Bayesian Hierarchical Zero-Inflation Models

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Josh Yamamoto

May 2023

Approved for the Division
(Mathematics & Statistics)

Leonard Wainstein

Acknowledgements

I want to thank a few people.

Table of Contents

Chapter 1: Introduction **1**
1.1 Zero-Inflated Data 1
1.2 Clustered Data 3
1.3 Forestry Setting 3
1.3.1 Immediate Modeling Struggles 5
1.3.2 The New Model 6
1.3.3 Building the Model 7
1.4 Looking Ahead 7

Chapter 2: Frequentists and Bayesians **9**
2.1 Bayesian “v.s.” Frequentist: Cryptic Definitions 9
2.2 Worked Example: A better way to learn 10
2.2.1 Frequentist Version 10
2.2.2 Bayesian Version 13
2.3 Takeaways 18

Chapter 3: Methods **19**
3.1 Notation 19
3.2 Model Structure Formalized 19
3.3 Specific Models 20
3.4 Model Fitting: Two Ways 22
3.4.1 Frequentist 22
3.4.2 Bayesian 22
3.5 Evaluation of the Bayesian Model: Posterior Predictive Distribution . 28
3.5.1 Theoretical version 29
3.5.2 MCMC version 31
3.5.3 Combining the Model Predictions 32
3.6 Simultaneous v.s Separate 32

3.6.1 Practical Backing 34
3.7 Posterior Predictive Distribution Finalized 35
3.8 Prediction Intervals 35

Chapter 4: Simulation Study 37
4.1 Aims 37
4.2 Data and Dials 38
4.2.1 Estimands 38
4.2.2 Data Generating Process 38
4.2.3 Dials 40
4.3 Methods 43
4.4 Performance metrics 44

Chapter 5: Results 47
5.1 Expectations 47
5.2 Model Failures 47
5.2.1 Model Regularization through Informative Priors 50
5.3 Root Mean Squared Error 51
5.4 Empirical Variance 52
5.5 Empirical Bias 52
5.6 Coverage 53

Appendix A: The First Appendix 55

Appendix B: The Second Appendix, for Fun 57

References 59

List of Tables

List of Figures

1.1 Zero-Inflated data 2
1.2 Zero-Inflated Forestry data. 4
1.3 The shortcomings of a simple linear regression model in this context. 5

2.1 A visualization of Frequentist confidence intervals. 12
2.2 The Prior, Likelihood, and Posterior for this example all plotted together. 17

3.1 The Gamma(a, b) Distribution 27

4.1 The distribution of the response variable in the population that we generated for our simulation. 39
4.2 Scatterplot of the main covariate against the response variable in Group 1 of our generated population data. 39

5.1 Model Failure rates across all simulation settings. 48
5.2 Percent of Results we would get to use if we only included simulation iterations where none of the models failed. 49
5.3 Model Failure Rate comparison between the Bayesian Normal model with flat priors and with informative priors. 50
5.4 Root Mean Squared Error 51
5.5 Empirical Variance 52
5.6 Empirical Bias 53

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Chapter 1

Introduction

In this thesis, I will concerned with modeling a response variable from various explanatory variables in data that exhibits two distinctive features: (i) the response variable that is “zero-inflated”, and (ii) the data is “clustered”.

1.1 Zero-Inflated Data

As the name suggests, data are canonically classified as being zero-inflated when they contain a significant proportion of zeroes. While it’s hardly ever very productive to spell out a definition for a phrase that is its own definition, I do so here to emphasize the fact that to call data zero-inflated is to only say something very broad about how that data is distributed. There is no commonly accepted cutoff for at what proportion of zeros our data deserves the label zero-inflation, and there is no restriction on the distribution of the non-zero data. While the work done in this thesis concerns zero-inflated data with no constraint on the level of “zeroness”, I do require that the non-zero data is positive and continuously distributed. For example, the response variable might look like this:

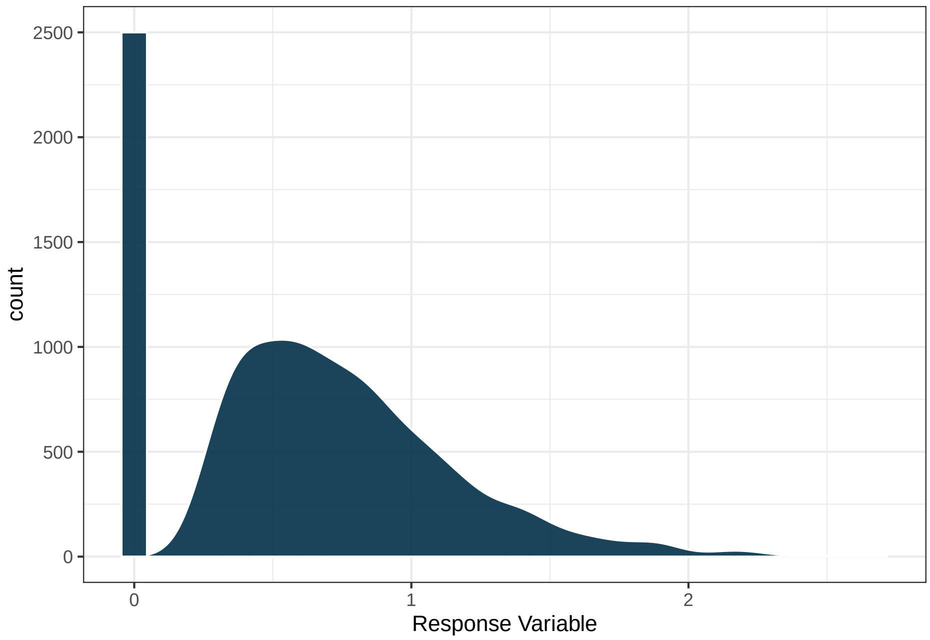


Figure 1.1: Zero-Inflated data

And in reality, this form of zero-inflated data is one that we see quite often in the real world. Importantly, an abundance of zeros in a measured variable might come about for a variety of different reasons. Sometimes it could be a characteristic of the data itself, for example if we collected data on the total weight of fish caught at a lake by individuals on a given day, we likely would see a lot of individuals who caught zero fish leading to a significant portion of zeros in our data. But other times it could be a characteristic of the data collection process itself, for example a measurement error or a sampling error could cause data to be zero-inflated as well.

Importantly, as I am working in a modeling setting, *when I say that data is zero-inflated I mean that the response variable is zero-inflated.* Although I will simply refer to my data as being zero-inflated and my models as being suited for zero-inflated data for the duration of this thesis, this is simply a matter of convenience and not a statement that the methods work for any situation in which data can be considered zero-inflated. Put simply, *I explore, present, and evaluate a model that is suited for a response variable which has a significant portion of zeroes, with the non-zero portion of that variable belonging to a positive continuous distribution.*

1.2 Clustered Data

Furthermore, we will be operating in a setting where the data is not only zero-inflated, but where it also exhibits a clustered structure. Again, the notion of data being clustered is a very non precise one. For this thesis we will not put a very strong restriction on what this looks like. Our data will be clustered in the sense that there is meaningful grouping in the data structure that makes data points within the same cluster more alike on average than points across clusters.

For example, going back to the fishing in a lake example. If we looked at data on the weight of each individual fish caught, we would imagine that fish of the same species would generally be more similar in weight than two fish from different species.

1.3 Forestry Setting

One specific setting that exhibits both of these data features is data on the United States' forests. In particular, I will focus on forestry data collected by the Forestry Inventory & Analysis Program (FIA) of the U.S. Government. The FIA monitors the nation's forests by collecting data on, and providing estimates for, a wide array of forest attributes. Not only is this work vitally important, but it's essential that it be done accurately and efficiently: "The FIA is responsible for reporting on dozens, if not hundreds, of forest attributes relating to merchantable timber and other wood products, fuels and potential fire hazard, condition of wildlife habitats, risk associated with fire, insects or disease, biomass, carbon storage, forest health, and other general characteristics of forest ecosystems."[1].

These sampled locations are referred to as plot-level data and the FIA sends a crew out to physically measure a wealth of forest attributes at that location. As you might expect, not only is this method extremely time intensive, but it is also very expensive. The vastness of the nation's forests in tandem with the resources needed to collect plot-level data, make it impossible to collect census level data on forest metrics. Thus, the need for additional data sources as well as statistical models are vital to the work that the FIA does. The main secondary data source that the FIA employes is remote sensed data. The remote sensed data typically includes climate metrics (e.g. temperature and precipitation), geomorphological measures (e.g. elevation and eastness), as well as metrics like tree canopy cover which can be measured from a satellite.

---

[1] McConville, Moisen, & Frescino (2020)

While the main use of the additional remote sensed data sources are to increase the accuracy of the estimators that the FIA builds, they are also used to make rational decisions about the aforementioned plot-level data collection. Before sending a crew out to a given sampled location, the FIA will first look at the remote sensed data for that location. If that location happens to be in a place where there is clearly no forest, for example in the middle of a parking lot, the FIA will not send a crew out and instead will mark all forest attributes for that location as being zero. As you might imagine, this happens quite a bit, and so an interesting characteristic of many forest attribute variables collected by the FIA is that they are zero-inflated. Importantly, this is an example of where the data is zero-inflated because of the data collection process.

Importantly, the FIA groups the continental U.S. into smaller domains called Eco-Subsections. These Eco-Subsections are drawn with the goal of maintaining internal ecologically homogeneity as best as possible. Thus each data point belongs to a specific Eco-Subsection and it's this grouping that gives us a clustered data structure.

If we look at the distribution of the FIA collected forest attribute "Dry Above Ground Biomass From live Trees", we can see that it is indeed quite zero-inflated.

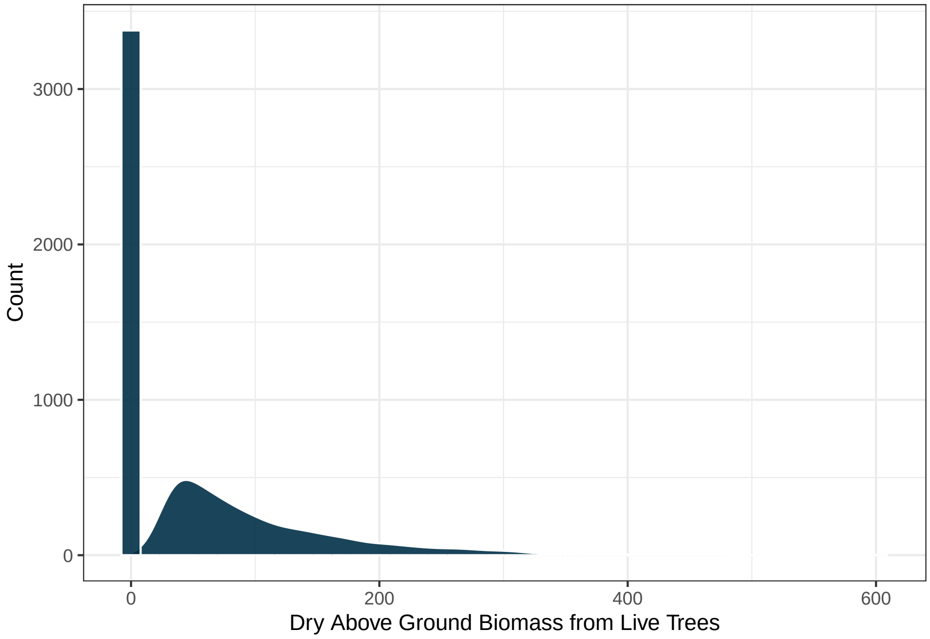


Figure 1.2: Zero-Inflated Forestry data.

Not only is the FIA data a good real life example of when we see this zero-inflated and clustered data structure, but it's also a setting in which it's quite important that the models used to estimate these forest attributes are sufficiently accurate and efficient.

1.3.1 Immediate Modeling Struggles

To motivate using a more complex method, I'll first show what happens when I try to just fit a simple linear regression to this type of data. If we regress our response variable on a useful covariate and plot both the data and the simple linear regression line together we get the following

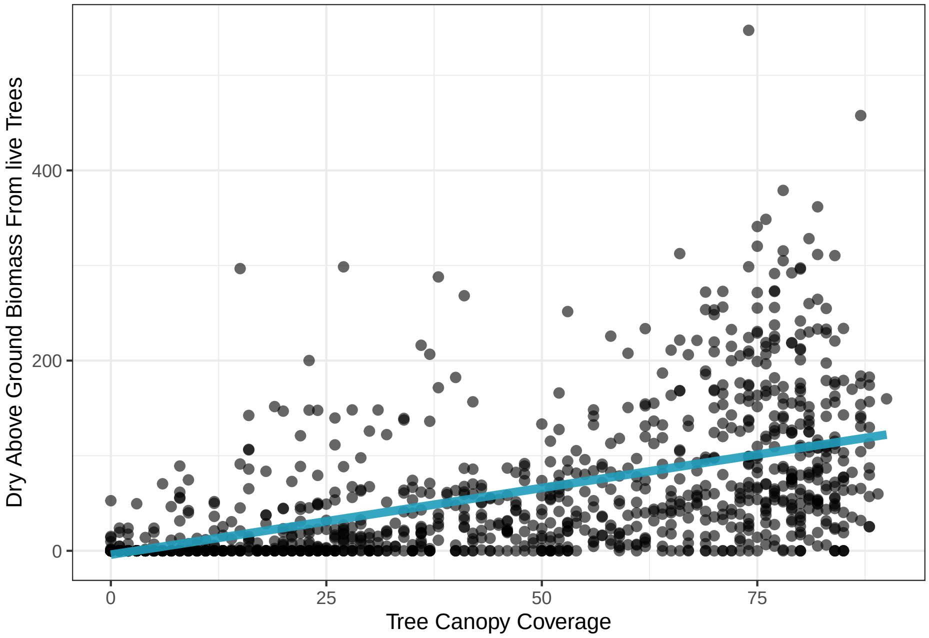


Figure 1.3: The shortcomings of a simple linear regression model in this context.

While this model isn't awful it's certainly misspecified. What I mean by this is that a simple straight line shown in Figure 1.3 doesn't appropriately capture the dynamics of the relationship between our covariate and our response. The zero-inflation in the response variable pulls the regression line down so that it doesn't properly capture the relationship between the explanatory variable and the *non-zero* response, but more importantly it doesn't capture the structure of zeros in the response at all. We can see that the only time this model will predict a near zero response is when the covariate value is very close to zero, but this is an extreme limitation of the model since we observe zero response values across almost the entire range observed values for the covariate. What's more, a simple linear regression model does not allow us to understand how the probability our response variable being zero, changes with our covariate.

We would call this model statistically biased, as it is overly simple and thus doesn't properly capture the structure of the data. While it's perhaps feels unfair to motivate my method by piting it against the simplest of statistical models, the reality is that linear regression is a very powerful and widely used model. Moreover, in a setting such as this one where the data looks plausibly linear, the principle of parsimony might make a linear regression model a well reasoned choice. While there's certainly a need for a model that is better fit to the data, I won't go down the route of constructing an incredibly opaque and complex deep learning model to do so. Instead they model I present is interpretable and intuitive while being flexible enough to capture the structure of the zero-inflated data.

1.3.2 The New Model

While I will exhaustively describe details of, and the math behind, the exact model in a later section, I'll go through a non-technical overview of how it will function here.

The defining characteristic of the model is that it is a two-part model. Instead of trying to fit the data with a singular model, we instead fit two different models and then combine them at the end. The two models are

1. A classification model fit to the entire data set that predicts how likely it is that a certain data point has a non-zero response value.
2. A regression model fit to the non-zero portion of the dataset that predicts the continuous response variable.

To get a final prediction for a data point we take the prediction from model (1) and multiply it by the prediction from model (2).

$$\text{final prediction} = \underbrace{\Big(\text{regression model output}\Big)}_{\text{Model (2)}} \times \underbrace{\Big(\text{how likely it is that that point is non-zero}\Big)}_{\text{Model (1)}}$$

The intuition here is that if our classification model is sufficiently accurate, then points that indeed have zero-response will get sent towards zero due to the fact that the regression output will be multiplied by a number close to zero, while points that have a non-zero response will remain unchanged when multiplied by a number close to one. This method also operates under the idea that a regression model can be well fit to the non-zero portion of the response variable, and thus our regression component should be less biased than if we just fit a single regression model to the entire data set like we described in the previous section.

1.3.3 Building the Model

Now, as the title suggests I'll be building these models in a Bayesian frame. But what does that even mean and why would one want to do that? While most of the thesis will be devoted to answering the second question, I'll spend some time in the next section describing Bayesian methods, and walking through how they differ from a Frequentist approach.

Importantly, since this thesis is simply an earnest exploration of a Bayesian method, it has no intention to participate in the deep and opaque philosophical dialogue regarding whether Bayesian or Frequentist methods are a more "correct" way to do statistics.

That being said, the word Bayesian is so overwhelmingly ideologically tied to this statistical dichotomy that it is, by nature, very difficult to talk about a Bayesian method without talking about Frequentism as well. Because classical statistical methods are all Frequentist ones, there is often a pressure to validate a Bayesian method by standing it next to its Frequentist counterpart. While this Bayesian thesis will indeed feature an alternative Frequentist method, it does so, not to argue for one side or the other, but rather to illustrate some of the key differences in, and logic behind, Bayesian and Frequentist analyses.

1.4 Looking Ahead

In order to introduce, study, and implement these models I will structure the research in the following way:

- Chapter 2 gives a thorough functional overview of how Bayesian and Frequentist methods differ in the simple setting of inference for a mean. The goal here is primarily to provide a gentle introduction to Bayesian data analysis, so as not to drop the reader into the deep end when the main model is introduced.

- Chapter 3 gives a detailed overview of all the methods employed in the thesis. It starts with a high-level description of the Zero-Inflation model, before moving on to detailed descriptions of how each model will be built. Next, prediction for Bayesian models is illustrated both theoretically and computationally. Finally, a mathematical proof is presented to justify building each part of the Bayesian two part model separately.

- Chapter 4 sets up the simulation study that serves as the main process by which we evaluate the various models.
- Chapter 5 showcases the results of each model's performance. Beyond comparing the performance metrics of each model, I also describe the challenges associated with making comparisons between Bayesian and Frequentist models in a complex setting like this one.
- Chapter 6 gives an overview of the R Package written to accompany the methods explored in this thesis. A vignette is provided the applies the R Package to the Forestry data setting.

Chapter 2

Frequentists and Bayesians

2.1 Bayesian "v.s." Frequentist: Cryptic Definitions

Perhaps the biggest roadblock for understanding how a Bayesian methodology differs from a Frequentist one stems from the fact that most of the statements you find on the internet are short cryptic quips that, while true, are largely unhelpful for someone just starting to dig in.

For example, a simple Google search for "Bayesian v.s. Frequentist" will tell you that this statistical philosophic divide is mainly a question of what we mean by probability. The top search result will likely say that for Frequentists, probabilities are fundamentally related to the frequencies of repeated events, while for Bayesians probabilities are related to one's own certainy or uncertainty about events. Again, while this statement is correct and does lead to many of the main functional differences between the two methods, it's nearly impossible to translate this statement into an understanding of how the methods differ in practice.

If you dig a bit deeper and refine your Google search, you'll eventually come across a more technical definition such as this one from Gelman:

> "Bayesian statistical conclusions about a parameter θ are made in terms of probability statements. These probability statements are conditional on the observed value of [x], and ... are written simply as $p(\theta \mid x)$ ... It is at the fundamental level of conditioning on observed data that Bayesian inference departs from the approach to statistical inference described in many textbooks, which is based on a retrospective evaluation of the procedure used to estimate θ over the distribution of possible [x]

> values conditional on the true unknown value of θ" [1]

It's not important to understand what this is saying right now, but I include it here because in just a few sentences Gelman fully lays out the core difference between Bayesian and Frequentist methods. While it is not a good entry point for someone just beginning to learn, it will be helpful to return back to portions of this excerpt as we work through an extended example.

2.2 Worked Example: A better way to learn

At a very high level, the fact that one should always return to when comparing a Bayesian and Frequentist methodology is that in an analysis for a parameter θ

- Frequentists treat the observed data as random and the parameter as fixed. Thus they aim to quantify how the data might vary around the fixed (but unknown) parameter value.
- Bayesians treat the observed data as fixed and the parameter as random. Thus they try to quantify how the parameter might vary based on the fixed observed data.

With this in mind, we now turn to a simple inference example.

Suppose we are interested in estimating the average weight of squirrels in a given park, let's call this θ. Moreover, suppose that we want to somehow quantify our uncertainty for that estimate. Suppose that the distribution of the weight for the entire squirrel population in that park is $\mathcal{N}(\theta, 1)$ (we treat the standard deviation as being fixed and known so as to simplify our example) and that we've properly collected a random sample $\{X_1, X_2, ..., X_n\}$ from the population.

2.2.1 Frequentist Version

We choose $\bar{X}$ as our point estimate and because of the Central Limit Theorem we can say that it is distributed $\mathcal{N}(\theta, 1/n)$. And indeed, as laid out above, by using asymptotic theory to place a distribution on the data, we are treating the data as random and the unknown parameter as fixed. Some shifting and scaling tells us that,

$$\frac{\bar{X} - \theta}{1/\sqrt{n}} \sim \mathcal{N}(0, 1)$$

---

[1] Gelman, Carlin, Stern, & Rubin (1995)

Furthermore, properties of the Normal distribution tell us that,

$$P\left(-1.96 < \frac{\bar{X}-\theta}{1/\sqrt{n}} < 1.96\right) = 0.95 \implies P\left(\bar{X} - 1.96 \cdot \frac{1}{\sqrt{n}} < \theta < \bar{X} + 1.96 \cdot \frac{1}{\sqrt{n}}\right) = 0.95$$

So a Frequentist would end up with what is called a 95% confidence interval for θ of:

$$\left(\bar{X} - 1.96 \cdot \frac{1}{\sqrt{n}},\ \bar{X} + 1.96 \cdot \frac{1}{\sqrt{n}}\right) \tag{2.1}$$

Let's pause to ask ourselves what is random in Equation (2.1). For starters, $\bar{X}$ is certainly random since it came from a random sample from the population, but an immediate implication of this is that the interval itself is actually random too. The very first step in our process was to use asymptotic theory (the Central Limit Theorem) to place a distribution on our observed data $\bar{X}$. This randomness in the data thus carries through to our confidence interval and we end up with an uncertainty statement about the **procedure** being performed and not the parameter itself. Different samples will results in different $\bar{X}$s which will result in different confidence intervals.

It's perhaps easiest to understand how to interpret Equation (2.1) through a quick simulation and visualization. If we generated a 100 new samples from the population with the true parameter θ being 5, and computed a confidence interval for each, we could then plot all 100 intervals and count how many of them contain the true parameter.

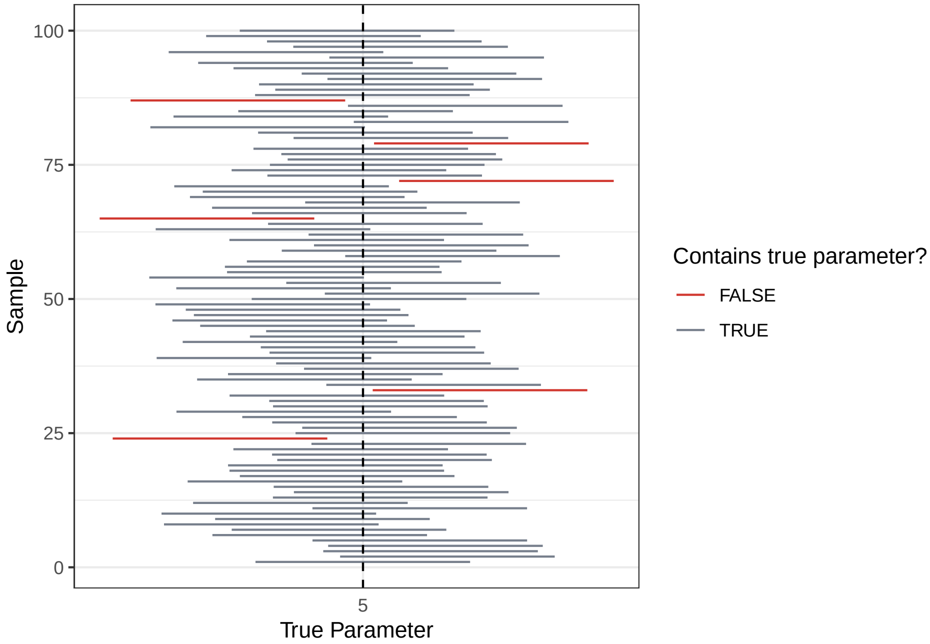


Figure 2.1: A visualization of Frequentist confidence intervals.

First of all, 2.1 really drills home the point that Frequentists treat the parameter as fixed and try to quantify how the data might vary around it. Here we see that in 6 of our iterations of this sampling procedure, the confidence interval did not contain the true parameter, giving us coverage of 94%. The reason we did not observe 95% coverage, is again due to the fact that the intervals are random and Equation (2.1) is an asymptotic statement about the process of resampling data and computing a new confidence interval. Thus, the correct interpretation of Equation (2.1) is that if we repeated the sampling procedure many times, we'd expect the true mean, θ, to be captured by such an interval 95% of the time.

With this new understanding, we can unpack why it's incorrect to interpret Equation (2.1) as saying that "the interval contains the true parameter with probability 0.95". As we just saw, our confidence interval is really a statement about a result we'd see if we repeated the whole procedure many times, and it is absolutely not a statement about any singular instance of the procedure. For a given confidence interval, the true parameter either lies within the interval or it doesn't, with 100% certainty.

Revisiting the Cryptic Definitions

Before moving on to exploring how a Bayesian would tackle this inference problem, think back to the beginning of the section where we described how an internet search might tell you that a Frequentists conceptualize probability in terms of frequency of related events. While this statement is largely unhelpful on it's own, it actually becomes quite helpful when taken together with the example that we've just walked through. Through a Frequentist lense, the probability statement in Equation (2.1) must be conceptualized in terms of frequenct of related events, which in this case is hypothetical resampling of the data.

What's more, while it's still overly complicated in it's language, Gelman's statement that Frequentist inference is "based on a retrospective evaluation of the procedure used to estimate θ over the distribution of possible [x] values conditional on the true unknown value of θ"[2], can at least partially be understood. Each of these pieces are things ideas that we've developed through our example:

- The retrospective evaluation of the procedure used to estimate θ: In the example this was the resampling of our data
- over the distribution of possible [x] values: In our example this involved using the Central Limit Theorem to place a distribution on our data
- conditional on the true unknown value of θ: Treating θ as fixed but unknown.

While these definitions are still found to be lacking when trying to absorb them on their own, in the context of our example, we can start to understand and appreciate the things that they are saying.

2.2.2 Bayesian Version

Instead of first describing the nuts and bolts of how we estimate both the parameter of interest and the uncertainty in that estimate as we did in the previous section, we start at the end with the final expression for both of these things in order to draw some of the most important similarities and differences between the two methods.

General Form

A good place to start whenever performing a Bayesian analysis is to remember that "the guiding principle for bayesian statistics is that the state of knowledge about anything unknown is described by a probability distribution."[3]. In the context of inference where we're interested in an unknown parameter θ, a Bayesian describes all of their knowledge about θ using a probability distribution. In particular, Bayesians use a specific distribution to do so- the posterior distribution. The posterior distribution is the center of interest for all Bayesian analyses and it is simply the distribution of the parameter of interest, conditional on the fixed observed data: $p(\theta \mid x)$.

---

[2] Gelman et al. (1995)

[3] Gelman et al. (1995)

Let's step back for a moment and think about what the immediate implications of this framework are. By describing our knowledge about θ using a distribution we are already doing something very different form in the Frequentist version. Here we are treating our observed data as fixed and our parameter of interest as random. In particular, we make conclusions about the distribution of θ, a.k.a how it varies, conditional on our observed data which we treat as fixed. For a Bayesian, this distribution holds all of the available information about θ and thus is the focus of their attention. So in the context of our example, when a Bayesian wants to perform inference for the mean weight of squirrels in a park, they will do so by constructing a posterior distribution that describes how the *true* mean weight might vary conditional on the fixed observed data.

Once we have an expression for out posterior, we quantify our uncertainty by creating what are called credible intervals. We do so by finding an interval C such that,

$$\int_C p(\theta \mid x) d\theta = 0.95$$

Again, notice how drastically this differs from the Frequentist calculation of an uncertainty estimate. No longer do we rely on asymptotic theory about the randomness of the data sampling process, but instead, since we treat the parameter as random, our uncertainty pertains to the fact that we have uncertainty about what that true parameter is. While estimation of the parameter and uncertainty in that estimate is a two step process in the Frequentist framework (first calculate $\bar{X}$, then use theory to calculate the confidence interval), both things are baked right into the posterior distribution in the Bayesian framework. This is by far one of the most appealing aspects of using a Bayesian method: because we make conclusions in terms of probability statements, we get uncertainty estimates "at no extra cost" in all of our analyses. While increasingly complex Frequentist methods might require increasingly complex procedures for estimating uncertainty, a Bayesian model can be expanded in complexity with no extra work required to acquire uncertainty estimates.

Furthermore, once we compute C, then we can correctly say that.

$$P(\theta \in C \mid x) = 0.95 \tag{2.2}$$

And here we really do mean that the probability that our interval C captures θ is 0.95. **Bayesians conceptualize probability in terms of certainty, or uncertainty, about events, meaning that their probability statements can be about the unknown parameter itself**.

Back to the Cryptic Definitions!

Again, it's the fact that Bayesians conceptualize probability as being related to one's own certainy or uncertainty, that allows us to interpret (2.2) in the way that we do. Furthermore, if we revisit Gelman's quote as well that "Bayesian statistical conclusions about a parameter θ are made in terms of probability statements. These probability statements are conditional on the observed value of [x], and ... are written simply as $p(\theta \mid x)$"[4], we can directly tie it into what we showed above.

Now that we've explained the gist of Bayesian analysis at a high level, we'll dive into the nuts and bolts of how the posterior is actually computed.

Building an Estimate

So how do we actually calculate and estimate $p(\theta \mid x)$? As the name of the framework suggests, we leverage Bayes Theorem as a way to try to quanitfy the posterior distribution. Bayes Theorem tells us that we can break it down into three separate pieces.

$$p(\theta \mid x) = \frac{p(x \mid \theta)p(\theta)}{p(x)}$$

So the problem of quantifying $p(\theta \mid x)$ is really a problem of quantifying these three other pieces. Traditionally $p(x \mid \theta)$ is referred to as the likelihood function, which is treated as being a function of θ, and we can think of it capturing how likely it would be for us to observe the sample data x given a certain realization of θ. Next, $p(\theta)$ describes our belief about θ *before* we have performed any analysis, and thus it is aptly named the prior. $p(x)$ is just a function of the data and thus is referred to as, and treated like, a normalizing constant. Because of this we usually just ignore it and write,

$$p(\theta \mid x) \propto p(x \mid \theta)p(\theta)$$

In plain English, we can imagine a Bayesian approach progressing in the following way. First we supply a prior belief about the unknown parameter. Then, once we observe the data we can generate an expression for the likelihood and can update our belief by multiplying our prior by that likelihood to get our posterior $p(\theta \mid x)$.

---

[4] Gelman et al. (1995)

Importantly, while the likelihood function, $p(x \mid \theta)$, is a function of θ, it comes directly from the data, meaning that there isn't any flexibility in how we represent that term. But perhaps the largest, and most contentious, consequence of describing θ by a probability distribution conditioned on x is that Bayes Theorem forces us to supply a prior distribution $p(\theta)$ ourselves. The reality is that Bayes theorem places almost no restrictions on what $p(\theta)$ could be and this means that in certain cases, drastically different priors can lead to very different posterior distributions. There is a whole body of literature that talks about this "subjective" aspect of a Bayesian analysis, but as we will show later in this thesis, these priors can actually be very powerful in their ability to regularize our analysis. If we have some prior information about θ it makes sense to try to utilize it, and the prior distribution gives us a way to do so.

To clear up what all of this looks like in practice, we'll now walk through our squirrel weights example in this Bayesian setting. As we walk through this process, just remember that at the end of the day, all we're really doing is choosing a prior, computing the likelihood, and multiplying the prior by the likelihood.

We might start by guessing that squirrels might weight around 1.5 pounds on average, and attach a relatively large variance to that guess of 10. So the prior that we supply could be $f(\theta) = \mathcal{N}(1.5, 10)$. While there is a huge literature on how you should choose your priors, for now, all you need to know is that at the very least a prior should extend over the entire range of possible values that your unknown parameter could take on. While we could use a distribution that was strictly positive to emphasize the fact that θ is certainly positive, we'll stick with a normally distributed prior for the sake of simplicity (in fact what makes this a simple choice is that it achieves something called conjugacy which in this case just means that it guarantees that our posterior will also be a normal distribution).

Next we use Bayes Theorem to combine our prior with the observed data:

$$
\begin{aligned}
p(\theta \mid x) &\propto \mathcal{L}(\theta \mid x) p(\theta) \\
&= \left[\prod_{i=1}^{n} \mathcal{L}(\theta \mid x_i)\right] p(\theta) \\
&= \left[\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x_i - \theta)^2\right)\right] \frac{1}{\sqrt{2\pi \cdot 10^2}} \exp\left(-\frac{1}{2 \cdot 10^2}(\theta - 1.5)^2\right)
\end{aligned}
$$

a rather large amount of math will simplify this down to

$$p(\theta \mid x) \propto \frac{1}{\sqrt{2\pi\sigma_f^2}} \exp\left(-\frac{1}{2\sigma_f^2}(\theta - \theta_f)^2\right)$$

Which we can recognize as being a normal distribution with mean θ_f and variance σ_f^2 . In particular θ_f and σ_f^2 are

$$\theta_f = \frac{\frac{1}{10^2} \cdot 1.5 + n \cdot \bar{x}}{\frac{1}{10^2} + n}$$
$$\sigma_f^2 = \frac{10^2}{1 + 10^2 \cdot n}$$

where each have notably been influenced by both the prior and the likelihood. It can be helpful to visualize what has happened here with a plot.

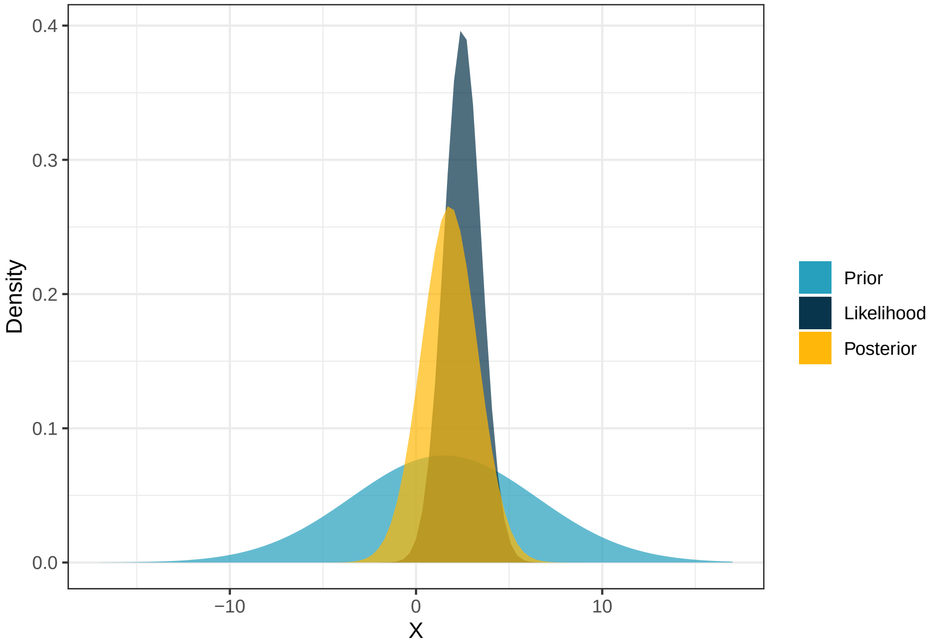


Figure 2.2: The Prior, Likelihood, and Posterior for this example all plotted together.

Above we see the three main components of a Bayesian analysis plotted all together. It's helpful to visualize what's happening here because it really drills home the idea that the posterior is, in a sense, a tradeoff between the prior and the likelihood function. More technically, the posterior is the normalized product of the likelihood and the prior. The result is that the prior information has pulled the posterior to the left and has introduced more variance than from the likelihood function alone.

Clearly, even in this simple example it takes quite a bit of work to derive the exact expression for the posterior. In fact, in practice the posterior is often times something that is incredibly complicated and so we don't even bother trying to simplify the product of the likelihood and the prior down into a distribution that we can recognize. Instead we make use of the massive computational advances that have been made that allow us to algorithmically approximate the posterior. Mainly, a process called Markov Chain Monte Carlo (MCMC) is used due to it's very powerful ability to produce samples from the posterior distribution.

2.3 Takeaways

In summary, in a Frequentist analysis the question being asked is "What sort of $\hat{\theta}$ would we expect to get under hypothetical resampling?", while in a Bayesian analysis the question is "What is our knowledge of θ based on the data and our prior information?"

While this thesis will focus on statistical predictive models rather than inference, the fundamental functional differences remain the same. To build a model in a Bayesian frame is to represent one's knowledge about the model parameters using probability distributions. What's more, predictions are no longer point estimates but rather distributions themselves. As stated earlier, *the state of knowledge about anything unknown is described by a probability distribution.*

This all seems fine and interesting, but at the end of the day the question remains- why bother doing statistical analysis in this way? This thesis should be read as a log of the time I spent trying to answer that question and not as some strongly opinionated piece about the quality of Bayesianism relative to Frequentism.

Chapter 3

Methods

We now move into a thorough description of all of the methods employed in this thesis. In subsections 3.1-3.3 the necessary notation and model structure are introduced. While the rest of the subsections intermittently mention the Frequentist model, they are primarily focused on the Bayesian models. In particular, 3.4 describes how the Bayesian models are fit, 3.5 describes what prediction with a Bayesian model looks like, 3.6 presents a theoretical backing for why the components of the Bayesian two-part model can be fit separately, and finally 3.7 wraps everything up.

3.1 Notation

Let U denote a finite population with N elements. U is broken into J meaningful domains U_j, $j = 1, 2, ..., J$, where each domain U_j is defined as having n_j sample observations. Let $p = 1, ..., P$ index the covariates. Each sample observation, i in domain j has auxiliary information x_{ij}^p for covariate p, response value y_{ij}, and indicator for being non-zero z_{ij}.

$$z_{ij} = \begin{cases} 1 & \text{if} \quad y_{ij} \neq 0 \\ 0 & \text{if} \quad y_{ij} = 0 \end{cases}$$

3.2 Model Structure Formalized

We now introduce the modeling technique that will be the main focus of the thesis. Wonderfully, there is some real mathematical backing for why we might build this model in the way that we do, and I think that looking at the steps draws out a lot of helpful intuition.

Let Y represent the response variable and X represent the covariates. We typically write $\mathbb{E}[Y \mid X]$ to denote the expected value of our response variable conditional on it's covariates. Since the separation of the response into values that are zero and those that are not is a finite partition, we can leverage the law of iterated expectation to expand our model structure:

$$\begin{aligned} E[Y \mid X = x] &= \underbrace{E[Y \mid X = x, Y = 0]}_{= \, 0} P(Y = 0 \mid X = x) + E[Y \mid X = x, \ Y > 0] P(Y > 0 \mid X = x) \\ &= E[Y \mid X = x, \ Y > 0] P(Y > 0 \mid X = x) \end{aligned}$$

Out of this equation comes a wonderful intuition for what our new modeling process will look like, what we end up with is something that is somewhat meta in that it doesn't tell us what the exact model will be, but rather it tells us what the structure of our model should look like. What we do know is that we should have one model that predicts our non-zero response using our covariates $E[Y \mid X = x, \ Y > 0]$, and another that predicts whether our response is non-zero or not, again using the covariates $P(Y > 0 \mid X = x)$.

Think back to the Introduction where we introduced the model as being:

$$\text{final prediction} = \Big(\text{regression model output}\Big) \times \Big(\text{how likely it is that that point is non-zero}\Big)$$

all we've done above is provided a formal theoretical backing for this structure and strategy.

We can imagine this structure functioning as follows: given a new data point we use a model that has been fit on the non-zero training data to predict the value of the response and then we weigh that first output by our predicted probability for whether that point is non-zero or not.

3.3 Specific Models

Although we could model these two parts however we wanted to, we will use a linear regression model for the first part and a logistic regression model for the second. Moreover, because the data we will be working with has a clustered structure (in the Forestry setting that comes from the ecologically homogenous domains), we attempt to capture that by including group-level random effects in both models. The precise models that we will be evaluating will be as follows.

We fit a generalized linear model with random intercepts to the **non-zero** portion of the data (the $*$ helps differentiate this model from our final model). The linear predictor is specified as follows

$$\mu_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + u_j + \varepsilon_{ij} \qquad \text{where} \qquad u_j \sim \mathcal{N}(0, \sigma_u^2)$$

With a link function g^{-1} and a probability distribution for the response, we get the final model as

$$y_{ij}^* = g^{-1}(\mu_{ij})$$

At this point we don't place a specific distributional assumption on the error term ε_{ij}, because this will vary with the specific generalized linear model being employed. To spoil the surprise, the two GLMs that we will look at will be one with identity link and Normal distribution (i.e normal linear regression), and one with the log link and a Gamma distribution. That being said, we keep the notation broad at this point so as to emphasize the fact that this portion of the model aims to capture the structure of the non-zero response, and even though we try to model it using various distributions, this is still the main goal.

Here, $\mathbf{x}_{ij}^T = (x_{ij}^1, ..., x_{ij}^P)$ is a $P \times 1$ vector of covariates, $\boldsymbol{\beta}$ is a $1 \times P$ vector of fixed effects, and u_j is the random effect associated with domain j. Finally, σ_u^2 is the between domain variance parameter. The distribution on ε_{ij} will change depending on the specifics of the glm that we employ.

Next we fit a logistic regression random intercepts model to the full data set

$$P(z_{ij} = 1) = p_{ij} = \frac{1}{1 + e^{-(\mathbf{x}_{ij}^T \boldsymbol{\gamma} + v_j)}} \qquad \text{where} \qquad v_j \sim \mathcal{N}(0, \sigma_v^2)$$

Here $\boldsymbol{\gamma}$ is a $1 \times P$ vector of fixed effects and v_j is the random effect associated with domain j. Again, σ_v^2 is the between domain variance parameter.

We will get into how each of these model pieces are estimated in the next section, but once we have estimates for all of our coefficients, we get our final model by taking the product of these two estimated models.

$$\hat{y}_{ij} = \hat{y}_{ij}^* \cdot \hat{p}_{ij}$$

And we get a prediction for a single domain by averaging the predictions over all samples in that domain

$$\hat{Y}_j = \frac{1}{n_j} \sum_{i \in n_j} \hat{y}^*_{ij} \cdot \hat{p}_{ij}$$

Importantly, while y^*_{ij} is fit on only the non-zero data, and p_{ij} is fit on the entire data, when we make predictions on new data, both models are applied to the entire new data set.

3.4 Model Fitting: Two Ways

3.4.1 Frequentist

As it is not the focus of this thesis, we will not go in depth into how these models are fit in a Frequentist setting. Still, it's important to at least provide a brief summary of how it is most often done.

In most cases (and in particular in most statistical software), Frequentist regression models are fit using something called Maximum Likelihood Estimation (MLE). Very broadly, MLE functions by first assuming that the observed data was sampled from some distribution. Out of that assumption we get a likelihood function $p(\text{data} \mid \text{parameters})$. And finally, as the name suggests, we choose parameter values that maximize the likelihood of the observed data given that parameter. The main gist of what's happening here is that we are answering the question: under what fixed parameter values would we be most likely to see the data that we observed? And we can connect this back to 2, where we talked about how Frequentists treat parameters as being fixed and the data as random. At it's core, MLE is doing just that.

3.4.2 Bayesian

Before diving into the specifics of the Bayesian model fitting, recall that a Bayesian analysis proceeds by treating the data as fixed and the unknown parameters as random. Importantly we still are interested in estimating the posterior $p(\theta \mid \text{data})$, but now we have many parameters of interest $p(\beta_1, \beta_2, \sigma^2_u, ... \text{ etc} \mid \text{data})$, and so the expressions get a bit more complicated. As we get into all of the specifics that we lay out below, always remember that at the core of this process, we aree treating our parameters as random and trying to quantify how they might vary.

We'll start by describing the logistic regression model, before moving on to the two different versions of the generalized linear regression component.

Logistic Regression Component

Again we start by specifying the broad distribution of our response in this model

$$z_{ij} \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\mu_{ij}}}\right) \qquad \text{where} \qquad \mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\gamma} + v_j$$

So far we have done nothing differently than in the usual frequentist formulation of a model, but now instead of treating our model parameters as fixed but unknown, we treat them as random variables and attach priors to them. The random intercepts are given a normal prior centered on zero with hyper-prior σ_v^2.

$$v_j \mid \sigma_v^2 \sim \mathcal{N}(0, \sigma_v^2)$$

Although $\boldsymbol{\gamma}$ and σ_v^2 are different parameters than the one's in the previous model, we attach the same priors to them

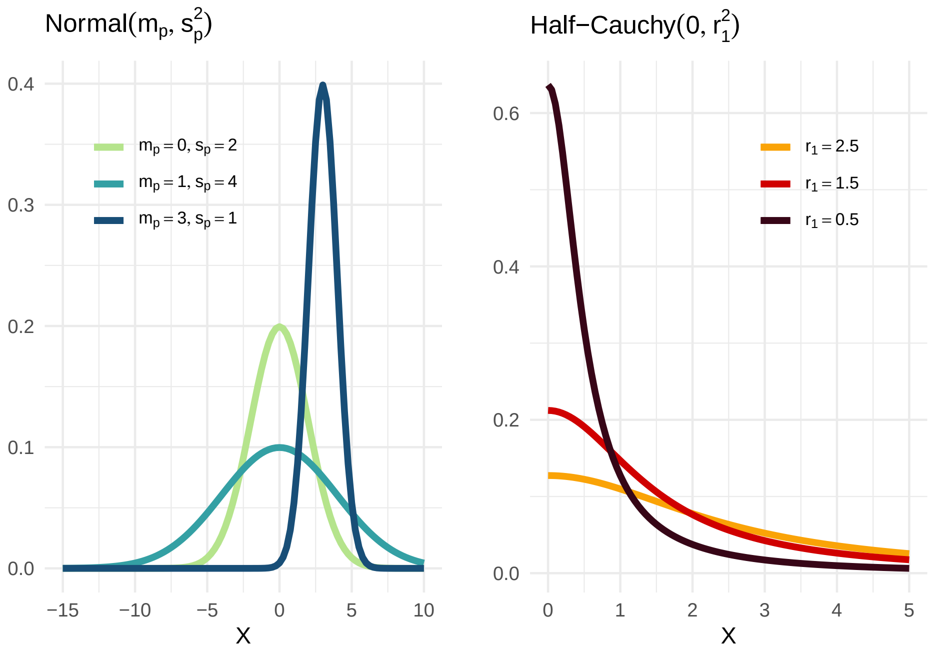
$$\begin{aligned} \gamma_p &\sim \mathcal{N}(m_p, s_p^2) \qquad \forall p \in 1...P \\ \sigma_v^2 &\sim \text{Half-Cauchy}(0, r_1^2) \end{aligned}$$

The prior parameters m_p, s_p^2, r_1^2 are real-valued numbers that center and scale the priors, thus they are chosen with the specifics of the data set in mind. A Half-Cauchy distribution is just a Cauchy distribution bounded to non-negative values, and is broadly utilized as a prior for variance parameters due to the fact that "even in the tail, they have a gentle slope (unlike, for example, a half-normal distribution) and can let the data dominate if the likelihood is strong in that region."[1]. While dependence between priors *can* be modeled in a bayesian frame, we will assume that all of our priors are independent of each other.

An example of what the prior distributions look like across a few of their parameter values is shown below.

---

[1] Gelman (2006)



Again, in estimating the actual **model parameters** in a Bayesian analysis the goal is to recover the joint posterior distribution of those parameters. Let $\boldsymbol{\gamma} = (\gamma_1, ..., \gamma_P)$ and $\mathbf{v} = (v_1, ..., v_J)$. The joint posterior distribution can be written as

$$\begin{aligned} p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2 \mid \mathbf{y}) &\propto \left[\prod_{i=1}^{n} p(y_i \mid \boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2)\right] \cdot p(\gamma_1, ...\gamma_P, v_1, ..., v_J, \sigma_v^2) \\ &= \left[\prod_{i=1}^{n} p(y_i \mid \boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2)\right] \cdot p(\gamma_1) \cdot ... \cdot p(\gamma_P) p(v_1) \cdot ... \cdot p(v_J) p(\sigma_v^2) \end{aligned}$$

If we were able to compute a closed form expression for the posterior, we could then attain posteriors for each of our individual model parameters by marginalizing- i.e integrating out all of the other parameters. For example, we might be interested in the posterior of only σ_u^2. In that case, the marginal posterior can be computed as follows:

$$p(\sigma_v^2 \mid \mathbf{y}) = \int_{\gamma_1} \cdots \int_{\gamma_p} \int_{v_1} \cdots \int_{v_J} p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2 \mid \mathbf{y}) d\beta_1 ... d\beta_P dv_1 ... dv_J$$

The result would be a probability density function that encapsulates all of our information about σ_v^2.

As it turns out, the models are often complex enough that the RHS will not result in a recognizable probability density function. Thus we employ a Markov Chain Monte Carlo (MCMC) algorithm to simulate draws from the approximate posterior. To do so we use the probabilistic programming language “Stan”. The specific version of MCMC algorithm that Stan runs is called "Hamiltonian Monte Carlo". While this thesis will not describe MCMC in depth, the short and sweet description is that an MCMC algorithm's strategy for drawing samples from an unknown probability distribution is to wander around the space in such a way that the amount of time spent in each location is proportional to the height of that distribution. The real nuts and bolts of the algorithm lie in how decisions are made about how and where to move around in the space so that the result is obtained. The especially powerful thing about MCMC algorithms is that under enough iterations, they construct a Markov Chain (random walk) that *has the desired posterior distribution as it's stationary distribution.* Thus we have to be a little bit careful with our language when interpreting the MCMC output. It's not that we are attaining samples from the actual posterior distribution (after all it is unknown), but rather stops along the random walk that is exploring the unknown posterior. That being said, if the MCMC algorithm converges properly then we will have samples from the approximate posterior distribution that should have characteristics similar to the actual posterior.

In it's most basic configuration Stan will output 2,000 sets of parameter draws which represent samples from the approximate joint posterior distribution.

$$\begin{bmatrix} \gamma_1^{(1)} & \dots & \gamma_P^{(1)} & v_1^{(1)} & \dots & v_J^{(1)} & (\sigma_v^2)^{(1)} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \gamma_1^{(2000)} & \dots & \gamma_P^{(2000)} & v_1^{(2000)} & \dots & v_J^{(2000)} & (\sigma_v^2)^{(2000)} \end{bmatrix}$$

One really nice aspect of this is that the while the combination of all of the columns in the output represent samples from the approximate **joint** posterior distribution, each column individually represent samples from the approximate **marginal** posterior distributions for that given individual parameter.

(Generalized) Linear Regression Component: Normal

The simplest way to model the non-zero response is through simple linear regression i.e generalized linear regression using the identity link and assuming a Normal distribution on the response. Again it may seem silly to introduce a simple linear regression model in this way, but we do so to stress that this form of the model still places just as many distributional assumptions as a more common GLM does.

$$y^*_{ij} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2 \sim \mathcal{N}(\mu_{ij}, \sigma_\varepsilon^2) \qquad \text{where} \qquad \mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j + \varepsilon_{ij}$$

In this setting, the error term is assumed to be distributed $\mathcal{N}(0, \sigma_\varepsilon^2)$.

Again the random intercepts have priors

$$u_j \mid \sigma_u^2 \sim \mathcal{N}(0, \sigma_u^2)$$

with hyperprior (i.e priors put on a hyperparameter) σ_u^2.

And the other model parameters are given the same class of priors as before.

$$\begin{aligned}
&\beta_p \sim \mathcal{N}(m_p, s_p^2) \qquad \forall p \in P \\
&\sigma_\varepsilon^2 \sim \text{Half-Cauchy}(0, r_1^2) \\
&\sigma_u^2 \sim \text{Half-Cauchy}(0, r_2^2)
\end{aligned}$$

Importantly while s_p^2, r_1^2, r_2^2 are given the same names as in the previous model, they should usually be chosen with the scale of the response in mind. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_P)$ and let $\mathbf{u} = (u_1, ..., u_J)$.

$$\begin{aligned}
p(\boldsymbol{\beta}, \boldsymbol{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y}) &\propto \left[\prod_{i:y_i>0} p(y_i \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) \right] \cdot p(\beta_1, ..., \beta_P, u_1, ..., u_J, \sigma_u^2, \sigma_\varepsilon^2) \\
&= \left[\prod_{i:y_i>0} p(y_i \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) \right] \cdot p(\beta_1) \cdot ... \cdot p(\beta_P) p(u_1) \cdot ... \cdot p(u_J) p(\sigma_u^2) p(\sigma_\varepsilon^2)
\end{aligned}$$

and we employ MCMC using Stan to simulate draws from it. Similarly, the Stan output will be 2000 draws from the approximate joint posterior distribution

$$\begin{bmatrix}
\beta_1^{(1)} & \dots & \beta_P^{(1)} & u_1^{(1)} & \dots & u_J^{(1)} & (\sigma_u^2)^{(1)} & (\sigma_\varepsilon^2)^{(1)} \\
\vdots & & \vdots & \vdots & & \vdots & \vdots & \vdots \\
\beta_1^{(2000)} & \dots & \beta_P^{(2000)} & u_1^{(2000)} & \dots & u_J^{(2000)} & (\sigma_u^2)^{(2000)} & (\sigma_\varepsilon^2)^{(2000)}
\end{bmatrix}$$

(Generalized) Linear Regression Component: Gamma

An alternative model that we considered in this thesis was a Gamma Generalized Linear Model. The motivation for this was to have a model that is more flexible to the distribution of the non-zero response. Figure 3.1 displays several versions of a Gamma distribution with various parameters. In particular, note that the Gamma distribution is able to capture the fact that the non-zero response can be skewed in a way that a Normal distribution simply can't.

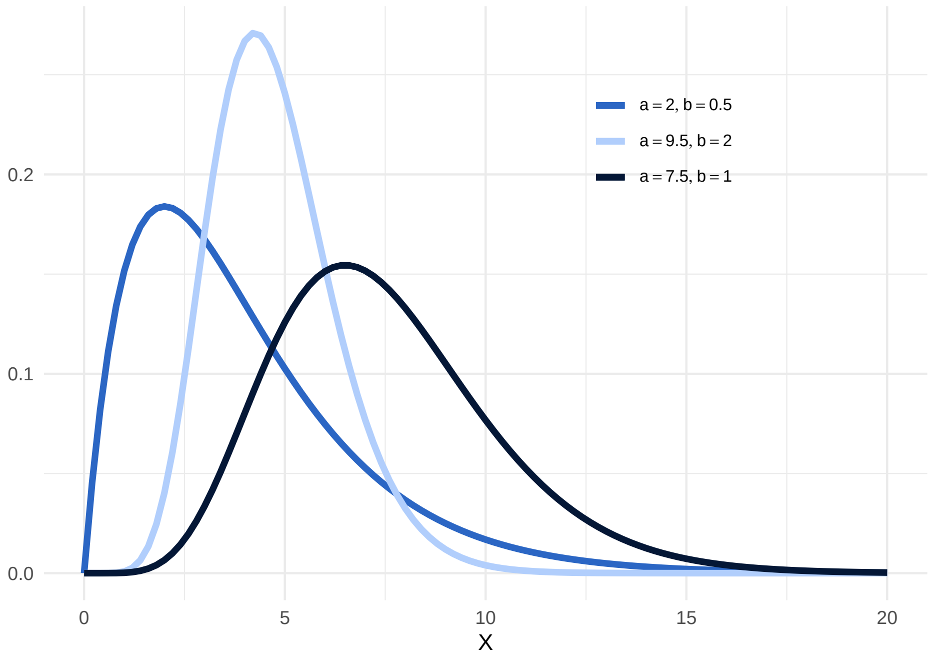


Figure 3.1: The Gamma(a, b) Distribution

Because the response variable we are working with is strictly positive and oftentimes right skewed, it's logical to try to model the response as coming from a Gamma distribution. The model is formulated as follows

$$y_{ij} \sim \text{Gamma}\left(\alpha, \frac{\alpha}{\mu_{ij}}\right) \qquad \text{where} \qquad \mu_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + u_j$$

We are operating under the shape and rate parametrization of a Gamma distribution, and this specific parametrization was chosen so the mean of our response would be the output of the linear component:

$$E[y_{ij}] = \alpha\left(\frac{\alpha}{\mu_{ij}}\right)^{-1} = \mu_{ij}$$

The random intercepts have the same types of prior as in the normal model:

$$u_j \mid \sigma_u^2 \sim \mathcal{N}(0, \sigma_u^2)$$

with hyperprior σ_u^2. We then define priors for the other parameters as

$$\beta_p \sim \mathcal{N}(m_p, s_p^2) \qquad \forall p \in P$$
$$\alpha \sim \text{Half-Cauchy}(t_1, r_1^2)$$
$$\sigma_u^2 \sim \text{Half-Cauchy}(0, r_2^2)$$

where $m_p, s_p^2, t_1, r_1, r_2$ are real valued numbers. While the Cauchy prior put on the shape parameter α is still bounded below by zero, we now allow it's mode to be set based on the given data.

Letting $\boldsymbol{\delta} = (\delta_1, ..., \delta_P)$ and $\mathbf{u} = (u_1, ..., u_J)$ we can write out the joint posterior distribution as

$$p(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \alpha \mid \mathbf{y}) = \left[\prod_{i:y_i>0} p(y_i \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \alpha) \right] \cdot p(\beta_1) \cdot ... \cdot p(\beta_P) p(u_1) \cdot ... \cdot p(u_J) p(\sigma_u^2) p(\alpha)$$

Hopefully at this point the repetitiveness of this process and of these formulations has helped to drill home the Bayesian method of model fitting. We've seen it in three different flavors: Logistic regression, Normal regression, and Gamma regression, but all that's really changed at each step has been the link function and the priors and parameters used.

3.5 Evaluation of the Bayesian Model: Posterior Predictive Distribution

In the Frequentist frame the model parameters are treated as fixed but unknown and so once we obtain estimates for them, we simply use those point estimates to make predictions.

$$\begin{aligned} \hat{y}_{ij}^* &= \mathbf{x}_{ij}^T \hat{\boldsymbol{\beta}} + \hat{u}_j \\ \hat{p}_{ij} &= \frac{1}{1 + e^{-(\mathbf{x}_{ij}^T \hat{\gamma} + \hat{v}_j)}} \\ \hat{Y}_j &= \sum_{i \in n_j} \hat{y}_{ij}^* \cdot \hat{p}_{ij} \end{aligned}$$

But in the Bayesian frame our model parameters are no longer fixed values, but are described by a posterior distribution. Instead of producing predictions that are single values, we construct what are called posterior predictive distributions. In fact it makes sense why we would end up with predictive distributions rather than single values when you consider that there are two main sources of variability that should be taken into account in our predictions:

1. Sampling variability in the data: we never expect our model to be perfectly deterministic, rather the real outcomes should be expected to vary around the model.

2. Posterior variability of the model parameters: we shouldn't go through all the trouble of constructing posterior distributions for our parameters to just throw out that information when it comes time to make predictions, rather we incorporate the variability in our posterior distributions into our predictions.

Within the Bayesian frame, these two sources of variability are combined to produce what is called a posterior predictive distribution.

To get a feel for how this works, I'll just focus on constructing a posterior predictive distribution for a new point $y_{ij,\ new}$ using each model separately.

3.5.1 Theoretical version

To really stress the logic of what we're doing, imagine that we haven't collected any data yet and that we only had one parameter A in our model. In any model we assume that the data for a fixed parameter A has distribution $p(y \mid A)$. Moreover, before having observed any data, all of our uncertainty about the value of A is contained by the prior $p(A)$. We can imagine $p(y \mid A)$ capturing the variability in (1) and $p(A)$ capturing the variability in (2).

To produce an estimate for the distribution of a new data point $y_{ij,\ new}$ we simply integrate the product of the previous two terms over A.

$$\int_A p(y_{ij,\ new} \mid A) p(A) dA = p(y_{ij,\ new})$$

This is sometimes called the prior predictive distribution for $y_{ij,\ new}$ as it represents our knowledge about $y_{ij,\ new}$ before observing any data.

But we can do much better at describing the variability of the model parameters than this. After observing the sample data we update our knowledge. Again we have the same $p(y_{ij,\ new} \mid A)$ which captures the sampling variability in the data, but now the variability of the model parameter is described by the posterior distribution $p(A \mid \mathbf{y})$. Again we get a distribution for $y_{ij,\ new}$ by integrating over A, except this time we end up with $p(y_{ij,\ new} \mid \mathbf{y})$ which is aptly named the **posterior predictive distribution**

$$\begin{aligned}\int_A p(y_{ij,\ new} \mid A, \mathbf{y}) p(A \mid \mathbf{y}) dA &= \int_A p(y_{ij,\ new} \mid A) p(A \mid \mathbf{y}) dA \qquad y_{ij,\ new} \text{ independent of } \mathbf{y} \\ &= p(y_{ij,\ new} \mid \mathbf{y})\end{aligned}$$

Note that the quality of this posterior predictive distribution depends strongly on the quality of our posterior distribution. In other words, this distribution will only accurately capture the structure of new data points, if the underlying posterior distribution correctly captures the structure of the parameters of interest.

As we move on to the expression for each full model, just remember that while the integral looks very complicated, all that we're doing is incorporating both sources of variability and averaging across the possible values of the model parameters. The posterior predictive distribution for a point $y^*_{ij,\ new}$ using the Normal regression model can be written as:

$$p(y^*_{ij,\ new} \mid \mathbf{y}) = \int_{\beta_1} \cdots \int_{\beta_P} \int_{u_1} \cdots \int_{u_J} \int_{\sigma^2_u} \int_{\sigma^2_\varepsilon} \left[p(y^*_{ij,\ new} \mid \boldsymbol{\beta}, \boldsymbol{u}, \sigma^2_u, \sigma^2_\varepsilon) p(\boldsymbol{\beta}, \mathbf{u}, \sigma^2_u, \sigma^2_\varepsilon \mid \mathbf{y}) \right]$$
$$d\beta_1 ... d\beta_P du_1 ... du_J d\sigma^2_u d\sigma^2_\varepsilon$$

Similarly, for the Gamma model it can be written as

$$p(y^*_{ij,\ new} \mid \mathbf{y}) = \int_{\beta_1} \cdots \int_{\beta_P} \int_{u_1} \cdots \int_{u_J} \int_{\sigma^2_u} \int_{\alpha} \left[p(y^*_{ij,\ new} \mid \boldsymbol{\beta}, \mathbf{u}, \sigma^2_u, \alpha) p(\boldsymbol{\beta}, \mathbf{u}, \sigma^2_u, \alpha \mid \mathbf{y}) \right]$$
$$d\beta_1 ... d\beta_P du_1 ... du_J d\sigma^2_u d\alpha$$

And for the classification model it can be expressed as

$$p(z_{ij,\ new} \mid \mathbf{y}) = \int_{\gamma_1} \cdots \int_{\gamma_P} \int_{v_1} \cdots \int_{v_J} \int_{\sigma^2_v} \left[p(z_{ij,\ new} \mid \boldsymbol{\gamma}, \mathbf{v}, \sigma^2_v) p(\boldsymbol{\gamma}, \mathbf{v}, \sigma^2_v \mid \mathbf{y}) \right]$$
$$d\gamma_1 ... d\gamma_P dv_1 ... dv_J d\sigma^2_v$$

The distributions $p(y^*_{ij,\ new} \mid \mathbf{y})$ and $p(z_{ij,\ new} \mid \mathbf{y})$ not only capture where we think each prediction might lie, but also how we would expect it to vary. Often in predictive modeling we're interested in quantifying the uncertainty in our model estimates, and in the Bayesian framework these are baked right into the predictions themselves.

While the theory behind constructing these posterior predictive is pretty intuitive, it's clear that even in the case of a fairly simple model, the actual computations are rather unwieldy. Again, we are saved by the fact that in practice the posterior is too complex to algebraically solve for, so we're already functioning in a setting where we use MCMC to simulate draws from the approximate posterior.

3.5.2 MCMC version

First I will describe how the posterior predictive distribution is derived from the MCMC draws, and then I will explain how it approximates the exact calculation above.

To generate a posterior predictive distribution for a new data point $y_{ij,\ new}$ using the normal regression model we simulate a prediction from the model for each parameter set of the MCMC output.

$$\begin{bmatrix} y^{(1)}_{ij,\ new} \sim \mathcal{N}\left(\mathbf{x}^T_{ij,\ new}\boldsymbol{\beta}^{(1)} + u^{(1)}_j, (\sigma^2_\varepsilon)^{(1)}\right) \\ \vdots \\ y^{(2000)}_{ij,\ new} \sim \mathcal{N}\left(\mathbf{x}^T_{ij,\ new}\boldsymbol{\beta}^{(2000)} + u^{(2000)}_j, (\sigma^2_\varepsilon)^{(2000)}\right) \end{bmatrix}$$

The result is a set $\left\{y^{(1)}_{ij,\ new}, y^{(2)}_{ij,\ new}, ..., y^{(2000)}_{ij,\ new}\right\}$ which approximates the posterior predictive distribution.

For the Gamma model we do the same thing but using the appropriate distribution

$$\begin{bmatrix} y^{(1)}_{ij,\ new} \sim \text{Gamma}\left(\alpha^{(1)}, \frac{\alpha^{(1)}}{\mathbf{x}^T_{ij,\ new}\cdot\boldsymbol{\beta}^{(1)}+u^{(1)}_j}\right) \\ \vdots \\ y^{(2000)}_{ij,\ new} \sim \text{Gamma}\left(\alpha^{(2000)}, \frac{\alpha^{(2000)}}{\mathbf{x}^T_{ij,\ new}\cdot\boldsymbol{\beta}^{(2000)}+u^{(2000)}_j}\right) \end{bmatrix}$$

And we do the same thing for the classification model

$$\begin{bmatrix} z^{(1)}_{ij,\ new} \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma}^{(1)}+v^{(1)}_j\right)}}\right) \\ \vdots \\ z^{(2000)}_{ij,\ new} \sim \text{Bernoulli}\left(\frac{1}{1+e^{-\left(\mathbf{x}^T_{ij}\boldsymbol{\gamma}^{(2000)}+v^{(2000)}_j\right)}}\right) \end{bmatrix}$$

While it may not be immediately clear, these processes are really just mimicking what the massive integrals above were computing exactly. By simulating realizations of the distribution behind each model, we are again capturing the sampling variability in the data, and by doing so across all of our MCMC parameter draws, the uncertainty about the model parameters is being incorporated as well.

3.5.3 Combining the Model Predictions

Now that we have two sets which represent approximate the posterior predictive distribution for unit i in domain j for each respective model, we have to think about how we combine them. After all, our final model prediction is the product of these two models, So we certainly need a posterior predictive distribution of $y_{ij,\ new} = y^*_{ij,\ new} p_{ij,\ new}$, but it's unclear how we should combine the predictive distributions from the individual models to get here. In the Frequentist version where our predictions are single point values, this poses no problem at all, but now that our predictions are themselves distributions, it's a little less clear how to proceed. We might just match MCMC iteration k from each model together, but what makes this matching more correct than shuffling the iterations and then matching them up?

One solution to this conundrum of combining the distributions is to simply build the models simultaneously. In practice this relies on a few tricks and definitely increases the complexity when actually writing code for it, but it can be done and it does allow us to avoid this problem. Moreover, as the two models grow to be more complicated this process of building them simultaneously grows much more difficult and so it isn't a very robust solution to the problem. In the next section we walk through a nice theoretical finding that offers a solution to this problem.

3.6 Simultaneous v.s Separate

As we just mentioned, to get around our problem of how we combine the MCMC iterations for the models built separately, we could fit the models simultaneously. The one major assumption that we will have here is that there is no dependence in the priors **between** models. While there are certainly cases where this doesn't hold, trying to incorporate these dependencies into the model incorporate a lot more complexity without much performance gain[2].

Finally, this result holds regardless of the particular models that we use, but for the sake of simplicity we'll use a logistic regression model with no random effects and a Normal linear regression model with no random effects. With all of that in mind, in this setting our posterior for both models would be:

---

[2] Pfeffermann, Terryn, & Moura (2008)

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2 \mid \mathbf{y}) \propto p(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2) p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2)$$

We can expand this by writing out the likelihood more fully based on whether y is zero or not:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2 \mid \mathbf{y}) \propto \left[\prod_{i:y_i=0} (1-p_i) \prod_{i:y_i>0} (p_i) p(y_i \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) \right] \cdot p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2)$$

While it wasn't too difficult to write this out up to a proportionality constant, in practice it can be very difficult to figure out how to combine the two models in such a way that the MCMC algorithm still converges once you start using models that are more complicated than these ones.

But, there's important insight still to be found here. Let's group these terms based on the parameters that they use. In particular we'll group by which individual model the parameter belongs to:

$$= \left[\Bigg(\prod_{i:y_i=0} (1-p_i) \prod_{i:y_i>0} p_i \Bigg) p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2) \right] \left[\Bigg(\prod_{i:y_i>0} p(y_i \mid \boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) \Bigg) p(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2) \right]$$

Again, we are able to split the joint prior in this way because we are assuming that there is no dependence in the priors **between** models.

But now, if we look at this closely we can see that what we really have here is a full separation into the posteriors for the individual models for p and y^* as seen in our derivation in the previous section. This means that we can write:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2 \mid \mathbf{y}) &\propto p(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y}) p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2 \mid \mathbf{y}) \\ &= C \Bigg[p(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y}) p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2 \mid \mathbf{y}) \Bigg] \end{aligned}$$

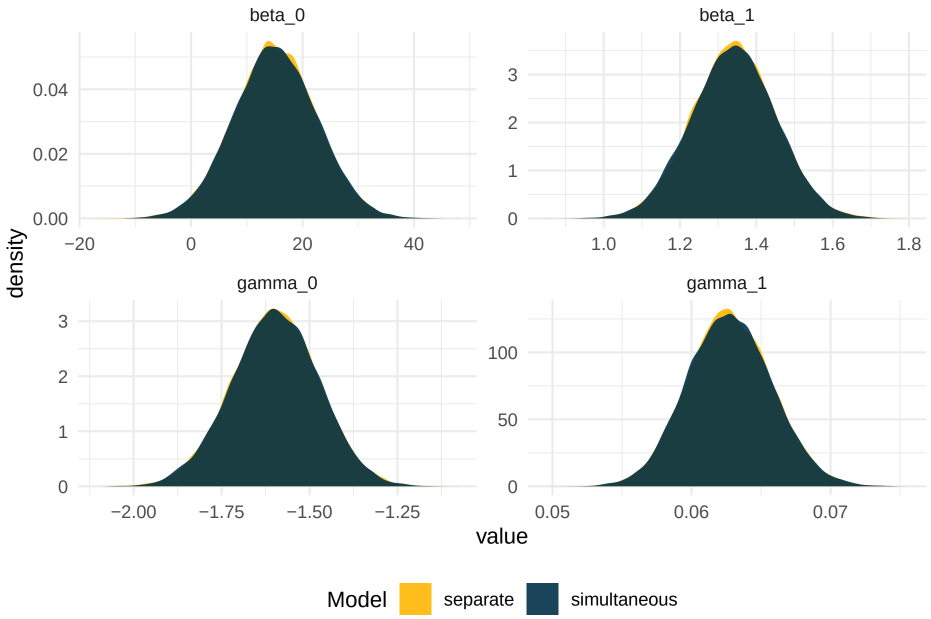
Finally, since these are all proper probability distributions we know that they should all integrate to 1 when integrated across all of their parameters. If we integrate both sides over all of the parameters from both models, its clear that the LHS would be one, and once we recall that there is no parameter dependence **between** the models, it is clear that the RHS does as well. And so we are left with the conclusion that $C = 1$ and thus that

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{u}, \mathbf{v}, \sigma_u^2, \sigma_v^2, \sigma_\varepsilon^2 \mid \mathbf{y}) = p(\boldsymbol{\beta}, \mathbf{u}, \sigma_u^2, \sigma_\varepsilon^2 \mid \mathbf{y}) p(\boldsymbol{\gamma}, \mathbf{v}, \sigma_v^2 \mid \mathbf{y})$$

So the full posterior for the model built simultaneously is equal to the product of the posteriors for each model built separately. The major upshot here is that we can fit the models separately, and then combine the results at the end to get our posterior predictive distributions. In other words *as long as we don't build in any correlations between the parameters in the two models, then we can build each model as complex as we might desire without having to worry about how we will eventually build the two models together. As we learned above, we can simply build them separately and combine the results at the end.*. This is a really nice theoretical finding as it alleviates the need to figure out how to build the models simultaneously and encourages us to have freedom in how we build each individual model.

3.6.1 Practical Backing

To test our theoretical work above, we fit simpler versions of the two models with no random effects. And indeed, when we fit the models simultaneously and also fit them separately making use of MCMC to simulate samples from their approximate posteriors, we find that they are nearly identical.



3.7 Posterior Predictive Distribution Finalized

Now that we know that we can fit the models separately and then combine them at the end, we are finally ready to describe how final predictions are made.

Since we are interested in a making predictions for the average response in domain j we obtain a set that approximates the posterior predictive distribution for each unit i in domain j. For example, if we fix j and let $n_j = 5$ we would have 5 sets:

$$\begin{gathered}\left\{y_{1j,\ new}^{(1)},\ y_{1j,\ new}^{(2)}, \dots,\ y_{1j,\ new}^{(2000)}\right\}\\ \left\{y_{2j,\ new}^{(1)},\ y_{2j,\ new}^{(2)}, \dots,\ y_{2j,\ new}^{(2000)}\right\}\\ \vdots\\ \left\{y_{5j,\ new}^{(1)},\ y_{5j,\ new}^{(2)}, \dots,\ y_{5j,\ new}^{(2000)}\right\}\end{gathered}$$

To get the posterior predictive distribution for $\bar{y}_{j,\ new}$ we take averages across units in domain j by indices of the MCMC draws. In other words we take the average of $\left\{y_{1j,\ new}^{(k)}, y_{2j,\ new}^{(k)}, \dots, y_{5j,\ new}^{(k)}\right\}$ for each MCMC draw k. In full we end up with the set:

$$\begin{bmatrix} \frac{1}{n_j}\sum_{i=1}^{n_j} \hat{y}_{ij,\ new}^{(1)} \\ \frac{1}{n_j}\sum_{i=1}^{n_j} \hat{y}_{ij,\ new}^{(2)} \\ \vdots \\ \frac{1}{n_j}\sum_{i=1}^{n_j} \hat{y}_{ij,\ new}^{(2000)} \end{bmatrix} = \begin{bmatrix} \hat{Y}_{j,\ new}^{(1)} \\ \hat{Y}_{j,\ new}^{(2)} \\ \vdots \\ \hat{Y}_{j,\ new}^{(2000)} \end{bmatrix}$$

Which is an approximation of the posterior predictive distribution for the mean of the response in domain j.

3.8 Prediction Intervals

In statistical modeling, another piece of information that we're often interested is a measure of our uncertainty in our predicted values. Often times these are referred to broadly as prediction intervals, and whereas the confidence intervals and credible intervals that we described in Chapter 2 provide uncertainty bounds for parameter estimates, these prediction intervals provide uncertainty bounds for a future observation or data point.

Immediately, we can see how straightforward it is to acquire these in the Bayesian setting. Because our predictions are not simply point estimates, but are predictions themselves, these prediction intervals are baked right into the prediction process.

Unfortunately it is far less straightforward in the Frequentist setting. While prediction intervals are straightforward to generate for singular regression models, things get a lot more complicated with a two part model like ours. The reality is that we spent a good amount of time trying to construct a bootstrap procedure to generate these prediction intervals and beyond being very computationally intensive, we also couldn't find one that actually worked correctly. While this was certainly frustrating, the struggle to try to develop a process for generating these prediction intervals for the Frequentist models, really highlighted how nice it is that you get them for free in the Bayesian model.

A note of caution

While it might be tempting to gleefully pronounce Bayesian models to be better than Frequentist simply for the ease of access to prediction intervals, the important caveat is that these prediction intervals will only be correct when the Bayesian model is correct. What I mean by that is that if our Bayesian model badly fits the data, then we still get uncertainty estiamtes for free, but a 95% prediction interval will likely not get 95% coverage, thus indicating that the intervals themselves are also incorrect.

Chapter 4

Simulation Study

We now introduce and the set up for the simulation study that will be the main avenue through which we evaluate the various models.

4.1 Aims

The first aim for this simulation study is to understand when we might benefit from using a Bayesian model.

In order to try to answer this question of *when* we will be turning a few dials. Because we are functioning in a setting with grouped data with moderately small sample sizes we will vary the data along two axes

- Number of groups
- Number of observations per group

The second aim is to get a sense for the role that the prior plays in a Bayesian analysis. To do so we will vary the quality of the priors put on the model parameters. The two levels will be

- Uninformative
- Informative

The idea to vary the quality of the priors comes out of the idea that while the likelihood does dominate the prior in large sample size settings, the prior plays a much larger role in smaller sample settings and thus we might expect a lot more regularization and model convergence in the Bayesian models when more information is added through the priors.

4.2 Data and Dials

4.2.1 Estimands

The primary estimand for this simulation study will be Y_j: the mean of our response variable in domain j. In order to stay consistent across all models and simulation runs, we only made predictions on group 1.

4.2.2 Data Generating Process

Our population data for the simulation was generated by the following process that takes as inputs the number of domains (J) and the number of total observations (N). The number of observations per group is thus ($n = N/J$). The population was generated using $N = 50 \cdot 500$ and $J = 50$, to give us 500 observations per group.

The DGP is defined as follows

$$\begin{aligned}
&z_{ij} \sim \text{Bernoulli}(p_{ij}) \quad \text{where} \quad p_{ij} = \frac{\exp\Big(1.5 + X_{ij} + v_j\Big)}{1 + \exp\Big(1.5 + X_{ij} + v_j\Big)} \\
&y^*_{ij} \sim \text{Gamma}(3,\ 1/b_{ij}) \quad \text{where} \quad b_{ij} = 10 + X_{ij} + u_j \\
&y_{ij} = z_{ij} \cdot y^*_{ij}
\end{aligned}$$

Where $X_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$, $v_j \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $u_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$

This data generating process checks all of our boxes as it produces zero-inflated data with a grouped structure and a moderate relationship between the response Y and a covariate X. For example if we examine the generated population data we see the following.

The response variable is indeed zero inflated with a strictly positive and continuous non-zero portion.

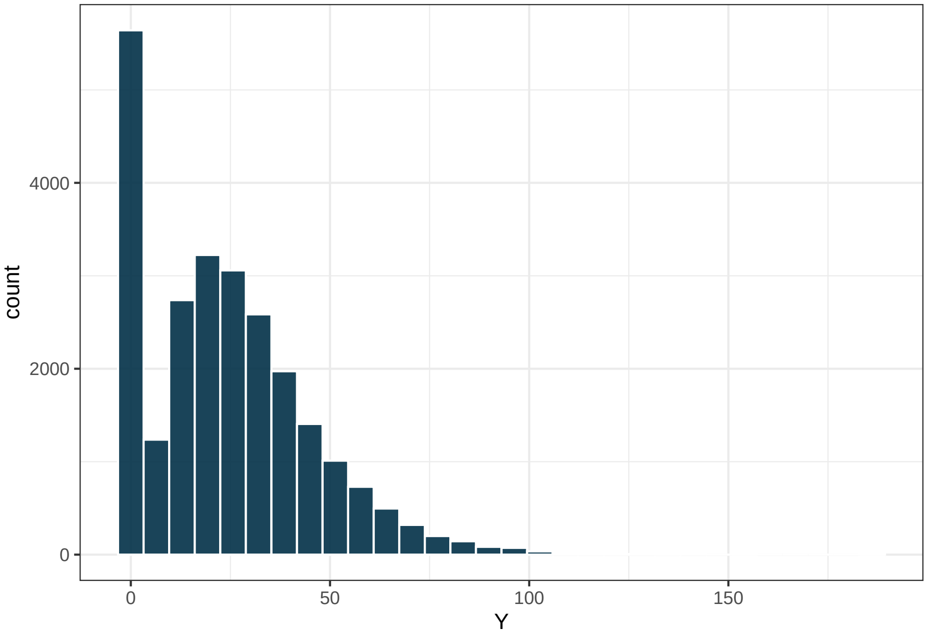


Figure 4.1: The distribution of the response variable in the population that we generated for our simulation.

If we examine the correlation between X and Y we get 0.305, for the entire data set. And indeed when we examine the plot of Y against X we do see a relationship. Here we just plot points from a single group so as not to overclutter the plot.

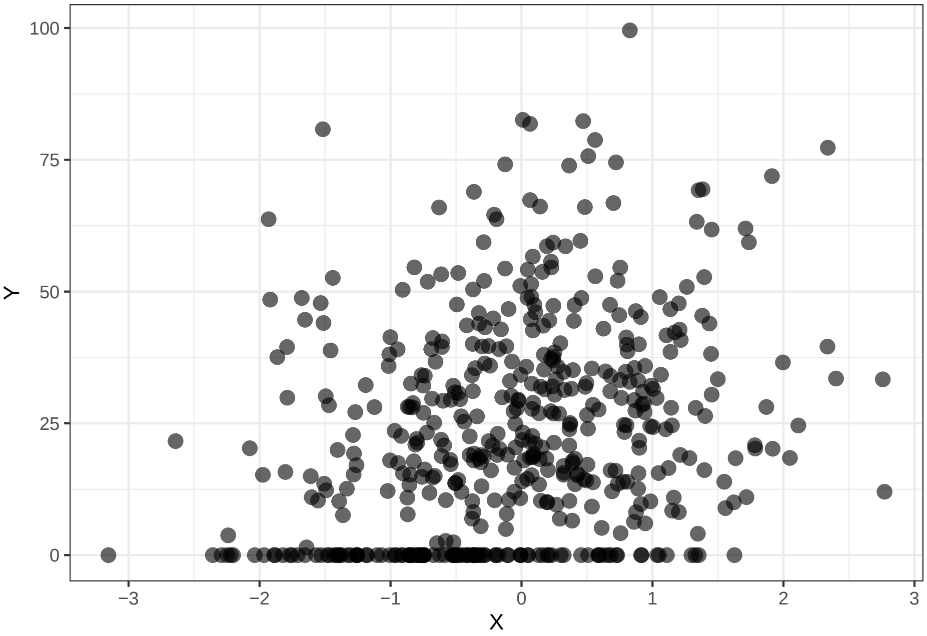


Figure 4.2: Scatterplot of the main covariate against the response variable in Group 1 of our generated population data.

4.2.3 Dials

The exact levels of the dials for number of groups and number of observations per group will be

- n $= \{15, 30, 50\}$
- J $= \{5, 10, 25, 50\}$

We ran a simulation for each combination of those levels, for a total of 12 runs, to get a full sense of how the models perform across each of these measures of how "small" the data is. In order to generate data sets for these settings we first sampled J groups from the population, and then within those groups we sampled n observations. Importantly, because we were always predicting on group 1, we always forced that group into each simulation iteration. For each of the 12 individual settings we ran 400 total iterations of the simulation.

Next, within each $(n,\ J)$ combination we built the Bayesian models with two types of priors: Non-Informative and Informative. Note that across our models we have a variety of parameters that require priors.

- For the Normal regression model we have the the fixed effects coefficients ($\boldsymbol{\beta}$), the variance parameter for the random intercepts (σ_u^2), and the variance parameter for the observation level model error (σ_ε^2).

- For the Gamma regression model we once again have the fixed effects coefficients ($\boldsymbol{\beta}$), and the variance parameter for the random intercepts (σ_u^2), but this time we also have the shape parameter (α).

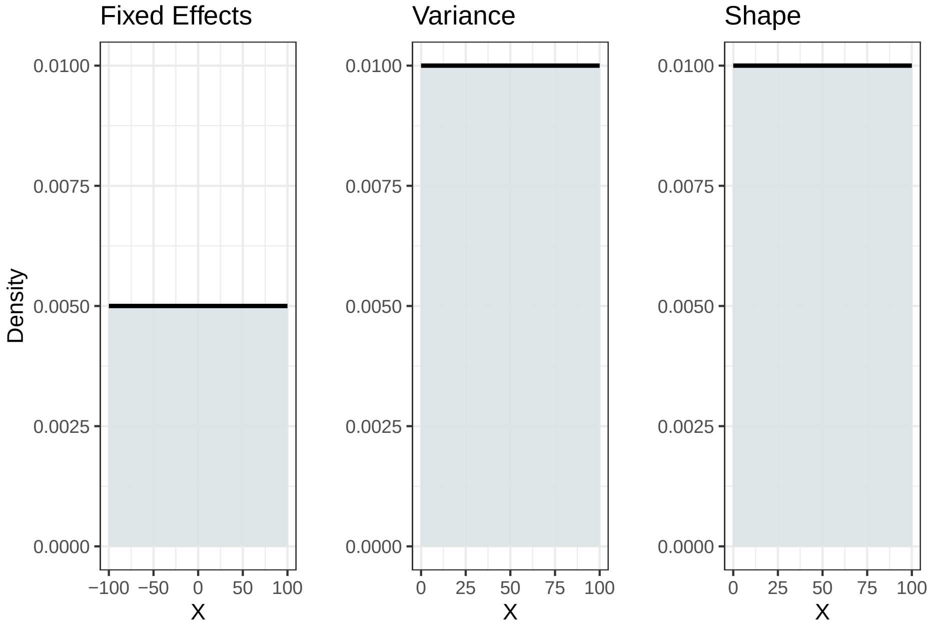
- For the Logistic regression model we have new fixed effects coefficients ($\boldsymbol{\gamma}$) and the variance parameter for the random intercepts (σ_v^2).

Note that across all models the random effects are assumed to be $\mathcal{N}(0, \tau^2)$ distributed and thus can be thought of as universally having a normal prior with a variance hyperparameter. While the distributions chosen for these priors were introduced in the methods section, we will now choose real valued numbers used to center and scale those priors.

These parameters that require priors can be placed into 3 main groups: fixed effects, variance, and shape. We simplified the process of assigning the levels of priors by not varying the priors within these groups within each bracket of quality.

Non-Informative

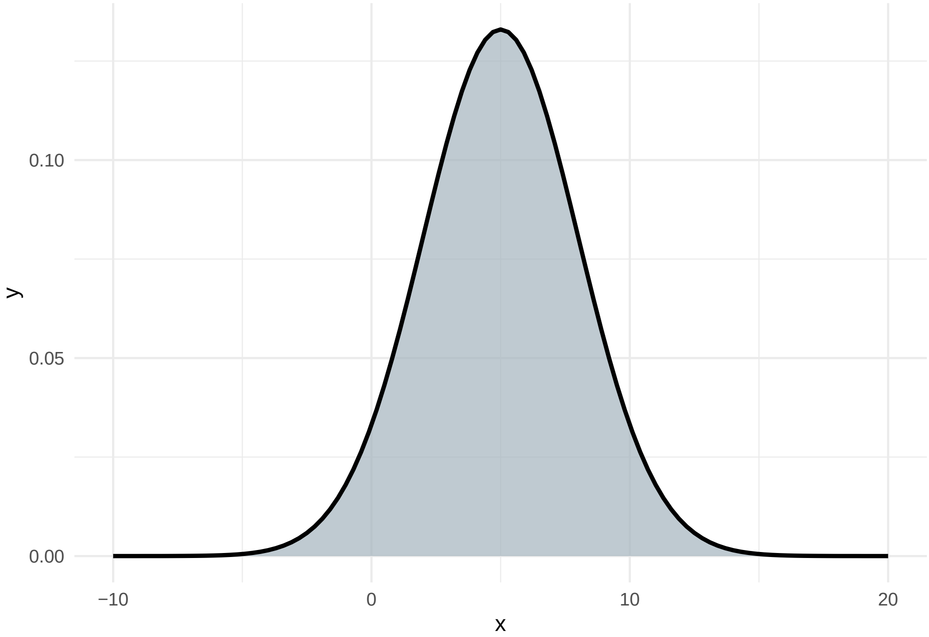
In order to include no prior knowledge in your analysis, a flat prior can be used: $p(\theta) \propto 1$. Recall that the posterior can be factored into a normalized product of the likelihood and the prior. Thus, to use a flat, non-informative, prior is to "let the data speak for itself" and utilize only the likelihood function in estimating the posterior. That being said prior distributions should still not cover parameter values that are impossible, and thus the flat priors will have supports that are restricted to the range of possible values that the parameter could take on. For the fixed effects parameters a support of $(-100, 100)$ was used and for the variance and shape parameters $(0, 100)$ was used.



Informative

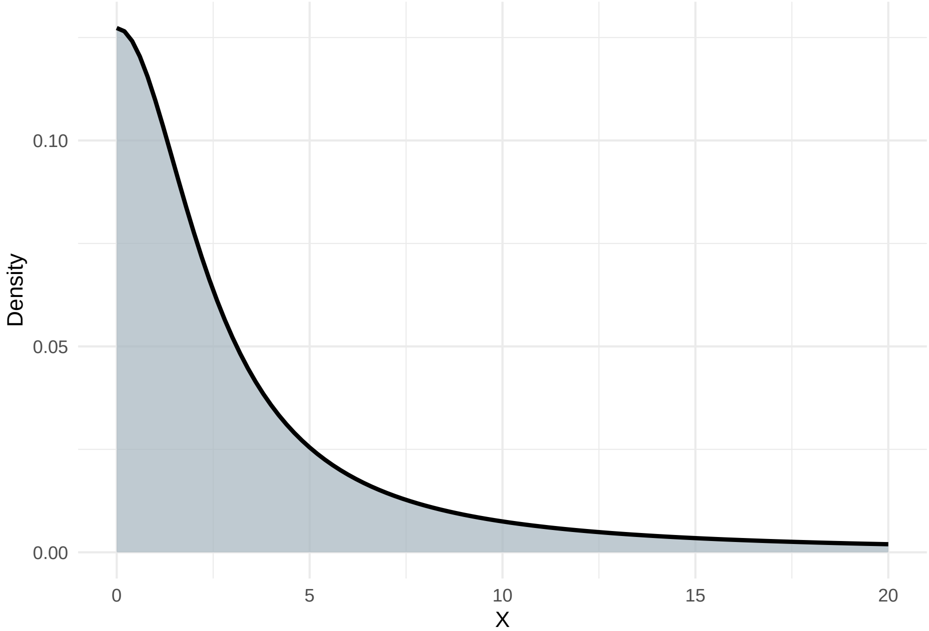
Next, the goal with our informative priors was add more information by choosing to center and scale the prior distributions with the observed data in mind. Naturally, since we know the exact process that generated the data it might be tempting to center the priors exactly on the correct parameter values, but to do so would certainly be both unfair and unrealistic. Instead exploratory data analysis was utilized to make the decisions.

An examination of a scatterplot of the response and covariate leads us to believe that the fixed effects coefficients for the models fitting the non-zero portion of the response are small *positive* numbers between the range of 0 to 10 and thus the prior $\mathcal{N}(5, 3^2)$ was used.



On the other hand, we tweaked the scale of this prior for the logistic regression model since the response being modeled is on a different scale. At this point we don't have much of an idea of how the chance of being non-zero changes as the covariate changes, so we use a $\mathcal{N}(1, 1)$ prior for the fixed effects coefficients in the logistic regression model (this may seem like a tight distribution, but in the context of logistic regression slope coefficients it's really quite wide).

Next, following the advice of Gelman 2013 the scale parameter of 2.5 was used to parametrize the Cauchy prior put on all variance parameters. In full the Half-Cauchy$(0, 2.5)$ prior was utilized.



Finally, since the non-zero portion of the response variable has a mean of 31.12, we can use the following “back of the napkin” calculation to get an idea for what the shape parameter should be centered on. Recall that a Gamma(α, β) distribution has mean α/β and variance α/β^2.

$$\bar{y} = 31.4 \approx \frac{\alpha}{\beta} \qquad \text{and} \qquad s^2 = 338.8 \approx \frac{\alpha}{\beta^2}$$

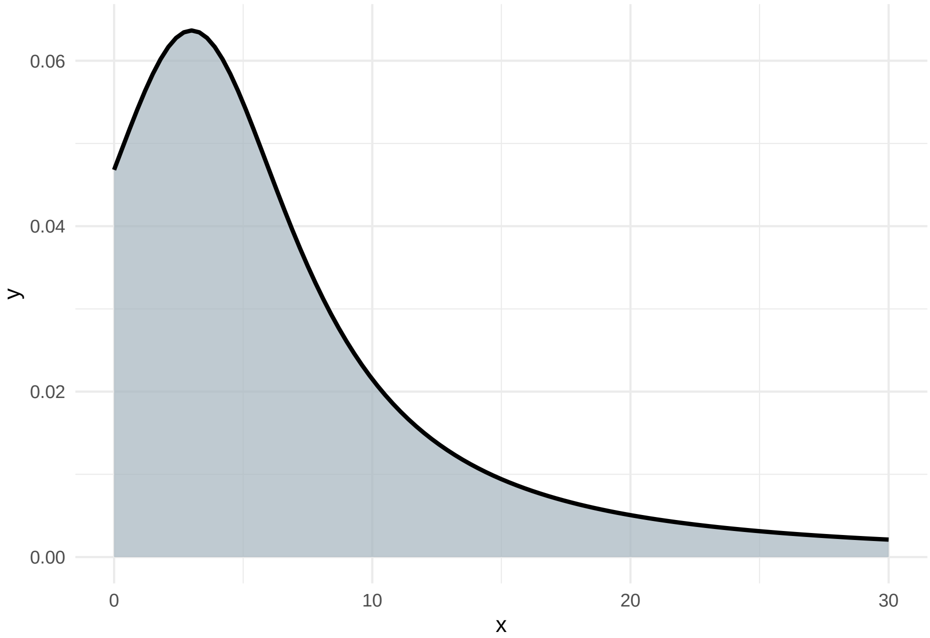
But some algebra tells us then that

$$\frac{\bar{y}}{s^2} = 0.09 \approx \frac{\alpha/\beta}{\alpha/\beta^2} = \beta$$

but now plugging in our approximate value for β gives us

$$\frac{\alpha}{0.09} \approx 31.4 \implies \alpha \approx 2.8$$

And so we might guess that the shape parameter is around 3. Obviously this calculation will vary quite a bit depending on the sample that we generate, so we use the scale parameter of 5 to emphasize our uncertainty in that estimate. Thus the complete prior used was Half-Cauchy$(3, 5)$.



4.3 Methods

There will be 3 different models run throughout the simulation.

- Frequentist two-part:
  - Linear Random Effects Model (`lme4::lmer`)
  - Logistic regression Random Effects Model (`lme4::glmer`)
- Bayesian two-part #1
  - Bayesian Linear Random Effects Model (`rstanarm::stan_lmer`)
  - Bayesian Generalized Linear Random Effects Model (Binomial) (`rstanarm::stan_glmer`)
- Bayesian two-part #2
  - Bayesian Generalized Linear Random Effects Model (Gamma) (`rstanarm::stan_glmer`)
  - Bayesian Generalized Linear Random Effects Model (Binomial) (`rstanarm::stan_glmer`)

4.4 Performance metrics

Importantly, due to the fact that our data has a clustered structure, we chose to evaluate all of our models on a single group. Thus all of the following performance metrics were computed for the same individual group (group 1) which had a true mean value of $Y_1 = 23.7$

The most broadly used and expected performance metric of an estimator is the Mean Squared Error (MSE). We use the Root Mean Squared Error (RMSE) which is computed as follows:

$$\text{RMSE}_j = \sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(\hat{\mu}^y_{j,s} - Y_j\right)^2}$$

Where S is the total number of simulation reps, $\hat{\mu}_{j,s}$ is the estimated mean of the response variable y in domain j for simulation rep s, and $\mu_{j,s}$ is the true mean of the response variable y in domain j for simulation rep s.

Next we will be examining the Empirical Bias of each model:

$$\text{Empirical Bias}_j = E[\hat{\mu}_j] - Y_j \qquad \text{where} \qquad E[\hat{\mu}_j] = \frac{1}{S}\sum_{s=1}^{S}\hat{\mu}_{j,s}$$

As well as the Empirical Variance of each model:

$$\text{Empirical Var}_j = \frac{1}{(S-1)} \sum_{s=1}^{S} \Big(\hat{\mu}_{j,s} - E[\hat{\mu}_j] \Big)$$

We will look at Predictive Interval (PI) coverage at a 95% confidence level for the Bayesian Models:

$$\text{Coverage}_j = \frac{1}{S} \sum_{s=1}^{S} \mathbb{I}(Y_j \in \text{PI}(\hat{\mu}_{j,s}))$$

And finally we will look at the total number of model failures.

- For the Frequentist models, a model failure was registered whenever one of the “Model failed to converge” warnings from `lme4`, although even in small sample sizes our data was robust enough that we essentially saw zero model failures.
- For the Bayesian models, a model failure was registered whenever the MCMC algorithm failed one or more of the diagnostics that indicate that the algorithm has converged and is stable. The Stan programming language is robust enough that you should never trust the results when one of these warnings are administered, and so while potentially overly harsh in some cases, we chose to count any convergence warning as a model failure.

Importantly, one solution to combating MCMC convergence issues is to bump up the number of iterations that the algorithm progresses through. That being said, this is not a sure fire solution, and always comes at the cost of longer computation time. The default number of iterations is 2000 and is considered a reasonable number of iterations for most models. Because the models we are fitting are fairly complex, ran a few tests to guage how many iterations we should use in order to try to minimize model failures while keeping model run-times reasonable. We saw good improvements up to 5000 iterations and then diminishing returns afterwards, so we used 5000 iterations for all of our Bayesian models.

Chapter 5

Results

5.1 Expectations

Before diving into the actual results, we'll first give a brief overview of what we were expecting to see.

We now present the results from the simulation study. Each section compares each model with regards to a specific performance metric as laid out in section 4.5. Across all of these sections we will label simulation settings like $n = \cdot$ - $g = \cdot$. For example if we have $(n = 15, g = 10)$ it's important to remember that this means that the data sets used in that simulation setting had 10 groups each with 15 observations, and thus the models were trained on $n \times g = 150$ data points.

5.2 Model Failures

Before we get into the bulk of the performance metrics, we will first examine how often each model failed in each setting across the 400 simulation reps that we ran.

This grid is organized so that the setting with the smallest sample sizes and smallest number of groups is in the top right hand corner. As we move left and down across the grid we move into settings with larger sample sizes and larger number of groups. We can understand the "size" of our data in each simulation setting to be a combination of both of these dials and thus we have the following structure.

$$\begin{bmatrix} & & \text{smallest} \\ & \swarrow & \\ \text{largest} & & \end{bmatrix}$$

Importantly we omit the Frequentist model in the following plot because it registered no model failures across the entire simulation. The model failure-rate results are as follows:

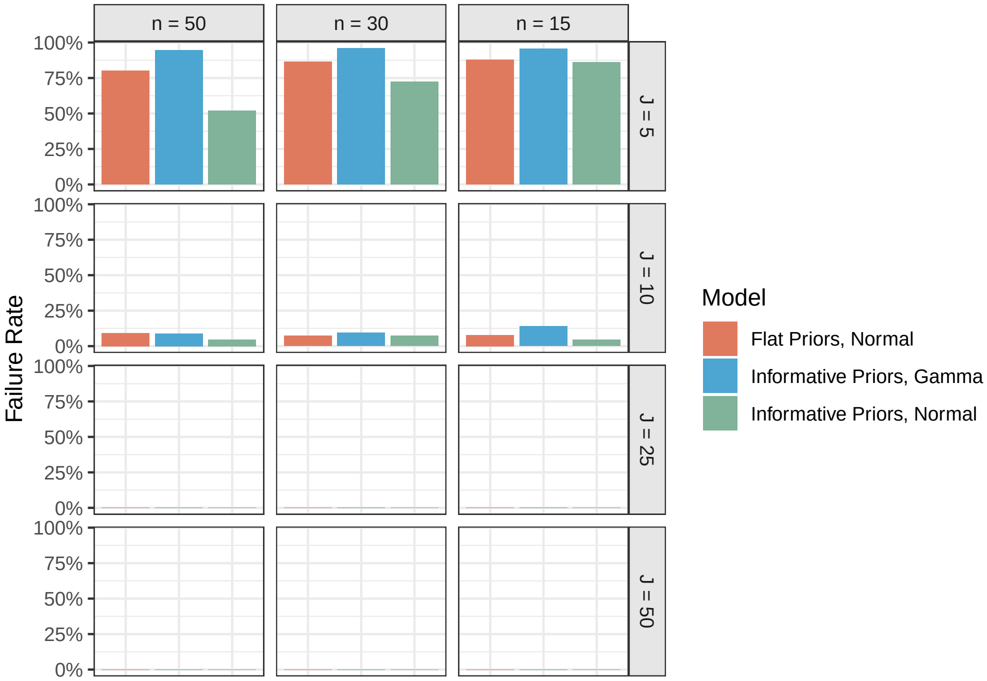


Figure 5.1: Model Failure rates across all simulation settings.

Perhaps the most notable takeaway from Figure 5.1 is that the number of groups that seems to drive failure rates much more than the number of observations per group. Moreover there is a massive jump in model failure rates from $J = 10$ to $J = 5$ that is quite alarming.

What's more, in order to avoid biasing our results, whenever we evaluate some of the models we only do so across simulation reps in which none of the models that we are comparing failed. For example, if we wanted to compare all 4 of the different models using the truncated results in the table below we might be tempted to simply remove individual rows in which the model failed. But doing so would mean that each model would be evaluated on a different number of results. To avoid this issue, we simply disregard all results from Simulation Rep 1 and only use Simulation Reps where none of the models failed (e.g Simulation Rep 2).

Simulation Rep	Model	Prediction	Model Failure
1	Informative Priors, Normal	29.19	0
1	Flat Priors, Normal	29.18	1
1	Informative Priors, Gamma	29.14	1
1	Frequentist	29.05	0
2	Informative Priors, Normal	27.63	0
2	Flat Priors, Normal	27.02	0
2	Informative Priors, Gamma	27.05	0
2	Frequentist	26.98	0

Of course this means that the individual model failure rates have a large impact on how much of the simulation results we are able to utilize. Below, we show what percent of the simulation results we'd be able to retain in each setting.

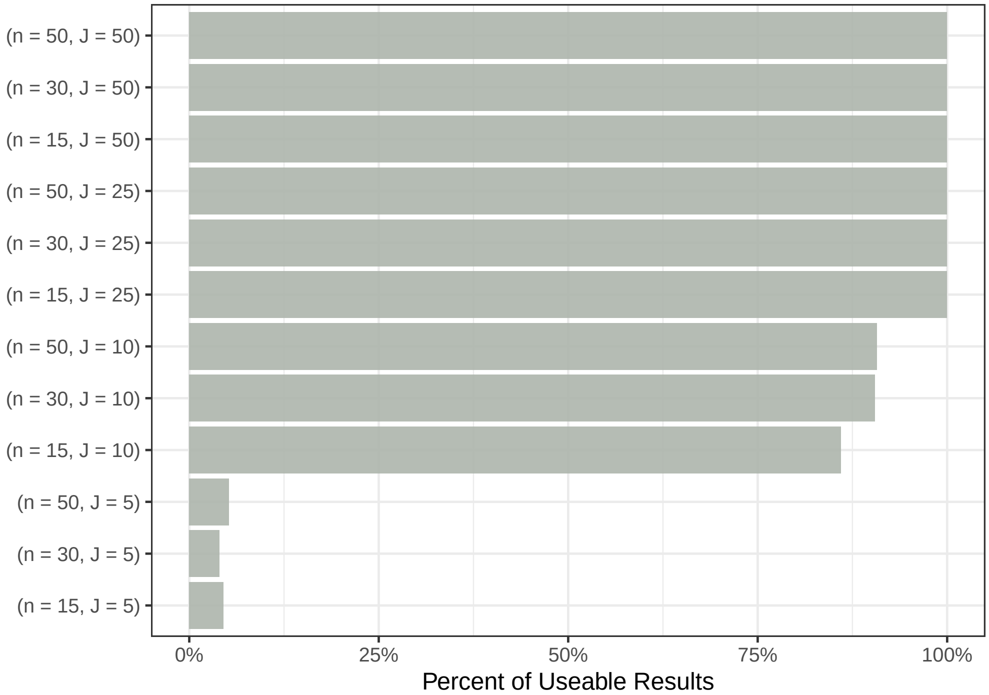


Figure 5.2: Percent of Results we would get to use if we only included simulation iterations where none of the models failed.

Clearly the Bayesian models fail far too much for any trustworthy and meaningful conclusions to be made in any of the $g = 5$ settings and so they will essentially be omited from the rest of the evaluation. While this may seem very disappointing, there is still an interesting takeaway that can be extracted from the high failure rates in those settings.

5.2.1 Model Regularization through Informative Priors

If we only compare the "Flat Priors, Normal" and "Informative Priors, Normal" failure rates, we are able to get a sense for the role that the prior distributions play here. After all these are the same underlying models just with different priors, and so their failure rates give us good insight into the power of priors to stabilize and regularize models.

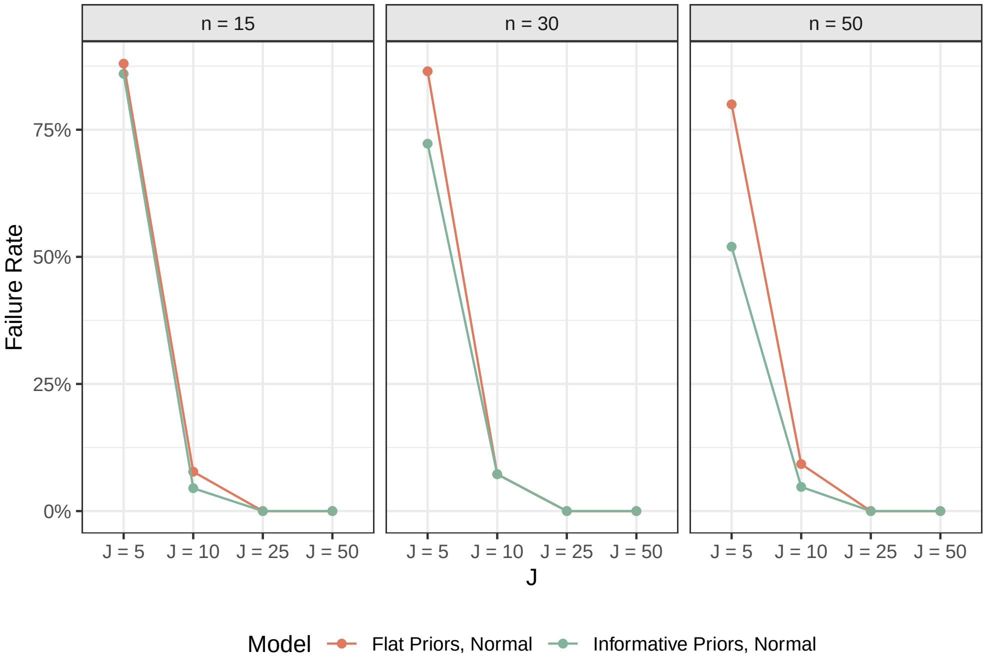


Figure 5.3: Model Failure Rate comparison between the Bayesian Normal model with flat priors and with informative priors.

While in the $J = 25$ and $J = 50$ settings both models exhibit no failures, we do indeed see that in the smaller settings where $J = 5$ and $J = 10$ the model with informative priors does have lower failure rates than the model with flat priors. While informative priors do not completely ameliorate the model failure issues, it's clear that they do have a positive impact on how often a Bayesian model converges. Although we aren't saying anything about the actual performance metrics of the models here, a Bayesian model that does not converge is a model that cannot be used. This is not to say that one should always use informative priors in every Bayesian analysis setting, in fact sometimes you have no prior knowledge to employ. Rather, the lesson here should be that if you are experiencing model convergence issues in a Bayesian analysis setting, one potential fix could be to add more information into your priors.

5.3 Root Mean Squared Error

We begin by comparing the RMSE of our models. As a reminder, due to the high model failure rates, the settings with $g = 5$ are not included in theses evaluations. The performance metrics below are calculated over the simulation reps in which none of the models being evaluated failed to converge (although this is only drops results in the settings where $g = 10$ since we observed no model failures at larger values of g).

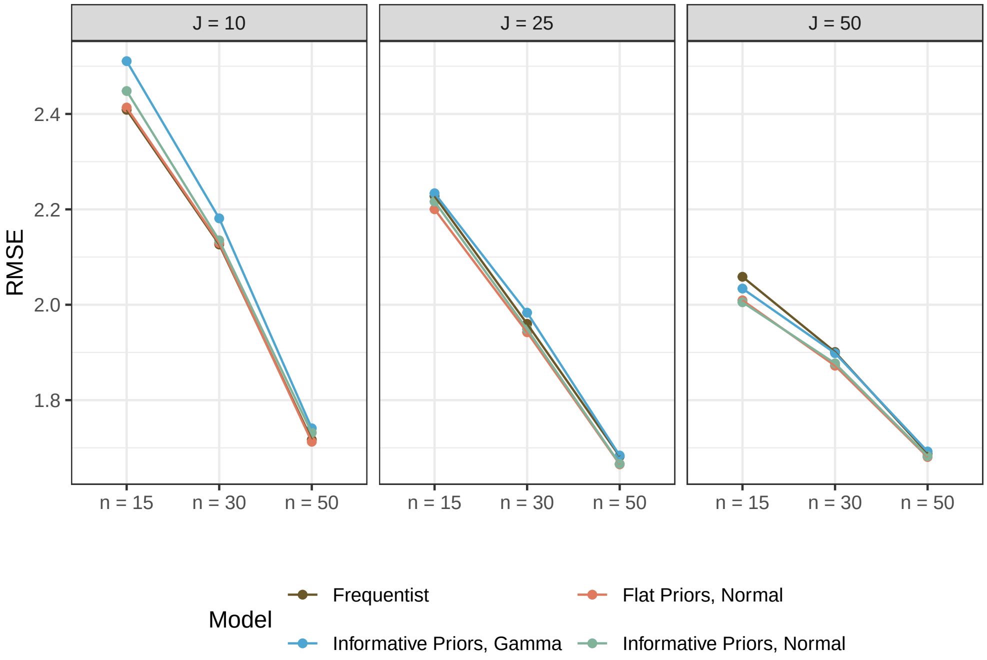


Figure 5.4: Root Mean Squared Error

As we'd expect, we see decreases in RMSE across all models as the number of observations per group and the number of groups increases. Essentially as the model has access to more data, we'd expect it to be able to "learn" the data structure better and thus have lower prediction error. We also see that the models that consistently have some of the lowest RMSE are the Bayesian Normal model with flat priors and the Bayesian Normal model with informative priors.The only setting in which they are outperformed is when $g = 10$ where they are marginally outperformed by the frequentist model. Note that the true value of the response for group one was 23.7 which gives a sense of scale for this model prediction error metric. While there are small differences in how the models performed, the differences are pretty marginal given the scale of our response.

5.4 Empirical Variance

Next we look at the empirical variance which gives a sense for how variable the predictions of the various models were across the different data sets in each simulation run.

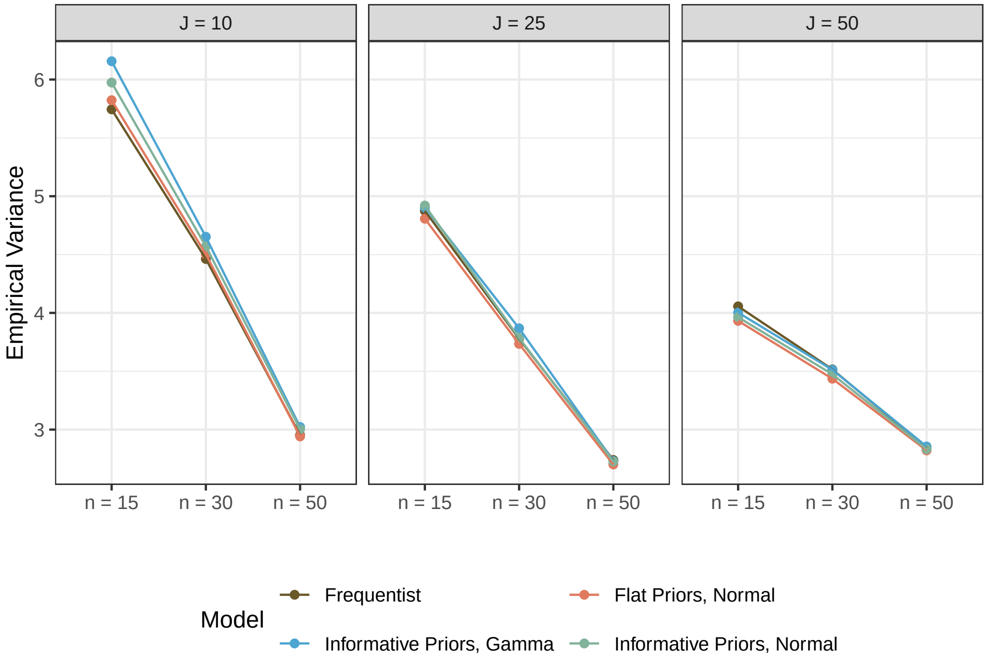


Figure 5.5: Empirical Variance

Again we see the decrease in variance as n and J increase that we'd expect to see. And besides small amounts of separation between the models in the setting $(n = 15,\ J = 10)$, they all seem to perform very similarly.

5.5 Empirical Bias

Next we examine the Empirical Bias of the same three models. These results are by far the most unexpected as we do not see clear decreases in bias as n and J increase:

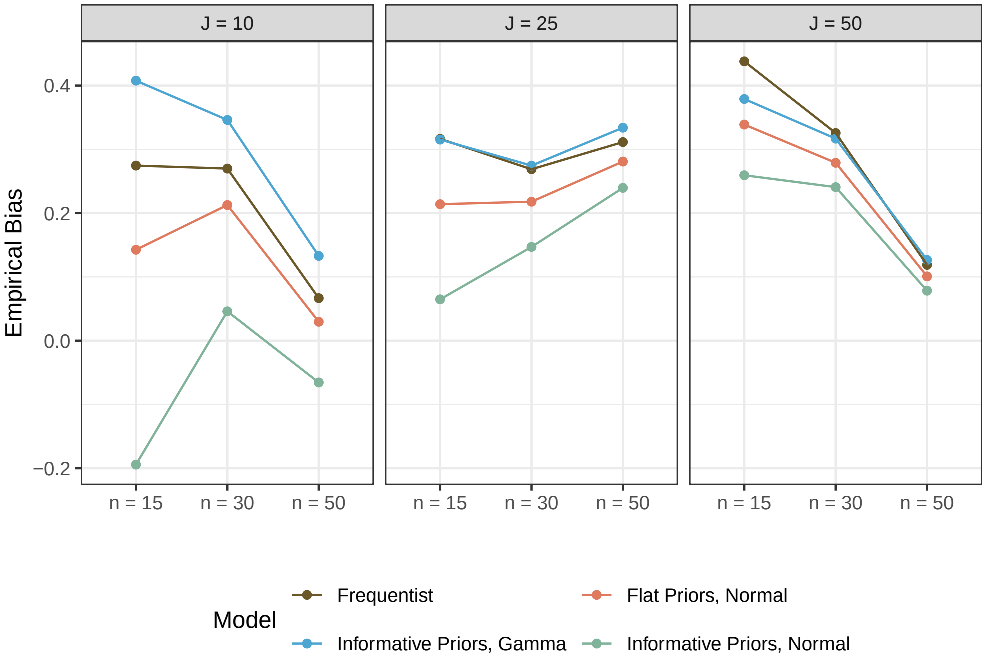
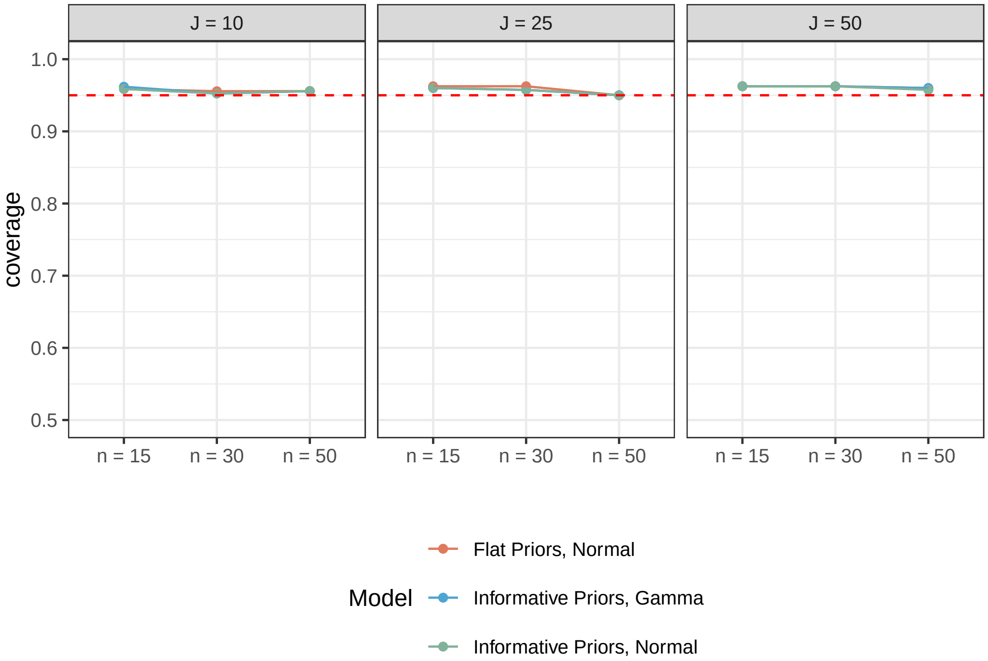


Figure 5.6: Empirical Bias

That being said, we do see that the the Bayesian Normal model with informative priors seems to consistently have the lowest bias, with the Bayesian Normal model with flat priors performing next best. These two models perform more similarly than it may initially seem in the settings where $g = 10$ due to the fact that the Bayesian Normal model with informative priors actually has negative bias in some of those settings. If we looked at the absolute empirical bias we would see that they perform very similarly there. Finally, even though we do see lower bias for the two Bayesian Normal models, note that the scale on the y-axis of these plots is very small. I truly wish that I could take this plot and proclaim the Bayesian Normal model with informative priors to perform far and away the best in terms of bias, but the reality is that the scale of the bias is so small for each model that there really isn't much here. In fact, it's likely this small scale that gives us the patterns that initially were a bit befuddling in Figure 5.6.

5.6 Coverage

And finally, we examine the coverage of the Bayesian models. Note that we do not include the frequentist model here.



Interestingly, while we get close to 95% coverage in all of our Bayesian models, we do see consistent slight over-coverage in almost every setting. While there are a certainly a couple of different reasons why this might have happened, the most prominent one in my mind is that our simulation data was generated in a Frequentist way. What I mean by that is that all of the parameters used to generate the data were fixed constants, and not random variables (as a Bayesian conceptualizes them to be). For this reason all of the Bayesian models are, at least slightly, misspecified.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

```
# This chunk ensures that the thesisdown package is
# installed and loaded. This thesisdown package includes
# the template files for the thesis.
if (!require(remotes)) {
  if (params$`Install needed packages for {thesisdown}`) {
    install.packages("remotes", repos = "https://cran.rstudio.com")
  } else {
    stop(
      paste('You need to run install.packages("remotes")",
            "first in the Console.')
    )
  }
}
if (!require(thesisdown)) {
  if (params$`Install needed packages for {thesisdown}`) {
    remotes::install_github("ismayc/thesisdown")
  } else {
    stop(
      paste(
        "You need to run",
        'remotes::install_github("ismayc/thesisdown")',
        "first in the Console."
      )
    )
  }
}
library(thesisdown)
# Set how wide the R output will go
options(width = 70)
```

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Angel, E. (2000). *Interactive computer graphics : A top-down approach with OpenGL*. Boston, MA: Addison Wesley Longman.

Angel, E. (2001a). *Batch-file computer graphics : A bottom-up approach with QuickTime*. Boston, MA: Wesley Addison Longman.

Angel, E. (2001b). *Test second book by angel*. Boston, MA: Wesley Addison Longman.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper).

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis*. Chapman; Hall/CRC.

McConville, K. S., Moisen, G. G., & Frescino, T. S. (2020). A tutorial on model-assisted estimation with application to forest inventory. *Forests*, *11*(2), 244.

Pfeffermann, D., Terryn, B., & Moura, F. A. (2008). Small area estimation under a two-part random effects model with application to estimation of literacy in developing countries. *Survey Methodology*, *34*(2), 235–249.