: PSIS diagnostic plot(hierarchical)

While the model shows some tendency to overestimate survivors, it's important to note that we used the same dataset for both training and testing. As a result, we cannot definitively assess the presence of overfitting in the model, as the model may have been over-optimized on the data it was evaluated on.

To compare both models, we used Leave-One-Out Cross-Validation (LOO). The Pareto k estimates for both models were found to be less than 0.7, indicating the absence of influential outliers and suggesting that the models are well-behaved in terms of outlier influence.

Table 7: Model Comparison

Model	**elpd_diff**	**se_diff**
fit_dead	0.0	0.0
fit_dead_hier	-0.2	0.4

However, when comparing the models using the expected log pointwise predictive density (elpd-LOO), the difference between the models was minimal. Given that the difference is smaller than the threshold for meaningful comparison ($|\text{edpd\_diff}| < 4$), we cannot definitively choose between the two models based on elpd-LOO.

4 Conclusion

Through this project, we aimed to identify the main factors leading to newborn mortality and analyze how gender influences mortality rates. We tested our hypothesis using both hierarchical and non-hierarchical models, but there were no significant differences in their results.

Our analysis identified pneumothorax, IVH, platelet count, and birth weight as the strongest predictors of mortality. Other variables either had insignificant correlations or were excluded due to high multicollinearity.

Contrary to our hypothesis, gender did not significantly influence newborn mortality. This is because biological differences between males and females are not prominent during early development stages. These differences typically emerge later in life, particularly during puberty.

5 Reflection

This project deepened our understanding of conducting a bayesian analysis project. One of the primary challenges was identifying and narrowing a suitable topic to analyze. Defining and applying priors based on existing literature was definitely the most demanding part of this project. Additionally, we also encountered challenges with calculations when performing LOO cross-validation and conducting posterior predictable checks. Through these challenges we were able to develop our analysis skills and gain more insight about this course's topics.

References

[1] Abeer Abd Elmoneim, Mohammed Zolaly, Ehab Abd El-Moneim, and Eisa Sultan. Prognostic significance of early platelet count decline in preterm newborns. *Indian journal of critical care medicine: peer-reviewed, official publication of Indian Society of Critical Care Medicine*, 19(8):456, 2015.

[2] Aimin Chen, Shingairai A Feresu, Cristina Fernandez, and Walter J Rogan. Maternal obesity and the risk of infant death in the united states. *Epidemiology*, 20(1):74–81, 2009.

[3] Michael O'Shea, David A Savitz, Marvin L Hage, and KA Feinstein. Prenatal events and the risk of subependymal/intraventricular haemorrhage in very low birthweight neonates. *Paediatric and perinatal epidemiology*, 6(3):352–362, 1992.

[4] Jayani Pathirana, Flor M Muñoz, Victoria Abbing-Karahagopian, Niranjan Bhat, Tara Harris, Ambujam Kapoor, Daniel L Keene, Alexandra Mangili, Michael A Padula, Stephen L Pande, et al. Neonatal death: Case definition & guidelines for data collection, analysis, and presentation of immunization safety data. *Vaccine*, 34(49):6027–6037, 2016.

[5] World Health Organization. Pneumonia, 2023.