DEPARTMENT OF COMPUTER SCIENCE

Controlling for Multiple Comparisons

A Comparative Study of Statistical Models with an Emphasis on the Bayesian Multilevel Model

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Bachelor of Science in the Faculty of Engineering.

May 3, 2023

Abstract

The issue of multiple comparisons is a significant problem in statistical analyses and arises as a consequence of several hypothesis tests being conducted simultaneously. The chance of at least one Type I error among the tests increases as more hypotheses are tested simultaneously. To adjust for this problem, frequentist approaches including the Bonferroni correction [3] and the Benjamini-Hochberg procedure [5] have been suggested. However, the conservative nature of them leads to an increase in the number of Type II errors.

In this project, I investigate how a Bayesian multilevel model [14] can perform a multiple comparisons correction while still maintaining a low Type II error rate. My research hypothesis is that the incorporation of Bayesian multilevel modelling techniques will result in a more effective approach to control for multiple comparisons, as compared to traditional frequentist methods, due to accounting for group-level variance and correlation.

- I wrote the Bayesian multilevel model [14] in Stan and used it to fit data from The Infant and Health Development Program (IHDP) [18] in R. I also fitted a linear regression with frequentist corrections and compared the models' error rates.
- I simulated data based on data from the IHDP but with more and less variation between sites and found that it is more important to choose the Bayesian multilevel model when there is more variation between sites.
- I compared models with simulated data in which correlation had been added and investigated how well a Bayesian multilevel model that incorporates this correlation performs.
- I wrote over 1000 lines of R / Stan code in total

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

[redacted] May 3, 2023

Contents

List of Figures

List of Tables

Ethics Statement

A compulsory section

This project did not require ethical review, as determined by my supervisor, Dr Conor Houghton.

Supporting Technologies

- I used R (4.2.3) in order to fit all the models and run all the simulations.
- I used Stan (2.21.0) to create the Bayesian multilevel model with and without correlation included.
- I used the University of Bristol HPC, BluePebble, to run some of the more computationally extensive simulations.
- I used the following packages within R:
  - foreign (0.8-84) to read the .xpt file of the IHDP data
  - dplyr (1.1.1) to group the IHDP data
  - broom (1.0.4) to summarise linear regression fits to the data
  - sgof (2.3.3) to perform the Benjamini-Hochberg procedure
  - rstan (2.21.8) to compile the models from a Stan file and produce posterior samples of the fit
  - shinystan (2.6.0) to access convergence diagnostics of the fitted Bayesian models
  - HDInterval (0.2.4) to calculate the High Posterior Density intervals
  - ggplot2 (3.4.2) to plot all of my graphs
  - ggpubr (0.6.0) and gridExtra (2.3) to arrange my graphs
  - MASS (7.3-58.2) to produce samples from a multivariate normal distribution

Notation and Acronyms

CI	:	Confidence Interval
HDI	:	Highest Posterior Density Interval
FDR	:	False Discovery Rate
FWER	:	Family-wise Error Rate
B-H	:	Benjamini-Hochberg
H_0 vs H_1	:	Null Hypothesis vs Alternative Hypothesis
IHDP	:	Infant Health and Development Program
MCMC	:	Markov Chain Monte Carlo
HMC	:	Hamiltonian Monte Carlo

Chapter 1

Introduction

In a statistical analysis, researchers often come across a situation in which multiple groups are compared simultaneously or multiple hypothesis tests are carried out [9]. However, the more tests there are or the more groups there are to be compared, the higher the chance of at least one false positive among the tests. A false positive suggests an effect when there is none and can lead to incorrect conclusions or time wasted to conduct further tests. This is known as the multiple comparisons problem and it has significant implications on the reliability of research results.

1.1 Motivating Examples

This is a prevalent issue that arises in a variety of contexts, including in medical research [16] [32] and social studies [2]. The following examples provide an introduction to some of these contexts.

1.1.1 Clinical Trial

Suppose that we would like to find the best possible treatment for a disease and we decide to test drug A and drug B. The patients with the disease are randomly separated into two groups, g_A and g_B, receiving drugs A and B respectively. In order to determine which drug performs more effectively we conduct a hypothesis test to compare the two groups. This involves $H_0 : \mu_A = \mu_B$ and $H_1 : \mu_A \neq \mu_B$ where μ_K is the mean of g_K's health after a fixed period of time. We test this at a 5% significance level such that if we were to repeat this test a large number of times, we would expect 1 in 20 (5%) of the results to return a false positive [20]. This is not a particularly large error rate but it is necessary in order to accept a certain level of uncertainty.

Suppose that we would now like to test $n = 5$ different drugs and compare them all. There would then be $\frac{n(n-1)}{2} = \frac{5 \cdot 4}{2} = 10$ pairwise comparisons. If each one was tested at a 5% significance level, then each one would result in a false positive 5% of the time. As a consequence, for any one of them to return a false positive, the chance is not so low:

$$\begin{aligned}
\mathbb{P}(\text{at least one false positive}) &= 1 - \mathbb{P}(\text{no false positives for all tests}) \\
&= 1 - \underbrace{(1 - 0.05)...(1 - 0.05)}_{\text{\# of tests times}} \\
&= 1 - (1 - 0.05)^{10} \\
&= 0.401
\end{aligned}$$

If we were only comparing one more drug and increase n to 6, we would introduce $n - 1 = 5$ more comparisons and thus increase the probability of at least one false positive even more considerably. With 10 drugs, we would end up with 45 comparisons, resulting in a large 0.901 chance.

1.1.2 Genetics

In genetic research, it is common for scientists to conduct association studies in which they examine if genetic variants are connected to a certain disease and determine what their specific impacts are [32].

This inevitably requires many tests to be carried out in order to estimate the association of the wide range of variants, thus leading to the multiple comparisons problem.

This problem is a particularly significant issue in large-scale genetic association studies, in which a large number of genetic variants may exist in a genome and the underlying architecture of a disease or trait may not be well understood. Therefore, it is quite unlikely that any single genetic variant is associated with a particular disease. Thus, we need to account for the multiple comparisons in order to prevent false positives.

1.1.3 Politics

We may want to compare the mean political ideology of Democrats and Republicans, ranging from 1 being extremely conservative to 7 being extremely liberal [2]. This involves comparing the two parties. However, as soon as we introduce more parties into consideration, we encounter more pairwise comparisons.

1.1.4 Infant Health and Development Programme

The Infant Health and Development Program (IHDP) [18] was an eight-site study in which 1090 low birth weight, premature infants received early intervention over three years to determine how effective these techniques are. The infants were randomly assigned to two groups: the Follow-up Group and the Intervention Group. The Follow-up Group received a paediatric follow-up only, while home visits and attendance at a special child development centre were included for the Intervention Group as well. The follow-ups led to test scores for a range of measures including analysing cognitive development, health status and other variables in order to assess the effect of the intervention. Since this study was conducted across eight sites, determining whether the treatment was effective for each one gives rise to the issue of multiple comparisons.

1.2 Approaches to Finding a Solution

The issue of multiple comparisons has been recognised as a problem in a variety of contexts, including those described above. From Sjölander and Vansteelandt's experience [28], it is generally believed among researchers that an adjustment is sometimes, if not always, needed. However, while many approaches in the classical setting have been suggested to address the issue, it is not clear what the ideal method is.

This project intends to look at two of these classicalprocedures that aim to control different error rates. The Bonferroni correction directly controls the family-wise error rate (FWER), the probability of at least one false positive [3]. It does this by dividing the significance level by the number of tests performed, making it more difficult to detect a statistically significant result. As an alternative, I look at the Benjamini-Hochberg (B-H) procedure which controls the false discovery rate (FDR) [5]. This is an approximation of the FWER but by controlling it, we allow for more false positives.

These approaches, especially the Bonferroni correction, have the limitation of producing an influx of false negatives. This is due to the correction making it hard to spot a positive result, even with a strong, true effect. Consequently, with enough comparisons, a researcher may feel pressure to manipulate the presentation of the data to reduce the apparent number of comparisons and increase the adjusted significance level, for example, by combining several groups into a single group. This can introduce bias and weaken the validity of the conclusions.

I investigate the Bayesian multilevel model put forward by Gelman [14] and how it naturally provides a solution by producing estimates that are shrunk towards each other and have more certainty. In this way, it reduces the number of false positives and false negatives. I look at the performance of the different corrections and the Bayesian model based on various error rates when fitted to data from the IHDP. Furthermore, I explore how adaptations of the Bayesian model can alter how well it fits a variety of data. More specifically, the concrete objectives are as follows:

- Research the classical and Bayesian approaches and how they correct for the numerous false positives.
- Research the limitations of the classical techniques in terms of the false negatives that are produced.
- Use simulations to explore each model's performance on the IHDP data based on various error rates.

- Conduct further simulation studies to explore how the Bayesian model performs and how it can be adapted when fitting data in which there is:
  - more or less variance between sites.
  - correlation between parameters.

Chapter 2

Background

In this section, I explore the frequentist and Bayesian approaches to multiple comparisons and how they differ. This is done in the context of data from the IHDP, as mentioned in the Introduction, and using models put forward by Gelman [14].

2.1 Frequentist Approach

The frequentist perspective of statistics is a classical view that interprets the long-term frequency of events as probabilities [25]. It involves the use of hypothesis testing, as mentioned in 1.1.1, to make inferences about population parameters based on some sample data [20].

In a hypothesis test, the null hypothesis, H_0, is assumed to be true and a statistical test can be conducted with the sample data to determine the probability of rejecting H_0. This probability is known as the p-value. It introduces the significance level, α, which is fixed and chosen by the researcher. It is equal to the allowed probability of committing a Type I error in which we reject H_0 when it is in fact true. Therefore, it represents the false positive rate. If the p-value $\leq \alpha$, then there is sufficient evidence to reject the null hypothesis and a significant result is obtained. Otherwise, if the p-value $> \alpha$, the null hypothesis is assumed to be true.

An alternative way to determine whether H_0 should be rejected is by looking at the $100(1-\alpha)\%$ confidence interval (CI) [10], as determined by the significance level α. If the test was carried out multiple times, this represents the interval that contains the true parameter value $100(1-\alpha)\%$ of the time. If the null hypothesis falls outside of this interval, then H_0 is rejected.

In terms of multiple comparisons, frequentist statisticians work by controlling the error rates that arise as a consequence of multiple hypothesis testing [3] [5].

Classical Linear Regression Model

In order to fit data from the IHDP to a model from the frequentist perspective, a classical linear regression model can be used on each site separately [14]:

$$y_{ij} = \gamma_j + \delta_j P_{ij} + \epsilon_i \qquad \epsilon_i \sim \texttt{Normal}(0, \sigma^2) \tag{2.1}$$

$$P_i = \begin{cases} 1 & \text{if infant } i \text{ from site } j \text{ is in the Treatment Group} \\ 0 & \text{if infant } i \text{ from site } j \text{ is in the Follow-up Group} \end{cases}$$

For each site j, infant i's test score, y_{ij}, is modelled as an outcome of the indicator variable P_i that represents the Program Status. The intercept, γ_j, represents the average test score in site j for the infants not receiving the treatment in the Follow-up Group and δ_j gives the average effect of the treatment in site j. Randomness is introduce by ϵ_i to account for the variability in the relationship between the variables.

2.1.1 Family-wise Error Rate

As was shown earlier, the issue of multiple comparisons is due to the high probability of making at least one Type I error. This is known as the family-wise error rate (FWER) [14] and is directly controlled in common frequentist solutions [3] [1] [27]. It is a good option to control this measure in situations when

even one false positive has an incredibly costly outcome that we would want to avoid. For example, a false positive in a clinical trial could lead to a potentially harmful treatment being described as effective. Due to the significant consequences of giving someone this treatment, it would be crucial to avoid the possibility of this happening even once and so it would be sensible to control the FWER to minimise this probability.

Bonferroni Correction

The Bonferroni Correction controls the FWER by adjusting the significance level used for each individual test [3]. The desired significance level is divided by the number of tests performed simultaneously so that each test is performed at an $\frac{\alpha}{n}$ level. In the case of the IHDP data, and an original target false positive rate of $\alpha = 0.05$, the Bonferroni correction would result in a significance level of $\frac{0.05}{8} = 0.00625$. This results in a more appropriate expected FWER:

$$\text{FWER} = 1 - (1 - 0.00625)^8 = 0.0489 \approx 0.05 \tag{2.2}$$

As we have seen, it has the advantage of being a very simplistic and easy to implement solution. It is a very conservative method in that it is highly cautious when making inferences and does not want to say an alternative hypothesis is true unless it is very sure. The consequence of this is that there is greater uncertainty about the estimate and the null hypothesis is more likely to be accepted when it is in fact false. This gives a greater chance of a Type II error in which H_0 is accepted when it is false.

2.1.2 False Discovery Rate

We may also consider to adjust our tests in order to control the false discovery rate (FDR) [5]. This is the expected proportion of rejected tests (discoveries) that are false. It is more useful to control than the FWER in situations in which we have a few false positives among lots of statistically significant results and the cost of a false positive is relatively low as compared to the potential benefits of discovering true positives. In this case, we are more focused on having a large proportion of true significant results rather than preventing one or more false positives [30]. For example, it is a good option in genomics where thousands of tests are often conducted simultaneously [32]. In this case, it would be too strict to control the FWER and instead more appropriate to focus on the proportion of positives that are true.

Benjamini-Hochberg Procedure

There are various procedures to control the FDR [29] [6] including the common Benjamini-Hochberg (B-H) procedure [5]. The steps are as follows:

1. Suppose we have N tests to conduct simultaneously. Perform the test for each hypothesis and compute the p-values for each one.

2. Rank the p-values in ascending order. We will call these $P = \{p_{(1)}, p_{(2)}, ..., p_{(N)}\}$, corresponding to the hypotheses $H = \{H_{(1)}, H_{(2)}, ..., H_{(N)}\}$ where $p_{(i)}$ is the p-value for $H_{(i)}$ for all $i \in \{1, ..., N\}$.

3. For each $p_{(i)} \in P$, calculate the critical value with the formula:

$$c_i = \frac{i \cdot q}{N} \tag{2.3}$$

   where i is the rank and q is the desired FDR chosen by the person conducting the test.

4. Find the highest rank i for which the p-value, $p_{(i)}$, is smaller than its corresponding critical value, c_i. Call this rank i^*.

5. All p-values, $p_{(1)}, ..., p_{(i^*)}$, are said to be significant as they are less than or equal to $p_{(i^*)}$. All hypotheses, $H_{(1)}, ..., H_{(i^*)}$, are accepted

By using the BH procedure, the expected proportion of false discoveries among all discoveries is guaranteed to not exceed q. In this way, it controls the false positive rate, while maintaining a large number of true positives. Therefore, it is considered less conservative than the Bonferroni correction.

2.2 Bayesian Approach

The Bayesian approach to statistics uses a subjective interpretation of probability in which probabilities relate to our beliefs and not to real physical properties [13]. Therefore, parameters are thought to be random variables with a prior distribution representing the beliefs. Inferences about the parameters can be made by updating the priors using the observed data, resulting in the posterior distribution. This is done with Bayes rule in which we find the posterior probability of a parameter, θ, given the data, y. This is represented as $p(\theta|y)$ and is calculated as follows [13]:

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)} \tag{2.4}$$

In this formula, $p(\theta)$ represents our prior belief that we have about the parameter; $p(y|\theta)$ is the likelihood function which is a function of θ that gives the probability of a fixed y. The denominator is given by $p(y) = \sum_\theta p(\theta)p(y|\theta)$ in the discrete case or $p(y) = \int p(\theta)p(y|\theta)$ in the continuous case and represents the probability of the data across all values of the parameter. In this way, Bayes rule provides a framework to update our beliefs about a parameter θ based on new evidence, y.

However, the constant $p(y)$ is often difficult to calculate in practice and so summarising posteriors using purely mathematical methods may be too complex [31]. Therefore, we can use numerical techniques such as Markov Chain Monte Carlo (MCMC) algorithms to draw samples from a distribution and estimate the posterior [31]. While Monte Carlo methods typically draw random independent samples [24], MCMC uses a Markov chain such that a sample depends on its predecessor. In this way, the sequence eventually converges towards the target distribution. After a sufficient number of iterations to ensure convergence, the following states can then be thought of as samples from the posterior distribution which accurately represents the parameters.

2.2.1 Bayesian Multilevel Model

The Bayesian approach to multiple comparisons involves using a multilevel model so that information is shared between tests [14].

In the IHDP, the infants are taken over eight different sites which introduces some hierarchical structure to the data. Individuals in various sites may have access to different resources; for example, one site may have the money to hire more experienced staff and provide a higher quality of treatment. This leads to the possibility of some site-to-site variation.

A multilevel model takes the variation between sites as well as the whole set of data into account by considering two levels: the individual level and the site level [14].

At the individual level, the whole set of test scores, y_i, are assumed to be taken from a Normal distribution:

$$y_i \sim \texttt{Normal}(\gamma_{j[i]} + \delta_{j[i]}P_i, \sigma_y^2) \tag{2.5}$$

As with the linear model, γ_j, the average untreated score, and δ_j, the treatment effect, are site specific to account for differences between sites. Now, however, they are included in a single model with the whole group of infants and the index $j[i]$ is used to represent the site that infant i belongs to. The variance for y_i, σ_y^2, can be interpreted as the amount of noise and represents the variability of y that cannot be explained by the regression model.

To make the model multilevel, we include a site or group-level for the treatment effects, δ_j. They are modelled with a Normal distribution with a common mean, μ_δ and some group-level variance, σ_δ^2.

$$\delta_j \sim \texttt{Normal}(\mu_\delta, \sigma_\delta^2) \tag{2.6}$$

This allows us to capture the overall average effect the treatment has on an infant's health with μ_δ, along with the variation in the effect between the different sites with σ_δ^2. By modelling δ_j with a probability distribution, the model can estimate the distribution of site-level effects and share information across the different sites. This results in the estimates being shrunk towards each other. It is known as partial pooling and can be thought of as a compromise between complete pooling and no pooling [14]. Complete pooling is the case in which the separate sites are not even considered and the whole set of test scores is modelled with a singular γ and δ. The classical linear regression model in 2.1 is an example of no pooling since the test scores in each site are fitted to their own separate model, with each site using only their own information. However, with partial pooling, the model recognises that the sites are all using the same program and measuring the same thing. Therefore, considering the site itself, but also taking into

account information from the other sites seems appropriate to better estimate the effects. This structure allows for estimates of the treatment effect to be more accurate and closer together. This results in fewer false positives and reduces the issue of multiple comparisons.

How Much Shrinkage?

It is intuitive that a no pooling point estimate with a narrower uncertainty interval would not need to borrow as much information from the other sites since it is more sure of the estimate. In the multilevel model, this would lead to less shrinkage towards the complete pooling estimate.

Consider a simple normal-normal hierarchical model with $y_i \sim \texttt{Normal}(\alpha_j, \sigma_y^2)$ with $\alpha_j \sim \texttt{Normal}(\mu_\alpha, \sigma_\alpha^2)$ and known σ_y^2, σ_δ^2 and μ_δ. For each group j, the posterior estimate for α_j is [15]:

$$\begin{aligned} \textit{posterior} \quad \mathbb{E}(\alpha_j) &= (1 - \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\bar{y}}^2})\mu_\alpha + \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\bar{y}}^2}\bar{y}_j \\ &= \omega\mu_\alpha + (1-\omega)\bar{y}_j \end{aligned} \tag{2.7}$$

where $\bar{y}_j$ is the sample mean of data in site j and

$$\omega = 1 - \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\bar{y}}^2} \tag{2.8}$$

is called the pooling factor. It represents the extent to which partial pooling has taken place and where the estimate is on the scale between no pooling and complete pooling. If $\omega = 0$, we would have $\hat{\alpha}_j = \bar{y}_j$ which represents the no pooling estimate where only the data in that specific site is used. However, on the opposite end, $\omega = 1$ results in $\hat{\alpha_j} = \mu_\alpha$, corresponding to the common mean value that the estimates are pooled towards.

The uncertainty about the estimate, $\hat{\alpha}_j$, is given by the standard deviation:

$$\textit{posterior} \quad \sigma(\alpha_j) = \sigma_{\bar{y}}\sqrt{1-\omega} \tag{2.9}$$

If we increase the shrinking to a common value with a larger ω, then the standard deviation is shrunk towards 0 and the uncertainty about the estimates decrease. With no pooling, however, the standard deviation is appropriately equal to $\sigma_{\bar{y}}$.

The difference between any two α_j can also be looked at to determine how close together the various estimates are [14]:
The posterior estimate for $\alpha_j - \alpha_k$, $\mathbb{E}(\alpha_j - \alpha_k)$, is given by $\mathbb{E}(\alpha_j) - \mathbb{E}(\alpha_k)$:

$$\textit{posterior} \quad \mathbb{E}(\alpha_j - \alpha_k) = (1-\omega)(\bar{y}_j - \bar{y}_k) \tag{2.10}$$

$$= \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_{\bar{y}}^2}(\bar{y}_j - \bar{y}_k) \tag{2.11}$$

The posterior standard deviation, $\sigma(\alpha_j - \alpha_k)$, is given by $\sqrt{\sigma^2(\alpha_j) + \sigma^2(\alpha_k)}$:

$$\textit{posterior} \quad \sigma(\alpha_j - \alpha_k) = \sqrt{2(1-\omega)\sigma_{\bar{y}}^2} \tag{2.12}$$

$$= \frac{\sqrt{2}\sigma_{\bar{y}}\sigma_\alpha}{\sqrt{\sigma_{\bar{y}}^2 + \sigma_\alpha^2}} \tag{2.13}$$

These equations can then be used to find the z-score. This is equal to the expectation divided by the standard deviation as follows:

$$
\begin{aligned}
\textit{posterior} \quad \text{z-score}(\alpha_j - \alpha_k) &= \frac{\mathbb{E}(\alpha_j - \alpha_k)}{\sigma(\alpha_j - \alpha_k)} \\
&= \frac{(1-\omega)(\bar{y}_j - \bar{y}_k)}{\sqrt{2(1-\omega)\sigma_{\bar{y}}^2}} && \text{(Using 2.10 and 2.12)} \\
&= \frac{\bar{y}_j - \bar{y}_k}{\sqrt{2}\sigma_{\bar{y}}} \cdot \sqrt{1-\omega} && \text{(By rearranging)} \\
&= \underbrace{\frac{\bar{y}_j - \bar{y}_k}{\sqrt{2}\sigma_{\bar{y}}}}_{\text{unpooled z-score}} \cdot \underbrace{\frac{1}{\sqrt{1 + \sigma_{\bar{y}}^2/\sigma_\alpha^2}}}_{\text{partial pooling correction}} && \text{(Using 2.8)} \qquad (2.14)
\end{aligned}
$$

The posterior z-score for this model is made up of two parts. The first part is the z-score if there was in fact no pooling and the second part is the correction in order to share information across groups. This correction is always less than 1 such that the difference in point estimates can only decrease and cannot lead to the estimates spreading out. It can be seen that the ratio of $\sigma_{\bar{y}}^2/\sigma_\alpha^2$ is important in influencing how much pooling there is. If σ_α^2 was to decrease towards 0, the partial pooling correction would also approach 0. This would result in a smaller difference between estimates and thus more shrinkage.

This chapter has provided an insight into the various classical and Bayesian approaches to the issue of multiple comparisons. Throughout the course of the project, I implement and compare these approaches on a range of data.

Chapter 3

Project Execution

In this chapter, I use the models set out by Gelman [14] and as described in the Background 2, to fit data from the IHDP. I assess their performance by setting up a simulation framework in order to recreate the data and estimate the FDR, FWER and Type II error rate for each model.

3.1 Fitting Models to the Data

3.1.1 The IHDP Data

For this project, I used data from the IHDP [18] to investigate the issue of multiple comparisons. This was a large dataset including characteristics of the infants such as their birth weight and their parents' economic status along with many measures taken to assess the infants throughout the three years. More specifically, I used data from the Primary Analysis Group in which only one infant from each pair of twins was included. To simplify this for my project, I extracted three sets of values for each infant i:

- The infant's site, $j = 1, ..., 8$.
- The infant's Program Status indicating whether it received the treatment, $P_i = 0, 1$.
- The infant's Stanford-Binet IQ score [26] assessed at 36 months, y_i.

After omitting missing values, I obtained a set of 908 infants that I stored in the dataset, `data`. For each site, j, I tested whether the treatment was effective:

$$H_0 : \text{Treatment has no effect on IQ scores, } \delta_j = 0 \tag{3.1}$$
$$H_1 : \text{Treatment does have an effect on IQ scores, } \delta_j \neq 0 \tag{3.2}$$

3.1.2 Frequentist Models

Using the linear regression model, I fit the data with a significance level of $\alpha = 0.05$ to represent the standard version of this model and also with $\alpha = 0.05/8 = 0.00625$ to apply the Bonferroni correction. From the model fits, I extracted the estimates for δ_j and the corresponding confidence intervals for each site j.

As shown in 3.2, the point estimates are the same in both cases which is as expected since the Bonferroni correction only changes the significance level and not the estimate itself. This significance level adjustment results in wider confidence intervals that are more likely to include 0. This is shown by the blue estimates with CI including 0 and therefore not rejecting the null hypothesis. With the Bonferroni correction, three of the sites accept that there is no effect compared to only one site rejecting without the adjustment. This shows that the risk of a false negative is much higher in the Bonferroni case.

I also applied the Benjamini-Hochberg procedure to the classical linear regression model with an allowance for a false discovery rate of 0.05. This involved sorting the p-values and finding the largest one that was less than its corresponding critical value, $(\text{rank} \cdot 0.05)/8$. As shown in Table 3.1 and Figure 3.1, this is the seventh largest p-value, and all p-values less than it are significant. This provides sufficient evidence to reject H_0 for all sites, except site 3.

Table 3.1: **Table for the Benjamini-Hochberg procedure with p-value's and critical values for each site's treatment effect** The p-values have been sorted by their rank and the darkest colour is the smallest p-value that is less than its critical value and corresponds to site 7. The coloured cells contain all values less than or equal to this p-value, meaning that they are significant.

Site	Rank	Sorted p-value	Critical value
4	1	0.0000114	0.00625
1	2	0.0000319	0.0125
5	3	0.0000675	0.01875
2	4	0.0012800	0.025
8	5	0.0032500	0.03125
6	6	0.0159000	0.0375
7	7	0.0344000	0.04375
3	8	0.9160000	0.05

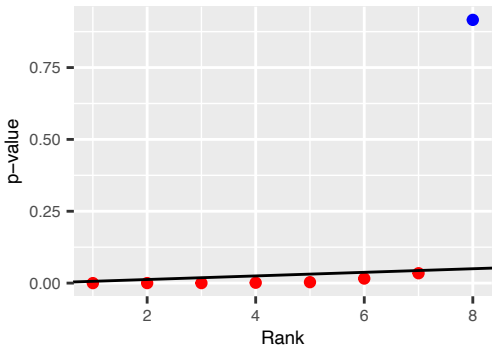

Figure 3.1: **Plot for the Benjamini-Hochberg procedure with p-value's for each site's treatment effect against their rank.** The black straight line represents the critical values. The largest p-value under its critical value is the one with rank 7. Red and blue points indicate significant and non-significant p-values.

3.1.3 Bayesian Model

Fitting the Model in RStan

I used the Bayesian multilevel model in 2.5 and 2.6 to fit the IHDP IQ scores from `data`. In order to fit to this model, I used Stan, and more specifically the R interface RStan. This is a probabilistic programming language which can be used for statistical inference for complex Bayesian models. It uses the Hamiltonian Monte Carlo (HMC) algorithm to generate proposal samples in MCMC. HMC is an efficient algorithm which uses principles from physics with a particle to represent the vector of parameters in the high-dimensional parameter space. The particle moves around the log posterior distribution such that it follows physical paths determined by the curvature of the distribution's gradient [8]. This results in intelligent proposals being generated.

I first specified the model in a Stan program which consisted of filling in the data, parameters and model blocks. The data block included the number of infants and the number of sites along with the IQ score, site and treatment group corresponding to each infant. All the parameters were in the parameter block and the model block included the relationship between the observed data and the unknown parameters. Since this is in the Bayesian framework, the model parameters are treated as random variables and prior distributions must be specified for them in the model block of the Stan program. I used mainly weakly informative priors as defined by Gelman in [12] to be a distribution that is "proper" but "set up so that the information it does provide is intentionally weaker than whatever actual prior knowledge is available". This was to avoid strong assumptions while still incorporating some initial knowledge.

$$\begin{aligned} \mu_\delta &\sim \texttt{Normal}(0, 15^2) & \gamma_j &\sim \texttt{Normal}(100, 15^2) \\ \sigma_\delta &\sim \texttt{Exponential}(1) & \sigma_y &\sim \texttt{Exponential}(1) \end{aligned} \tag{3.3}$$

The Stanford-Binet IQ scale is standardised such that the population has an average IQ of 100 and a standard deviation of 15 [26]. The prior on γ_j reflects this belief and is therefore quite informative. For δ_j, the prior has a mean of 0 and a standard deviation of 15 which allows for a wide range of possible values while still reflecting some belief about the parameters. Similarly, the priors for σ_y and σ_δ are weakly informative. The Exponential distribution is only defined for non-negative values and thus ensures that the standard deviations are always positive.

Using RStan, I compiled the Stan program in R to fit the model. This involved specifying arguments in the `stan` function from the package 'rstan', to run HMC and obtain posterior samples for the model parameters.

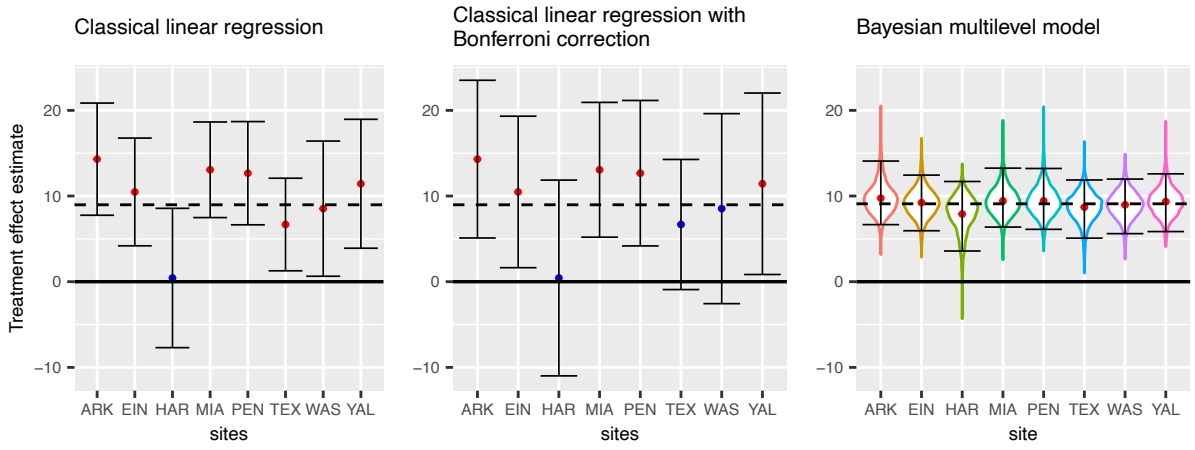

Figure 3.2: **Comparison of the treatment effect estimates and their 95% uncertainties intervals for the classical regression model with and without the Bonferroni correction and the Bayesian multilevel model, inspired by** [14]. The marginal distribution for each δ_j is shown for the Bayesian model. The colour of the point estimates reflect whether or not the corresponding uncertainty interval contains 0: red ones do not while blue ones do. The Bonferroni correction produces wider uncertainty intervals and has more blue points. The Bayesian multilevel model has closer together point estimates with narrower intervals, and thus more red points.

```
fit <- stan(
    file = "ihdp-iq.stan",   # Stan program
    data = data,             # list of data with IQ, site and
# treatment group for each infant
    chains = 4,              # number of Markov chains
    warmup = 1000,           # number of warmup iterations per chain
    iter = 2000,             # total number of iterations per chain
    cores = 4,               # number of cores
    refresh = 0              # no progress shown
)
```

I ran the algorithm on four Markov chains in order to check the reliability of the posterior estimates and that convergence to the target distribution had been achieved. This was done over four cores so that the chains could run in parallel and take less time. The samples taken in the warmup stage adapt the sampling such that the samples after are from the target distribution.

Diagnostic Checks

After fitting a model in RStan, diagnostics had to be checked to assess the reliability of the MCMC algorithm [25]. This included checking the trace plots, $\hat{R}$ and the number of effective samples.

In order to check the convergence of the chain, I started by looking at plots of the samples for each iteration, called trace plots. This included checking that all chains were stationary as shown by the chains staying within the same portion of the posterior distribution and not drifting too far away. I also checked that there was good mixing such that the chain explored the parameter space efficiently and was not clustered in a particular region. I also looked at the measure of convergence, $\hat{R}$, which is very close to 1 for chains that have converged well. In MCMC, it is also desired to have a large number of independent samples, known as the effective sample size, n_{eff}, to show the efficiency of the chains. Figure 3.3 shows the initial diagnostic checks and provides evidence that the model is fitted correctly. Throughout this project, I used these convergence diagnostics for each new fit to a Bayesian model.

However, when initially fitting the model, approximately 10% of the transitions were divergent. This is where a proposal state encounters numerical problems meaning that the region of the posterior distribution around that state is hard to explore [7]. This was resolved by reparameterising δ_j such that:

$$\delta_j = \mu_\delta + \sigma_\delta \xi_j \qquad \text{where} \quad \xi_j \sim \texttt{Normal}(0, 1) \tag{3.4}$$

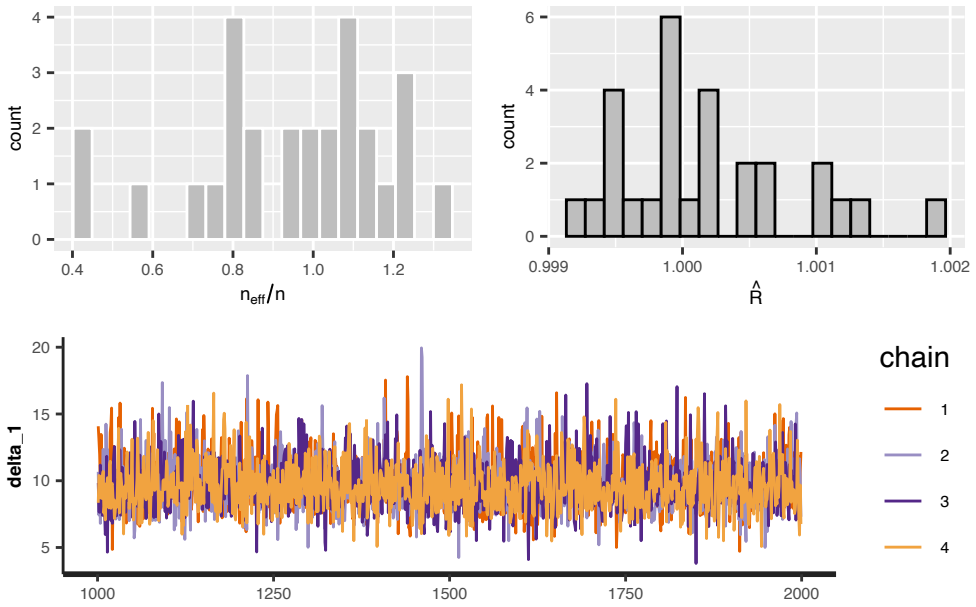

Figure 3.3: **Diagnostic checks after fitting `data` to the Bayesian multilevel model.** The top-plots are histograms of characteristics of all the parameters. The top-left plot shows the proportion of samples that are effective which is above 0.4 for all parameters. The top-right plot depicts $\hat{R}$ which is always very close to one. The bottom plot is a trace plot for δ_1, and shows good mixing around the same portion of the posterior distribution.

Having performed the diagnostic checks, posterior samples for the model could be obtained.

Interpreting the Estimates of δ_j

In order to extract estimates for the parameters from the Bayesian model, I used the posterior mean. In this way, I obtained point estimates for δ_j by taking the mean of the 1000 δ_j samples that were produced by the model. I also calculated the uncertainty intervals using the Highest Posterior Density Interval (HDI) as a Bayesian counterpart to the frequentist CI. The HDI is the narrowest interval containing a specified probability mass [25]. Just as with CIs, 95% HDIs were used with H_0 being rejected if the interval did not include 0.

From the classical linear regression to the Bayesian multilevel model, the point estimates have shrunk towards each other, as shown in Figure 3.2. More specifically, they shrink towards the dotted line. This line corresponds to the complete pooling estimate if the IQ scores had not been separated into their respective sites and a treatment effect, δ, for the whole set of infants had been estimated. The uncertainty intervals have also become narrower due to the more accurate estimates from the use of partial pooling [14]. Since there is more certainty around the estimates, fewer false positives and false negatives arise.

3.2 Framework for Simulation Studies

As described in the previous section, I began this project by fitting models to data from the IHDP. This included the Bayesian multilevel model and is illustrated on the left-hand side of 3.4. Using this Bayesian multilevel model, I extracted posterior samples of each parameter to provide estimates for their posterior distributions. These posterior samples are repeatedly used throughout my project and I will refer to them as `post` for convenience.

In my experiments, I simulated new data to assess model performance in various contexts. I used a generative model that was based on the Bayesian multilevel model with the posterior estimates from `post` for the parameters. By using the list of sites and treatment group corresponding to each infant in the IHDP data, the data produced was in the same form and of the same size, $N = 908$, as the original data. The model provided a flexible way to simulate data under various contexts as it could be adjusted to fit with the situation. The simulated data was then fitted to the various models described above and

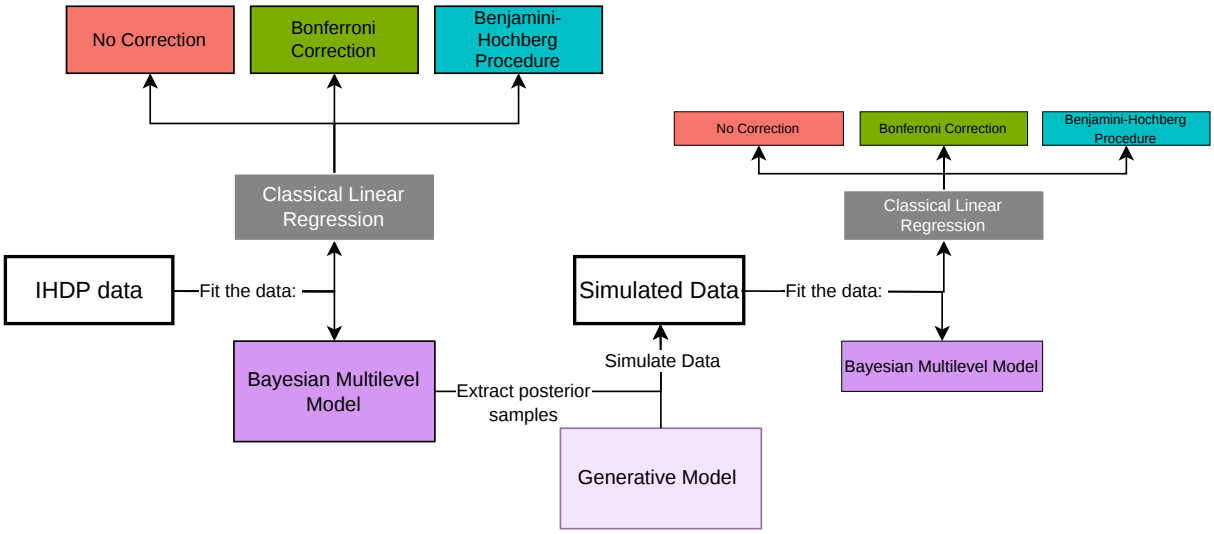

Figure 3.4: **Diagram representing the simulation framework.** The IHDP data is fitted to the models and the posterior samples from the Bayesian multilevel model are used along with the generative model to simulate more data. This data is then refit to the same models.

as before, I conducted tests to evaluate their performance. This is illustrated on the right-hand side of 3.4.

In the linear regression case, the hypothesis test described in 3.1 and 3.2 was conducted at a 95% significance level. The Bonferroni correction and B-H procedure were applied to determine which tests rejected H_0. As before, the Bayesian multilevel model used 95% HDI to determine whether H_0 should be rejected. I estimated various error rates to evaluate a model's performance on the simulated data,

- The family-wise error rate [4]. This is the proportion of sets of data simulated under H_0 that resulted in at least one rejection.
- The false discovery rate [5]. This is the proportion of the total rejections that were false. Therefore it is equal to:

$$\frac{\#\text{ of rejections from fitted data under } H_0}{\#\text{ of rejections from fitted data under } H_0 + \#\text{ of rejections from fitted data under } H_1} \tag{3.5}$$

- The Type II error rate [23]. This is the proportion of total tests for the sets of data simulated under H_1 which incorrectly accepted H_0.

By comparing these errors, I could evaluate model performance in a variety of situations.

3.3 Model Performance on the IHDP Data

3.3.1 Recreating the Data

As an introduction into my simulation studies, I looked at each model's performance on the IHDP data itself. This was done by simulating the IQ scores in an attempt to recreate the data and finding the error rates for the fit of each model. I used 1000 posterior samples from `post` to simulate 1000 sets of the data, each of sample size $N = 908$. This was done with the generative model equivalent to the fitted Bayesian multilevel model:

$$y_i \sim \texttt{Normal}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma_y^2) \qquad \text{with} \qquad \delta_j \sim \texttt{Normal}(\mu_\delta, \sigma_\delta) \tag{3.6}$$

In this case, each set of data was generated with different posterior samples of δ_j, γ_j and σ_y. For each infant i in the sample, an IQ score, y_i, was then simulated with i's treatment group and the parameters corresponding to i's site. This enabled the whole distribution of each parameter to be portrayed such that a better representation of the true data was provided across all the simulations. It provided samples

of IQ scores with the assumption that there is a treatment effect, since in all cases the posterior sample of δ_j was not equal to 0.

The simulations are shown below and stored in `recreate_scores`: a matrix of 1000 rows with the 908 IQ scores. I used the dataset, `data` from 3.1.1, containing the list of sites and treatment group for each infant.

```
recreate_scores <- matrix(nrow=1000, ncol=N)
for (sample in 1:nrow(recreate_scores)){
  gammas <- post$gamma[sample,]  # 8 posterior gammas for that sample
  deltas <- post$delta[sample,]  # 8 posterior deltas for that sample
  for (i in 1:N){
    site <- data$site[i] # infant i's site
    P_i <- data$P[i] # infant i's treatment group
    # simulate one IQ score using generative model:
    recreate_scores[sample, i] <- rnorm(1,mean=gammas[site]
                                + deltas[site]*P_i, sd=post$sigma_y[sample])

  }
}
```

By using site as an index, `[site]`, the correct parameter corresponding to i's site was accessed.

In order to find the number of false positives, data had to be simulated such that the null hypothesis was true. Therefore, I generated 1000 samples with the assumption that there was no treatment effect using this generative model:

$$y_i \sim \texttt{Normal}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma_y^2) \qquad \text{with} \qquad \delta_j \sim \texttt{Normal}(0, \sigma_\delta^2) \tag{3.7}$$

This was done in the same way as the previous simulation with the same posterior samples, but with the exception of δ_j. Instead, of using its posterior samples, I simulated 8 δ_j's with $\mu_\delta = 0$ for each set of data: `deltas <- rnorm(8, mean=0,sd=post$sigma_delta)`. Setting $\mu_\delta = 0$, rather than δ_j, provided treatment effects that had a mean of 0 but still some random site-to-site variation from σ_δ. This was different to data under H_1 in that I did not have to generate δ_j's with posterior μ_δ and σ_δ since the posterior δ_j's were already assumed to be taken from this distribution.

3.3.2 Comparing each Model's Performance on the Simulated Data

As described in the framework, each set of data, based on H_0 and H_1 was fitted to the models in the diagram 3.4 and the tests were conducted to find the number of rejections in each case. Using these, I found the various error rates.

The plots in 3.5 show that the classical linear regression model has a significantly higher FWER and FDR compared to the other three models. More specifically, the uncorrected linear model resulted in a FWER of 0.422, at least 5 times as large as the other three, and a FDR of 0.082, at least 3 times as large. These results provide evidence for the necessity of some form of multiple comparisons correction in order to ensure the results of the tests are valid.

The two corrections to the linear model perform quite similarly in terms of the FWER and FDR. Naturally, the Bonferroni correction leads to a slightly lower FWER while the lower FDR comes from using the B-H procedure. This is expected since the adjustments target the different error rates. However, where these corrections do less well is in the Type II error rate that arises. The Bonferroni correction especially leads to a very large Type II error rate of 0.570 compared to the linear model of only 0.292. This supports the idea that the Bonferroni correction is a highly conservative procedure and can be too cautious to reject the null hypothesis when it is in fact true. The B-H procedure resulted in a 0.351 probability of false negatives. This is much lower than with the Bonferroni correction and thus provides evidence for it being less conservative. However, it is still not as low as the Type II error rate for the uncorrected linear model.

While the linear model and its suggested corrections have advantages and limitations when looking at different criteria, the Bayesian multilevel model performs well throughout. Out of the four models, it has the smallest FWER and Type II error rate, with the Type II error rate being approximately 65 times smaller than that of a non-adjusted linear model. However, it does result in a slightly higher FDR than the linear model corrections. Since this increase is minimal and the Type II error is reduced so greatly, the Bayesian multilevel model can be thought of as to perform better on the whole.

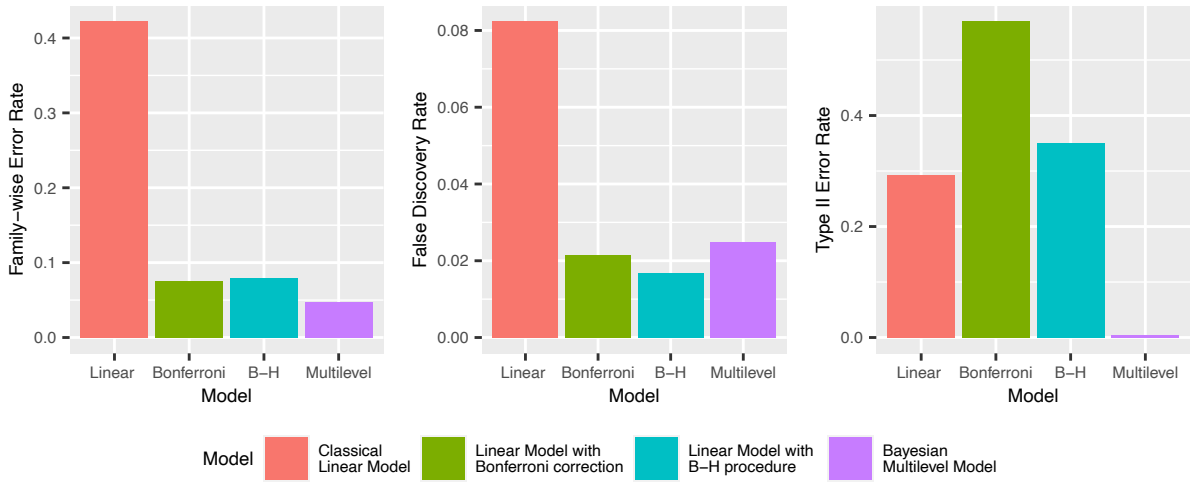

Figure 3.5: **A comparison of the four models based on error rates obtained after fitting simulated data based on data from the IHDP**. The uncorrected linear model displays a high FWER and FDR as compared to the other three. The Bonferroni correction and B-H procedure result in more Type II errors than the uncorrected linear model. The Bayesian multilevel model has an almost 0 Type II error rate.

In the following two chapters, I change various aspects of how the data is simulated. I see how well the models respond based on the error rates and whether the pattern between the models is maintained. This involves looking at data with alternative variance ratios, $\sigma_y^2/\sigma_\delta^2$ in 4, and adding in correlation between parameters in 5.

Chapter 4

Simulation Study 1: The Effect of group-level variance and its prior

In this section, I look at how the variation in IQ scores between sites affects the performance of each model. This is done by changing the group-level variance, σ_δ^2, of the true data. Following on from this, I look at how different prior variance ratios in the Bayesian multilevel model affect its fit.

4.1 Variation between Sites

4.1.1 Simulating Data with Different $\sigma_y^2/\sigma_\delta^2$

I started by simulating data with various σ_δ^2 and a fixed σ_y^2. With the Bayesian multilevel model as a generative model, we can look at the ratio between the variances, $\sigma_y^2/\sigma_\delta^2$ as a degree of heterogeneity in the data [19]. This implies that a large ratio indicates that there is a large amount of within group variance relative to the variance between groups. By focusing on the ratio rather than just σ_δ^2, I was able to examine the joint effect of both the group-level variance and the residual variance on the error rates.

Data from the IHDP produced posterior samples of σ_y^2 with a mean of 316 and σ_δ^2 with a mean of 2.38 from `post`. Therefore, the estimated ratio of variances was approximately 132. In order to recreate this and try out new values, I used a fixed $\sigma_y^2 = 300$ as a rough approximation as well as values of σ_δ^2 such that I had ratios of: $10, 25, 50, 100, 150, 200, 250, 300$.

I used the Bayesian multilevel model to generate 100 sets of data under H_0 3.1 and 100 sets under H_1 3.2. I set $\sigma_y^2 = 300$ and repeated this for each value of σ_δ^2.

$$y_i \sim \texttt{Normal}(\gamma_j + \delta_j P_i, 300) \qquad \text{with} \qquad \delta_j \sim \texttt{Normal}(\mu_\delta, \sigma_\delta^2) \tag{4.1}$$

While in the last simulation I used different posterior parameters for each set of data, in this one I chose a fixed value for each parameter to use across the simulations. γ_j was taken to be the mean of the corresponding posterior samples for each j: `gamma_means <- apply(post$gamma, 2, mean)`. This was still a good representation of the data but added simplicity in order to wholly focus on changing σ_δ^2. For each simulation, I generated 8 values of δ_j under a normal distribution. This corresponded to the 8 sites such that each site had its own treatment effect.

```
deltas <- rnorm(8, mean=mu_delta, sd=sqrt(sigma2_delta))
```

Under H_0, μ_δ was set to 0 and under H_1, μ_δ was set to 10. The mean of all δ_j from `post` was equal to 9.06 and so 10 was chosen as a rough approximation.
For each infant i, y_i was generated with `data` as defined before:

```
for (i in 1:N){
  site <- data$site[i]
  P_i <- data$P[i]
  ratio_sigma_delta_ys[sample, i] <- rnorm(1,mean=gamma_means[site] +
                        deltas[site]*P_i, sd=sqrt(sigma2_y))
}
```

This produced similar sets of data for each σ_δ^2 to fit to the models.

4.1.2 The Effect of the Variance Ratio on each Model

Each sample was fit to the models and the error rates were once again computed for the various ratios in the same way.

As the ratio increases, we can see that in 4.1 for each model, all the error rates tend to decrease. This decrease is steep and obvious for low ratios but as the ratio increases, the errors tend to level out. Data with a higher between-group variance, σ^2_δ, and so a lower variance ratio, produces more variability between sites and so greater differences in the estimated group-level effects, δ_j. This leads to a greater probability of observing a significant effect due to chance and results in a larger FWER and FDR. However, as the ratio increases, the simulated data is more similar across the sites and result in estimates for δ_j that are closer together. Therefore, when H_0 is true, the chance of a false positive decreases, since the uncertainty interval is more likely to include 0. This is shown in 4.2 in which the fitted data has been simulated as described in 4.1.1 with $\mu_\delta = 0$ but with $\sigma^2_\delta = 5$ for the top line and $\sigma^2_\delta = 50$ at the bottom. With more group-level variation, the point estimates are further apart for all three of the models. This leads to fewer uncertainty intervals including 0 as shown by an increased number of red dots and thus more false positives. The levelling out of errors in 4.1 suggests that once σ^2_δ becomes more than about 100 times as small as σ^2_y, the noise from σ^2_y overpowers any negligible effect of the small between-group variation.

Similarly, under H_1, a higher σ^2_δ would lead to estimates that are more spread out and uncertainty intervals that are more likely to incorrectly include 0. This corresponds to a greater chance of committing a Type II error. The change in σ^2_δ does not affect the Type II error rate as much as it does for FWER and FDR. This suggests that the error rate is influenced more by a large σ^2_y creating wide uncertainty intervals that often incorrectly include 0, regardless of how much variation there is. This is especially the case for the Bonferroni correction in which the wider CIs result in almost no decrease in error as the ratio increases.

The various models show similar patterns to the previous chapter, 3, when looking at how they compare for each error type. For the FWER and FDR, the classical linear model performs very badly while the other three perform similarly better. There is the same pattern as before in Type II error rate as well, with the Bayesian multilevel model performing considerably better.

There is some evidence that the Bayesian multilevel model can handle data with a greater complexity from a higher σ^2_δ as shown by the smaller error rates for lower ratios in 4.1. The hierarchical structure of the data becomes more important since there is more variability between the sites and the multilevel model can take this into account. However, when the between-group variance is small relative to the individual variance, the hierarchical structure of the model becomes less important and the Bayesian model and classical corrections perform similarly.

In the case of the IHDP data, I obtained a mean posterior ratio of 132 and even though a Bayesian multilevel model is preferred, the hierarchical structure is of less importance. These results show that if there is more variation between sites then using a Bayesian multilevel model to fit the data is even more important in order to accurately represent the data.

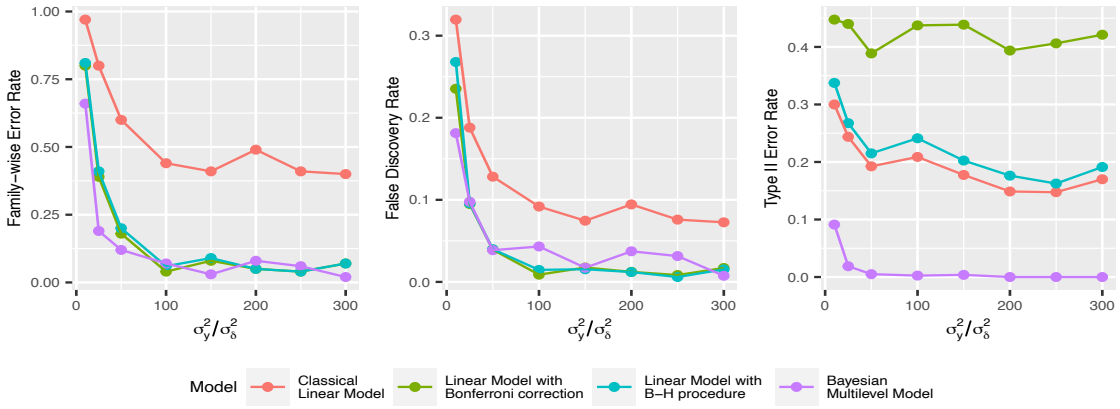

Figure 4.1: **A comparison of model performance based on the error rates and varying simulated data variance ratios.** Increasing the ratio decreases the error rates. Between models, the uncorrected linear model has high FWER and FDR. For large ratios, the other three models have similar FWER and FDR, but for smaller ratios, the Bayesian multilevel model has a lower error rate. The Bonferroni correction results in many more Type II errors than the uncorrected linear model. The Bayesian multilevel model has significantly fewer Type II errors.

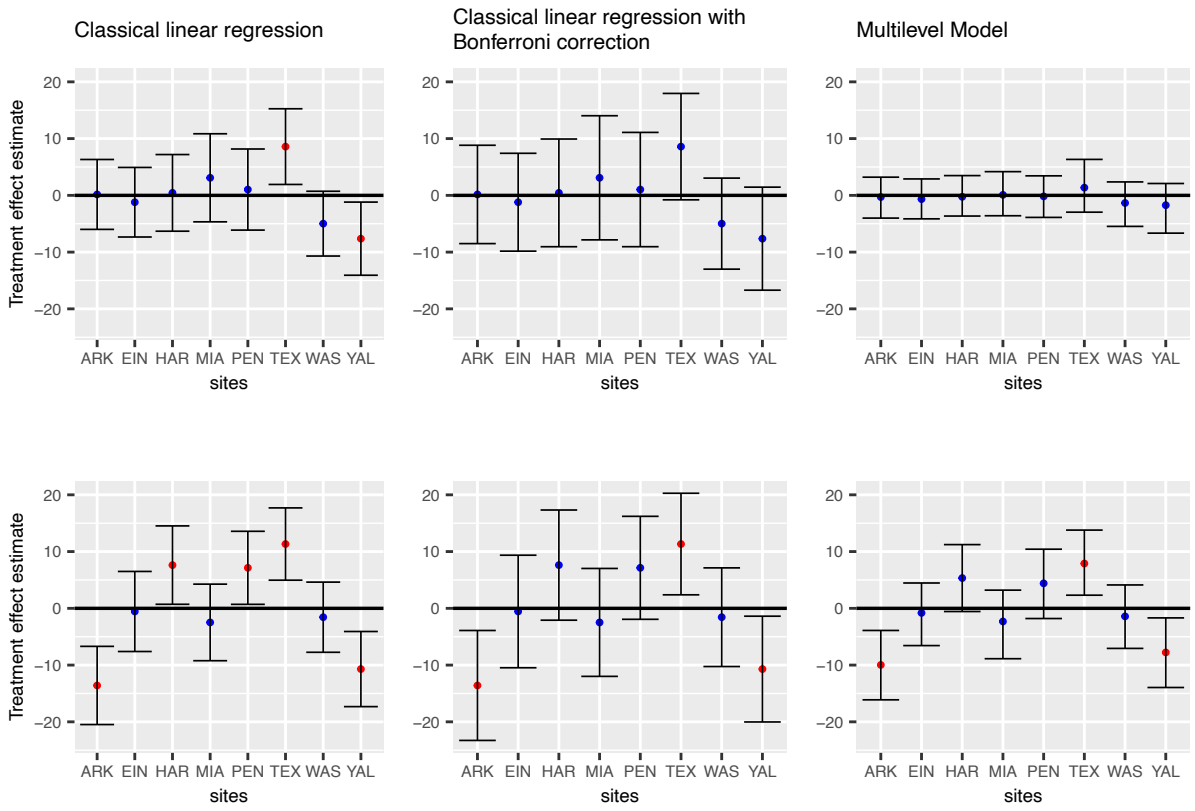

Figure 4.2: **A visualisation of treatment effect estimates and their 95% uncertainty intervals obtained from fitting data that had been simulated with different between-group variances.** The models fitted include the linear model with and without the Bonferroni correction and the Bayesian multilevel model. The data is simulated with $\mu_\delta = 0$ and with $\sigma_\delta^2 = 5$ for the top line, $\sigma_\delta^2 = 50$ for the bottom line. The red points represent significant estimates, the blue points represent insignificant estimates. There is more variation between sites across all the models in the bottom line. This results in more more red, significant estimates.

4.2 Changing the Prior on σ_δ^2

When fitting to the Bayesian multilevel model thus far, I have been using the model that was originally fit to the IHDP data. The group-level and individual level standard deviations were of the form: $\sigma \sim \texttt{Exponential}(\lambda)$ which gives an expected prior standard deviation of $1/\lambda$. I have been using $\lambda = 1$ for both σ_δ and σ_y and so $prior \mathbb{E}(\sigma_y)^2/\mathbb{E}(\sigma_\delta)^2 = 1$. However, the performance of this model is influenced by this prior variance ratio since it determines the amount of shrinkage of the estimates, as explained in 2.2.1. Therefore, I decided to try models with a wider and narrower prior on σ_δ to see how the error rates are affected:

- For a wide prior, I used $\lambda = 0.1$ which corresponds to $prior \quad \mathbb{E}(\sigma_y)^2/\mathbb{E}(\sigma_\delta)^2 = 0.01$.
- For a narrow prior, I used $\lambda = 10$ which corresponds to $prior \quad \mathbb{E}(\sigma_y)^2/\mathbb{E}(\sigma_\delta)^2 = 100$.

4.2.1 Larger Ratios

When there is a large ratio between σ_y^2 and σ_δ in the data, the narrow prior appears to have poor effects on the FDR especially, but also the FWER. As shown in 4.3, the model with $\lambda = 10$ results in a larger error than the other two priors. A narrow exponential prior is more concentrated on smaller values, resulting in a much smaller prior variance. This leads to more pooling of δ_j estimates towards a common value [14] as shown in 4.4. The results suggest that the pooling is very strong and that, occasionally, it is in fact pooling too much towards a value that is not in accordance with the value of μ_δ used for

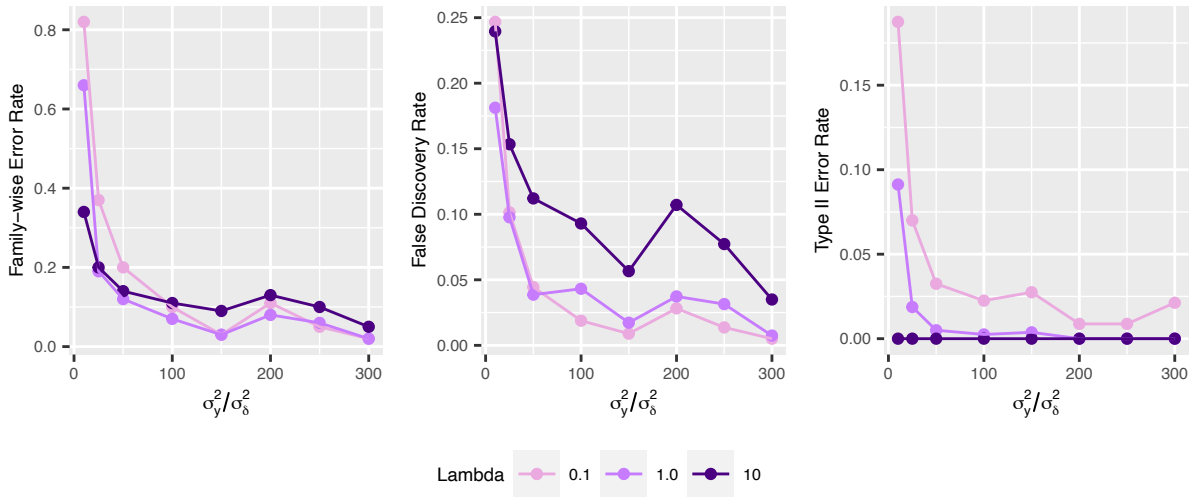

Figure 4.3: **Error rates from Bayesian multilevel models with different priors on $\sigma_\delta \sim \texttt{Exp}(\lambda$ with $\lambda = 0.1, 1.0, 10$. This is plotted against the variance ratio of the simulated data**. With increasing λ, FDR is greater and Type II error is smaller. When $\sigma_y^2/\sigma_\delta^2$ is large, a high λ results in a higher FWER, while with smaller values of $\sigma_y^2/\sigma_\delta^2$, the largest FWER comes from the smallest λ.

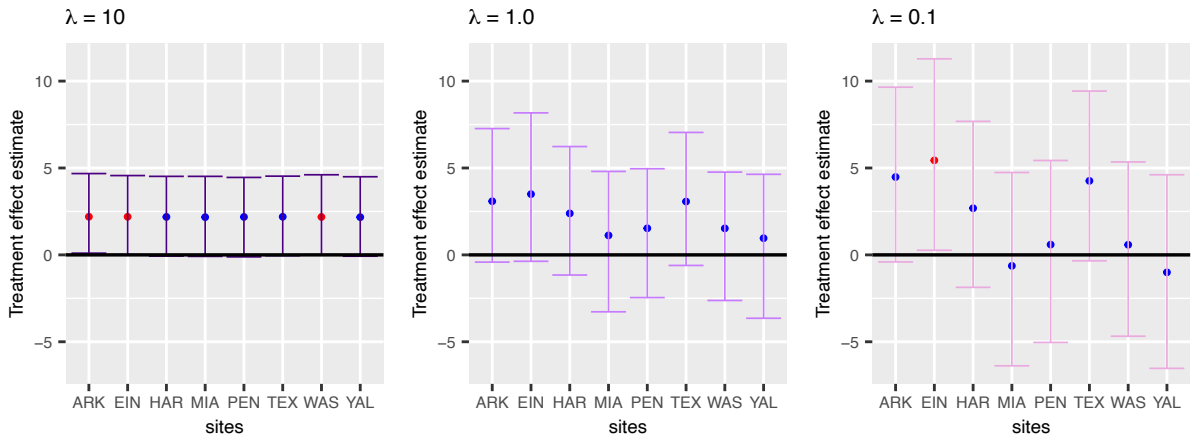

Figure 4.4: **Point estimates and their 95% HDIs from fitting data to a Bayesian multilevel model with differing λ's on the prior for $\sigma_\delta \sim \texttt{Exp}(\lambda)$.** The fitted simulated data is based on a ratio $\sigma_y^2/\sigma_\delta^2 = 25$ and $\mu_\delta = 0$. Smaller values of λ result in more spread out point estimates of δ_j and wider HDIs.

simulation. Since the simulations of the data are not based on δ_j but μ_δ being the true value, a random effect is added. Therefore, since we only have 8 sites, the simulated mean of the treatment effects may not be exactly μ_δ. With such a small prior variance, the pooling may be so strong towards the simulated mean that the 95% HDIs does not include the true μ_δ. When data is generated based on $\mu_\delta = 0$, we have an increased number of rejections due to an extreme amount of pooling towards a different value. This can be seen in the left-hand graph of 4.4.

When data is generated based on $\mu_\delta = 10$, it is likely that the pooled value will still be far from 0, and so the simulated mean of the treatment effects will not hugely affect the number of true rejections. In this case, more rejections arise when there is more shrinkage to produce uncertainty intervals that are less likely to incorrectly include 0. Therefore, the model benefits from a narrow prior with a small prior variance. We can see this from the negligible Type II error rate in 4.3 in which there are very few false negatives for $\lambda = 10$.

Therefore, with a narrower prior we have more false positives and more true positives. Since the FDR has increased, the results suggest that the increase of false positives must greatly outweigh the increase of true positives.

The model with a wider prior of $\lambda = 0.1$ generally produces a lower FDR than the original model but does not seem to considerably affect the FDR to the extent that the narrow prior did. Similarly, the difference between the FWER obtained from models with $\lambda = 0.1$ and 1 does not appear to be significant. Therefore, choosing between the two wider priors does not seem as crucial as ruling out the narrow one. However, the model with $\lambda = 0.1$ does lead to a slightly higher Type II error rate while the other two are similarly low. This suggests that $\lambda = 1$ is the best compromise when $\sigma_y^2/\sigma_\delta^2$ is large.

4.2.2 Smaller Ratios

When the ratio of $\sigma_y^2/\sigma_\delta^2$ is smaller, there is more variation between groups relative to σ_y^2. In this case, the narrow prior with $\lambda = 10$ appears to perform better and produces the lowest FWER 4.3. This suggests that the no-pooling estimates are so far apart that it becomes more important for the estimates to get shrunk towards each other so that the intervals are more likely to include 0. This ends up outweighing the limitation of shrinking towards a wrong value as before.

The beginning of this effect with small ratios can also be seen in the FDR plot. It can be explained by the huge increase of Type II errors when using a wide prior which is due to an insufficient amount of pooling. More false negatives imply fewer true positives, meaning that the FDR is larger and it starts to reflect the pattern in the FWER. This suggests that for a larger σ_δ^2, it is important to use a smaller prior on σ_δ to pool the estimates together and maintain low error rates.

4.3 Evaluating the Bayesian model based on Data with Different Variance Ratios

The IHDP data has an approximate ratio of 132. In this case and larger ratios, the model with $\sigma_\delta \sim$ `Exponential`(1), is generally the best choice since it maintains low values for each error rate. I have found that when there is little variance between sites, the effects of more pooling with a narrower prior actually have a negative effect on the FDR and FWER, despite increasing the certainty of the estimates. While we often want to reduce all error rates, it is sometimes important to focus on a specific one. For example, if the cost of a false negative was extremely great and false positives were of less importance then the narrower prior would be a good choice to almost rule out any possibility of a Type II error.

When the ratio is much smaller and there is more variation between sites, it is crucial to choose a narrower prior for σ_δ in order to produce more certain estimates that are closer together.

The smallest ratio of $\sigma_y^2/\sigma_\delta^2$ that I used to simulate the data was 10. Therefore, the effect discussed in 4.2.2 of using a large value of λ for data with large σ_δ^2 is not particularly obvious in my results. I also did not investigate the case in which there was more variance between sites than individual variance such that $\sigma_y^2/\sigma_\delta^2 < 1$. To examine this further, more research with smaller ratios would benefit this argument and emphasise my findings.

My σ_y^2 stayed fixed throughout, which allowed me to focus on how σ_δ^2 affects the models. However, since I was looking at the ratio between them, it would also be interesting to adjust σ_y^2 to examine how the models would perform with more or less noise.

Chapter 5

Simulation Study 2: The Effect of Correlation

Throughout this project, I have been assuming that there is no correlation in the data, meaning that none of the parameters at the individual level have been simulated to have any sort of association. However, in the case of the IHDP data, this may not be true. For example, of all infants without treatment, if the infants in site j have a higher average IQ score than in other sites, then they may not need the treatment as much and the treatment could be less effective. This creates a negative correlation between γ_j and δ_j in which if γ_j increases, δ_j decreases. In this section, I investigate the effect of the data having various correlations between these parameters. As well as the models used before, I also look at a Bayesian multilevel model that takes correlation into account. Following on from this, I look at how changing the prior parameters as well as the noise of the data can affect the fit of the model.

5.1 Correlation between γ's and δ's

5.1.1 Simulating with Correlation between γ's and δ's

To investigate how correlation between γ's and δ's may affect each model, I started by simulating data with a new generative model. This was done in a similar way to before but with γ_j and δ_j taken from a multivariate normal distribution in order to introduce an association between the two [25]:

$$y_i \sim \texttt{Normal}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma_y^2) \qquad \begin{bmatrix}\gamma_j \\ \delta_j\end{bmatrix} \sim \texttt{MVNormal}\left(\begin{bmatrix}\mu_\gamma \\ \mu_\delta\end{bmatrix}, \Sigma\right) \tag{5.1}$$

where Σ is the covariance matrix and can be decomposed such that it is written in terms of the parameters' standard deviations and the correlation matrix, R [25]:

$$\Sigma = \begin{pmatrix}\sigma_\gamma & 0 \\ 0 & \sigma_\delta\end{pmatrix} R \begin{pmatrix}\sigma_\gamma & 0 \\ 0 & \sigma_\delta\end{pmatrix} \qquad \text{with} \qquad R = \begin{pmatrix}1 & \rho \\ \rho & 1\end{pmatrix} \tag{5.2}$$

where ρ is equal to the correlation between the parameters.

I randomly generated values of y_i with this generative model and based on a range of $\rho = -0.9, -0.75, -0.6, -0.45, -0.3, -0.15, 0$, in order to introduce different negative correlations to the data. I used the posterior estimates from `post` to extract values for σ_δ and σ_y. Since μ_γ and σ_γ were not modelled in the original Bayesian multilevel model, I took their estimates to be the mean and standard deviation of all posterior samples of γ:

```
sigma_gamma <- sd(post$gamma)
mu_gamma <- mean(post$gamma)
sigma_delta <- mean(post$sigma_delta)
sigma_y <- mean(post$sigma_y)
```

The correlation matrix with the correlation, ρ, along with the standard deviation parameters were used to create the covariance matrix [25]:

```
# create a 2x2 correlation matrix with 1's at the diagonal and rho on
# the off-diagonals:
```

```
corr_matrix <- matrix(c(1, rho, rho, 1), nrow=2)

# creates a diagonal matrix with sigmas:
sigma_matrix <- diag(c(sigma_gamma, sigma_delta))

# use matrix multiplication to create the covariance matrix, Sigma:
Sigma <- sigma_matrix %*% corr_matrix %*% sigma_matrix
```

This allowed me to generate 8 values (one for each site) of γ and δ from the multivariate distribution: `gammas_deltas <- mvrnorm(8, mu=c(mu_gamma, mu_delta), Sigma=Sigma)` and then the y_i for each infant as done previously:

```
for (i in 1:N){
  P_i <- data$TG[i]
  site <- data$SITE[i]
  ys_cor[sample, i] <- rnorm(1,mean=gammas_deltas[site,1] +
                          gammas_deltas[site,2]*P_i, sd=sigma_y)
}
```

In alignment with Simulation Study 1 in Chapter 4, 100 sets of data were generated under H_0 with $\mu_\delta = 0$ and 100 under H_1 with $\mu_\delta = 10$

5.1.2 A New Model

As well as fitting each sample of data to the models mentioned previously, I also fit to an alternative Bayesian multilevel model. This model was of the same form as the generative model 5.1, so that the covariance matrix was also modelled to take correlation into account.

As in equation 3.4, I had to reparameterise to reduce divergent transitions. When δ_j and γ_j are modelled with a multivariate normal distribution, the reparameterisation uses the Cholesky factor L [25]. It comes from the Cholesky decomposition $R = LL^T$ where R is the correlation matrix. This results in δ_j and γ_j being modelled in this way:

$$\begin{bmatrix}\gamma_j \\ \delta_j\end{bmatrix} = \begin{bmatrix}\mu_\gamma \\ \mu_\delta\end{bmatrix} + \begin{pmatrix}\sigma_\gamma & 0 \\ 0 & \sigma_\delta\end{pmatrix} L \begin{bmatrix}\xi_j \\ \epsilon_j\end{bmatrix} \qquad \text{where} \quad \xi_j, \epsilon_j \sim \texttt{Normal}(0,1) \tag{5.3}$$

In order to fit this model, I had to provide priors on all the parameters. This was similar to the original fitted Bayesian multilevel model with the priors on μ_δ and σ_δ as shown in 3.3. The prior on γ in 3.3, $\gamma \sim \texttt{Normal}(100, 15^2)$, was deconstructed into μ_γ and σ_γ such that:

$$\mu_\gamma \sim \texttt{Normal}(100, 15^2) \qquad \sigma_\gamma \sim \texttt{Exponential}(1) \tag{5.4}$$

Additionally, a prior on L had to be given. This involved using an LKJ prior with a single parameter, η, that controls how much emphasis is put on different values of the correlation [22]. This can be seen in the density plots in Figure 5.1. I started by using $\eta = 1$ which defines a flat prior over all correlation values, like a uniform prior. As η increases above 1, the prior concentrates more on weaker correlations while if $\eta < 1$, then extreme correlations are more likely.

5.1.3 How Well Do the Models Perform?

Changing the correlation does not appear to have a considerable effect on the models' error rates, as shown in Figure 5.2. For each model, all the error rates are similar across all correlations with no obvious pattern and the behaviour between the original models is as in the previous chapters. However, the Bayesian multilevel model that takes correlation into account performs very similarly to the original one. This is surprising since it may be expected to model the estimates with more certainty since it can account for more characteristics in the data [11].

5.2 Investigating the Model accounting for Correlation

To check the reliability of the estimates obtained from this model, I looked at various diagnostic measures. For example, n_{eff}/n and $\hat{R}$ are shown in 5.3 for a fitted model and prove to have good results. n_{eff}/n

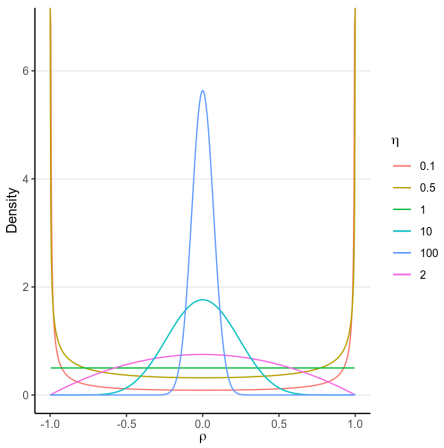

Figure 5.1: [21] **Density of the LKJ prior with various values of η.** As η increases, the distribution becomes more concentrated on weaker values of correlation, ρ

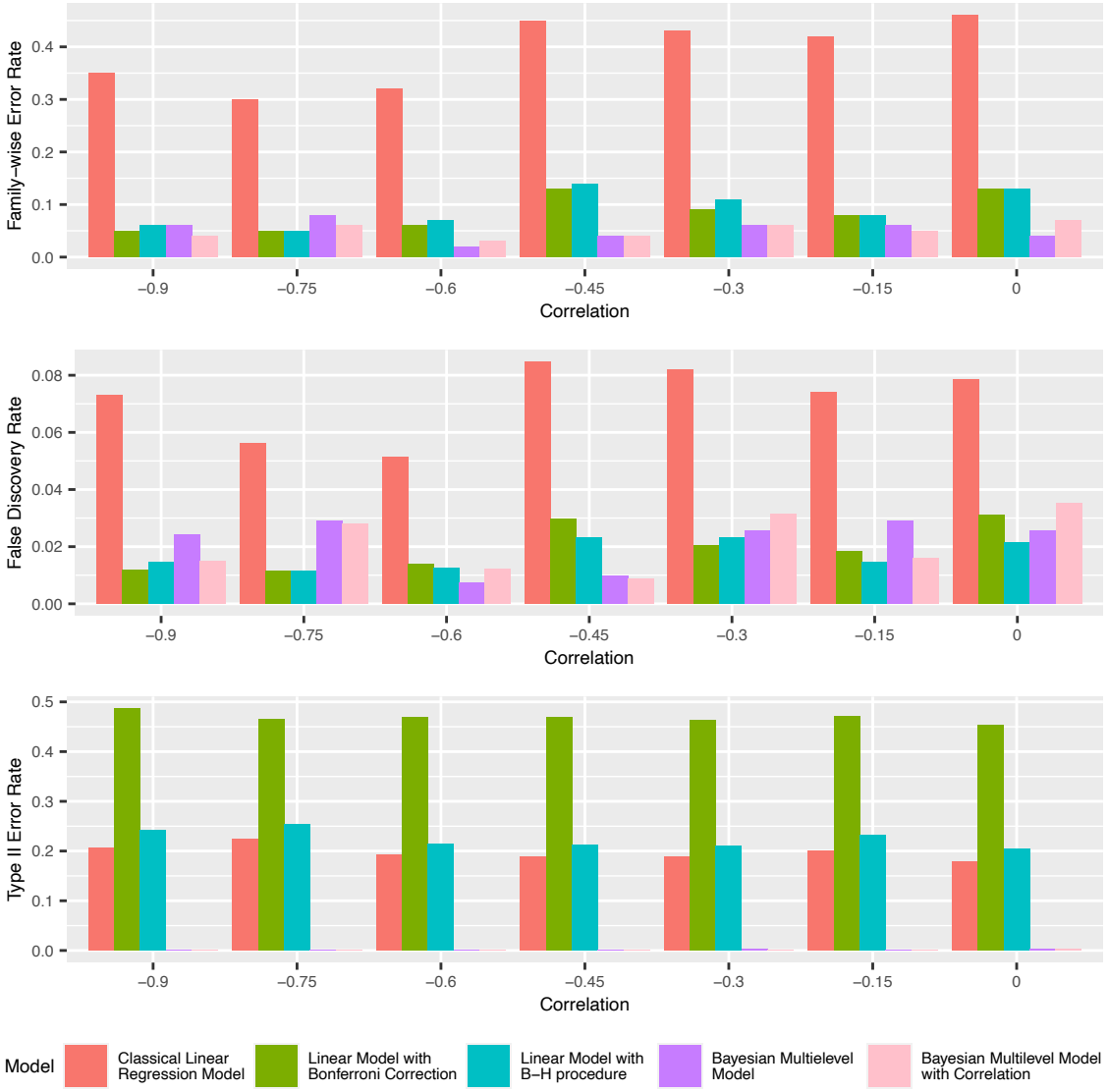

Figure 5.2: **Error rates for models fitted with data simulated with different values of correlation between γ and δ.** The models include the linear model with and without its two corrections and the Bayesian multilevel model with and without correlation. The data was simulated with a range of negative correlations but this does not appear to affect the errors. The Bayesian multilevel model accounting for correlation has similar error rates the one that does not.

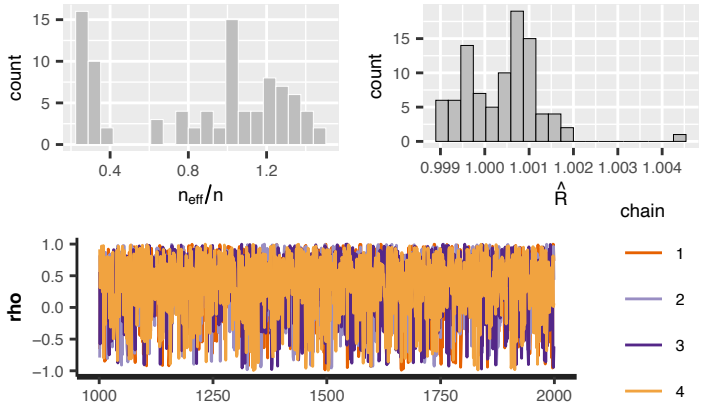

Figure 5.3: **Diagnostic checks after fitting one set of simulated data with correlation to the Bayesian multilevel model accounting for it**. The top-plots are histograms of characteristics of all the parameters. The top-left plot shows the proportion of samples that are effective which is above 0.2 for all parameters. The top-right plot depicts $\hat{R}$ which is in a range of 0.998 to 1.005 for all parameters. The bottom plot is a trace plot for δ_1, and shows good mixing around the same portion of the posterior distribution.

is always high and $\hat{R}$ is very close to 1 for all parameters. The trace plots also exhibited good behaviour as shown in the trace plot of ρ, for example, in 5.3, which further indicated good convergence.

Having carried out diagnostic checks, I proceeded to look at the posterior samples of ρ for each set of data to see how they compared to the corresponding correlation used to simulate that set. From looking at the first and third column in Table 5.1, it can be seen that there is very little correspondence between the target correlation and the mean posterior ρ, providing the posterior estimate. This is also shown in the left-hand plot of 5.4 in which posterior estimates of ρ are always around 0 and not in accordance with the true correlation shown by the black dots. Therefore, when the true correlation is strong, the posterior estimate is always very far off. More importantly, the shape of the distributions for each set of simulated data do not have a pronounced shape like a normal distribution. The oval shape across all values between -1 and 1 gives a mean close to 0 and indicates that the posterior samples are quite evenly distributed and not concentrated around one value. The wide distribution suggests a large amount of variability in the posterior samples and as a result, provides a less informative estimate of the correlation.

Original correlation	**Correlation after simulating γ's and δ's**	**Correlation posterior estimate, ρ**
-0.90	-0.916	-0.135
-0.75	-0.802	-0.254
-0.60	-0.571	0.204
-0.45	-0.435	-0.322
-0.30	-0.189	-0.067
-0.15	-0.224	-0.045
0.00	0.150	0.231

Table 5.1: **Table of values of correlation between γ's and δ's throughout the simulation and after fitting to the Bayesian multilevel model accounting for it**. The first column consists of the starting correlation that is used to simulate the data. The second column is the correlation after simulating just γ's and δ's and is very similar to the first column. The third column is the mean of the posterior samples of ρ and is quite different to the first column.

5.2.1 Changing the LKJ Prior

I investigated whether this was due to the LKJ prior on the cholesky correlation matrix, L, being poorly specified. If it was too weakly informative then a wide prior may have been the cause of a wide posterior. I fit each set of data with the various correlations to a model with $\eta = 0.5$ such that it performs better when data with strong correlation is fit. I compared this to a model with $\eta = 2$ to see if data with weaker

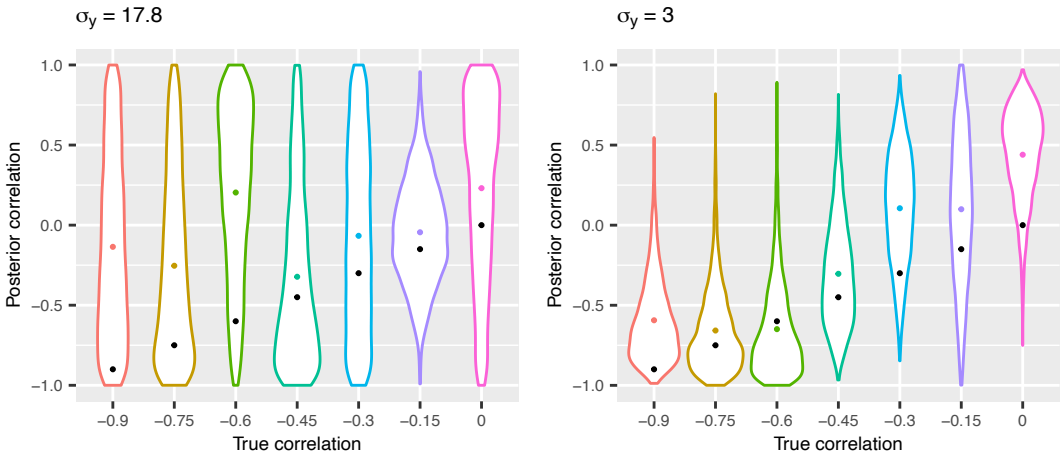

Figure 5.4: **Posterior distributions and posterior estimates of ρ after fitting data to the Bayesian multilevel model that accounts for correlation.** This is shown for data simulated with different values of ρ, as shown on the x-axis and the black dots. The coloured dots are the posterior estimates of ρ. The left-hand plots uses data simulated with $\sigma_y = 17.8$ and result in wider distributions and estimates that are further from the true value than the right-hand plot which uses $\sigma_y = 3$.

correlation produces smaller errors.

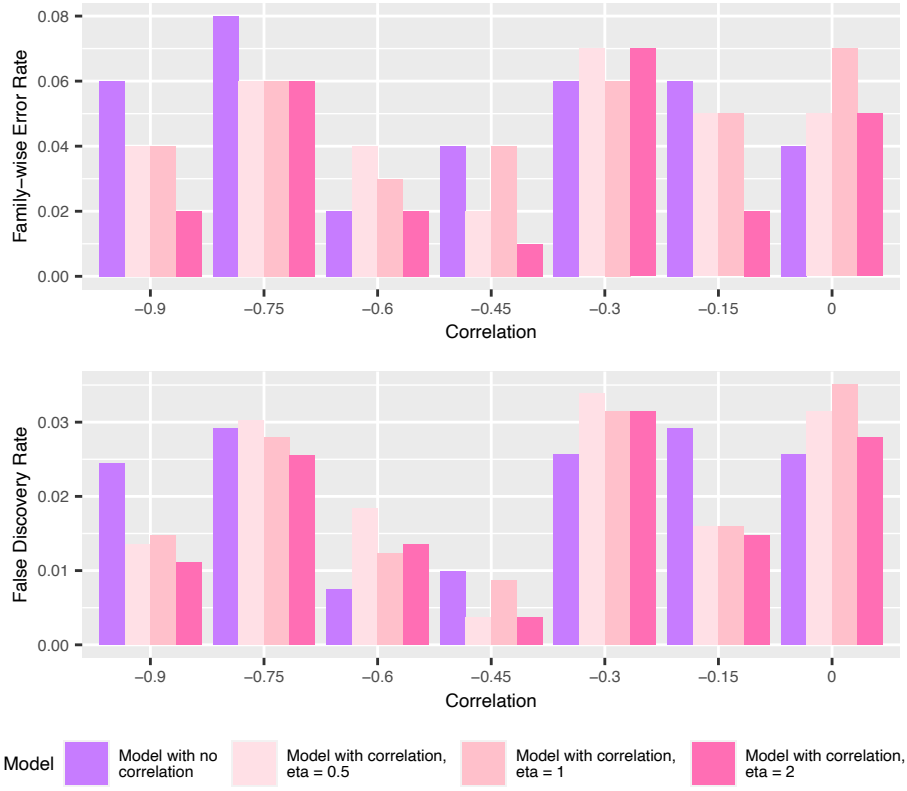

Figure 5.5: **FWER and FDR obtained after fitting data of various negative correlations between γ and δ to Bayesian multilevel models.** This included the model without correlation and three with correlation with different LKJ prior parameters: $\eta = 0.5, 1, 2$. There is no obvious pattern as the correlation weakens or across the models.

However, my results shown in 5.5 indicate that the value of η does not seem to affect the model significantly, since the error rates are similar across the models for each value of correlation. The LKJ distribution with parameter $\eta = 2$ is a curve concentrated on weaker correlations 5.1, but it is not

particularly pronounced and so may not have a large effect on the fit of the data. Even a strong correlation of $\rho = -0.9$ is not hugely affected and suggests that only even stronger correlations with $\rho > 0.95$ would be affected. Setting $\eta = 0.5$ creates the opposite effect where there is more concentration on the extreme correlations 5.1. In between these extreme correlations, however, the prior is quite flat. The strongest correlation value, -0.9, that I used to simulate the data is not extreme enough to be included in the huge upswing in the distribution. Therefore, the model with $\eta = 0.5$ performs similarly to the model with $\eta = 1$.

5.2.2 Noise in the Data

Looking further at the distributions of ρ in the left-hand plot of 5.4, I investigated whether the wideness was due to noise in the data, such that it was harder for the model to estimate the parameter accurately. From the middle column of 5.1, it can be seen that after simulating γ's and δ's in `mvrnorm(8, mu=c(mu_gamma, mu_delta), Sigma=Sigma)`, the correlation between the two sets of parameters is similar to the original correlation. This suggests that the data may become more complex when the y_i's are simulated.

The various standard deviation values used to simulate the data were obtained from `post` and were as follows: $\sigma_y = 17.8$, $\sigma_\gamma = 8.38$ and $\sigma_\delta = 1.17$. The high value of noise as compared to σ_γ and σ_δ may have disrupted the data and the true correlation between γ and δ may become harder to estimate accurately.

To investigate this, I simulated data in the exact same way but using a smaller value of $\sigma_y = 3$ such that less noise was introduced. This produced posterior estimates that were closer to the original ρ and distributions that were more concentrated on the corresponding mean, as shown in 5.4. The difference between the two plots is particularly apparent in stronger correlations. The posterior means are much closer to the true value when data was simulated with $\sigma_y = 3$, suggesting that the noise in the data did affect how well the multilevel model with correlation performed.

5.3 Evaluation of the Bayesian model based on Data with Correlation

My results have shown that the models in the previous chapters were not considerably affected by the fitted data containing correlation. I attempted to adapt the Bayesian multilevel model such that it was the same form as the generative model in order to provide a better fit to the data. However, this did not have the results that I was expecting and I have come to the conclusion that it was due to a large amount of variance at the individual level. When reducing the noise, the posterior estimates of ρ corresponded more with the original value, suggesting that the model would perform better if the data had a lower σ_y compared to σ_δ and σ_γ. To support this, further research could include finding the FDR and FWER for data simulated with the lower σ_y. A wider range of σ_y could also be used to simulate and fit the data.

While changing the parameter for the LKJ prior had a minor effect, I did not choose extremely different priors to the original one with $\eta = 1$. To examine this further, I would fit the same data to priors with larger values of η to determine whether the weaker correlations perform better. I would also simulate data with a very large amount of correlation, $\rho = -0.98$ for example, and fit to the model with $\eta = 0.5$ to see if the concentration on extreme correlations makes a difference.

Chapter 6

Critical Evaluation

Having conducted the experiments and analysed the results, this section focuses on evaluating the use of Bayesian multilevel models and their practical implications. Furthermore, I assess the use of FWER and FDR in simulation studies.

6.1 Evaluation of the Bayesian Multilevel Model

My simulation studies have provided evidence for the advantage of using Bayesian models to reduce false negatives as well as false positives, in contrast with linear model corrections that produce a large number of Type II errors. The Bayesian model benefits from having more flexibility in that it uses a multilevel structure to create a more complex representation of the data. It also incorporates any prior belief, meaning that the choice of prior distribution may further influence the results.

6.1.1 Complexity

The greater complexity of the Bayesian models results in more parameters to estimate, thus taking longer to fit the data. Furthermore, MCMC sampling, that provides the posterior estimates, is very computationally intensive and time-consuming and I quickly encountered this in my research. This led me to use the High Performance Computing system, BluePebble, to provide greater capabilities than just my own laptop. However, all my studies would benefit from repeated simulations in order to obtain more definitive conclusions. The time-consuming nature of these models may limit their practical use in certain applications in which outputs are required almost instantly.

6.1.2 Priors

The flexibility that comes from choosing the prior distributions can introduce belief about a parameter, but a poorly specified one may negatively affect the fit. This can be seen in 4.3 by the huge increase in Type II errors when there is lots of variation between the groups and a wider prior on σ_δ is used.

In order to model the standard deviations for each parameter, I have been using exponential priors in alignment with McElreath's Statistical Rethinking [25]. These priors generally work well for multilevel models [25] but it is important to note that other distributions can also be chosen. The exponential distribution can cause complications when there are very few groups since the data does not provide much information to estimate the variance between the different groups. Upon reflection, this may have affected fitting to the Bayesian multilevel models throughout my project. My simulation studies contained only eight groups since they were based on the IHDP in which data was collected over eight sites. This meant that the estimated variance between sites is based on little information and may be less accurate. An alternative approach would be to use a more informative prior such as the half-Normal prior which is a Normal distribution that cuts off values below zero [25].

6.2 Evaluation of the Use of Simulations

6.2.1 FDR

Throughout the simulations, I have been able to compare a variety of models with different error rates including FDR and FWER. However, I have not been directly comparing the error values themselves for a singular model, even though I mentioned in the Introduction that the FDR is an approximation of the FWER. This is because their values are equal only when all the null hypotheses are true [5].

When estimating the FDR with the use of simulations, the expectation sign is omitted such that:

$$\text{FDR} = \frac{\text{false rejections}}{\text{false rejections} + \text{true rejections}} \tag{6.1}$$

If I were to simulate with no false null hypotheses, then the estimate of the FDR would always be 1 since there would be no true rejections. Therefore, simulations under the alternative hypothesis have to be taken into consideration. In my simulations, I have been using the same number of true and false null hypotheses, but if the number of simulations under H_1 were to increase, then the number of true rejections would increase. This would result in a lower FDR. In order to produce a better approximate of the FWER with the FDR, I could simulate only 10 samples under the alternative hypothesis such that there were fewer true rejections. That way, I could compare the FWER and FDR more closely and see how the methods of corrections result in different error rates.

The FDR may not be the best criterion to use for simulations when focusing on the value and how it compares to FWER. However, if we only wanted to look at the pattern that arises from comparing models, then it is still useful.

6.2.2 Simulating under the Alternative Hypothesis

In my simulation studies, I have been using $\mu_\delta = 10$ to simulate data under the alternative hypothesis, since it was a rough approximation of the mean of all δ_j for all sites j. However, I could have chosen any value of $\mu_\delta \neq 0$ to be in correspondence with H_1. This could have led to dramatically different values for the error rates obtained. For example, for $\mu_\delta = 5$, there would be more false negatives and fewer true positives, since the uncertainty intervals would be more likely to include 0. This would lead to an increased Type II error rate and FDR. As before, it is important in these simulation studies that the values themselves are not the main focus but rather the pattern between them.

6.2.3 FWER

When analysing models based on the FWER, it is often difficult to spot any patterns in the errors. This is due to the very small errors that are obtained by some models, as well as an insufficient number of simulations of the data. With only 100 samples, as in 4 and 5, the FWER can only differ by 0.01 since one set of data can either produce at least 1 rejection or not. When looking at Bayesian multilevel models with correlation in 5, the errors produced were all less than 0.08 and so analysing models when they only differ by small amounts may not provide a valid comparison. The FDR can be represented with greater precision which suggests that simulation studies should use the FDR to compare models to obtain more accurate results. In my experiments, I only used 100 samples and so it is still difficult to draw concrete conclusions with the FDR. Therefore, more simulations of the data should be fitted to the models to be able to make inferences with more certainty.

Chapter 7

Conclusion

7.1 My Contributions

In this project, I have demonstrated the importance of the multiple comparisons issue and approaches to correct it with a focus on Bayesian multilevel modelling. This has been done in the context of data from the IHDP in which testing over the eight sites introduces the problem.

The classical linear regression model has consistently led to a high FWER and FDR throughout the project, thus emphasising the need for an adjustment of the model. While the Bonferroni Correction and Benjamini-Hochberg procedure have corrected this, they lead to a higher rate of Type II errors with the Bonferroni correction being especially conservative. However, the Bayesian multilevel model has performed better throughout, with a negligible Type II error rate and low FWER and FDR. This supports the idea of using a Bayesian multilevel model to correct for the issue of multiple comparisons.

I have found that the Bayesian multilevel model performs especially better when there is a large amount of variation between the sites. I investigated how the Bayesian multilevel model can be adapted depending on the variance ratio of the data. I have found that a Bayesian multilevel model with a narrow prior on the group-level variance is most likely to be the best choice when it is suspected that there is lots of group-level variation. However, with less variation, the prior is of less influence and depends on which error rate is the most important to reduce.

I obtained counter-intuitive results when exploring a Bayesian multilevel model that takes correlation into account. It did not perform any better on data with correlation than the original Bayesian multilevel model and I deduced that this was due to the large amount of noise in the simulated data.

7.2 Future Work

7.2.1 Alternative Correlation

While I considered the correlation between δ_j and γ_j, there is also the possibility of site-to-site correlation. For example, suppose the sites are in a circle and each site can access resources from the two sites either side of them. This may lead to similarities in site j's δ_j and the treatment effect of j's neighbours, thus introducing correlation. Further research into this type of correlation between δ_j's could also be done to see if similar effects to the correlation between δ and γ are observed on the models. This would follow the same framework as in 5 but with a generative model and fitted model adapted to include this type of correlation:

$$y_i \sim \texttt{Normal}(\gamma_{j[i]} + \delta_{j[i]} P_i, \sigma_y^2) \qquad \begin{bmatrix} \delta_1 \\ \vdots \\ \delta_8 \end{bmatrix} \sim \texttt{MVNormal}\left(\begin{bmatrix} \mu_\delta \\ \vdots \\ \mu_\delta \end{bmatrix}, \Sigma_2 \right) \tag{7.1}$$

where Σ_2 can be written as:

$$\Sigma_2 = \underbrace{\begin{pmatrix} \sigma_\delta & 0 & \cdots & 0 \\ 0 & \sigma_\delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_\delta \end{pmatrix}}_{\#\text{ of sites} = 8} R_2 \begin{pmatrix} \sigma_\delta & 0 & \cdots & 0 \\ 0 & \sigma_\delta & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \sigma_\delta \end{pmatrix} \quad \text{and} \quad R_2 = \begin{pmatrix} 1 & \tau & 0 & 0 & 0 & 0 & 0 & \tau \\ \tau & 1 & \tau & 0 & 0 & 0 & 0 & 0 \\ 0 & \tau & 1 & \tau & 0 & 0 & 0 & 0 \\ 0 & 0 & \tau & 1 & \tau & 0 & 0 & 0 \\ 0 & 0 & 0 & \tau & 1 & \tau & 0 & 0 \\ 0 & 0 & 0 & 0 & \tau & 1 & \tau & 0 \\ 0 & 0 & 0 & 0 & 0 & \tau & 1 & \tau \\ \tau & 0 & 0 & 0 & 0 & 0 & \tau & 1 \end{pmatrix}$$

with τ represents the correlation of treatment effects between a site and its neighbours.

7.2.2 An Alternative Model

I have been focusing on the Bayesian multilevel model as compared to the frequentist linear regression model. However, a more direct comparison could be to compare the Bayesian model to a frequentist multilevel model [17]. This allows for the hierarchical structure to still be modelled but does not have the flexibility that comes from a Bayesian model. Instead of the classical linear regression model, this would perhaps be more appropriate for the hierarchical IHDP data and result in more concrete conclusions of the effect of using a Bayesian model.

Bibliography

[1] Hervé Abdi. Holm's sequential Bonferroni procedure. *Encyclopedia of research design*, 1(8):1–8, 2010.

[2] Alan Agresti, Barbara Finlay, et al. *Statistical methods for the social sciences*, volume 207. Pearson Prentice Hall Upper Saddle River, NJ, 2009.

[3] Richard A. Armstrong. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics*, 34(5):502–508, 2014.

[4] Yoav Benjamini and Henry Braun. John W. Tukey's contributions to multiple comparisons. *The Annals of Statistics*, 30(6):1576–1594, 2002.

[5] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[6] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[7] John K Best and Andre E Punt. Parameterizations for Bayesian state-space surplus production models. *Fisheries research*, 222:105411, 2020.

[8] Michael Betancourt. A conceptual introduction to Hamiltonian Monte Carlo. 2017.

[9] Douglas Curran-Everett. Multiple comparisons: philosophies and illustrations. *American Journal of Physiology-Regulatory, Integrative and Comparative Physiology*, 279(1):R1–R8, 2000.

[10] Jean-Baptist Du Prel, Gerhard Hommel, Bernd Röhrig, and Maria Blettner. Confidence interval or p-value? Part 4 of a series on evaluation of scientific publications. *Deutsches Ärzteblatt International*, 106(19):335, 2009.

[11] Adelino R. Ferreira da Silva. A Bayesian multilevel model for fMRI data analysis. *Computer Methods and Programs in Biomedicine*, 102(3):238–252, 2011.

[12] Andrew Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515 – 534, 2006.

[13] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

[14] Andrew Gelman, Jennifer Hill, and Masanao Yajima. Why we (usually) don't have to worry about multiple comparisons. *Journal of Research on Educational Effectiveness*, 5(2):189–211, 2012.

[15] Andrew Gelman and Iain Pardoe. Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, 48(2):241–251, 2006.

[16] Jennifer S Gewandter, Shannon M Smith, Andrew McKeown, Laurie B Burke, Sharon H Hertz, Matthew Hunsinger, Nathaniel P Katz, Allison H Lin, Michael P McDermott, Bob A Rappaport, et al. Reporting of primary analyses and multiplicity adjustment in recent analgesic clinical trials: ACTTION systematic review and recommendations. *PAIN*, 155(3):461–466, 2014.

[17] Sander Greenland. Principles of multilevel modelling. *International Journal of Epidemiology*, 29(1):158–167, 2000.

[18] et al. Gross, Ruth T. Infant Health and Development Program (IHDP): Enhancing the outcomes of low birth weight, premature infants in the United States, 1985-1988, 1993.

[19] Julian P. T. Higgins and Simon G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11):1539–1558, 2002.

[20] John K. Kruschke and Torrin M. Liddell. The Bayesian new statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin &amp Review*, 25(1):178–206, 2017.

[21] Mark Lai. Course handouts for Bayesian data analysis class, 2020.

[22] Daniel Lewandowski, Dorota Kurowicka, and Harry Joe. Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, 100(9):1989–2001, 2009.

[23] Matthew D. Lieberman and William A. Cunningham. Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience*, 4(4):423–428, 2009.

[24] David John Cameron Mackay. Introduction to monte carlo methods. *Learning in graphical models*, pages 175–204, 1998.

[25] Richard McElreath. *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press, 2020.

[26] Gale H. Roid. *Stanford-Binet Intelligence Scales: Fifth edition*. Riverside Publishing Compagny, 2003.

[27] Zbynek Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633, 1967.

[28] Arvid Sjölander and Stijn Vansteelandt. Frequentist versus Bayesian approaches to multiple testing. *European Journal of Epidemiology*, 34(9):809–821, 2019.

[29] John D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(3):479–498, 2002.

[30] John D Storey. False discovery rate. *International encyclopedia of statistical science*, 1:504–508, 2011.

[31] Don van Ravenzwaaij, Pete Cassey, and Scott D. Brown. A simple introduction to Markov Chain Monte–Carlo sampling. *Psychonomic Bulletin &amp Review*, 25(1):143–154, 2016.

[32] Nengjun Yi, Shizhong Xu, Xiang-Yang Lou, and Himel Mallick. Multiple comparisons in genetic association studies: A hierarchical modeling approach. *Statistical applications in genetics and molecular biology*, 13:35–48, 2014.