ABSTRACT

Bayesian Methods for Hurdle Models

Joyce H. Cheng, Ph.D.

Chairpersons: David J. Kahle, Ph.D. and John W. Seaman, Jr., Ph.D.

Hurdle models are often presented as an alternative to zero-inflated models for count data with excess zeros. They consist of two parts: a binary model indicating a positive response (the "hurdle") and a zero-truncated count model. One or both parts of the model can depend on covariates, which may or may not coincide. In this dissertation, we explore the Bayesian approach to these models in detail, focusing on prior structures.

Many of the Bayesian hurdle models encountered in the literature fail to incorporate expert opinion into the prior structure. We consider how prior information can be elicited from experts and incorporated into the prior structure of a hurdle model with shared covariates through the use of conditional means priors. More specifically, we propose a prior structure that assumes an inherent functional relationship between the two parts of the model. Through simulations, we explore the potential gains, as well as the shortcomings, of the approach. We also consider a simulation algorithm for Bayesian sample size determination for such models. We illustrate the use of the new methods on data from a hypothetical sleep disorder study.

Bayesian Methods for Hurdle Models

by

Joyce H. Cheng, B.A., M.S.

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

David J. Kahle, Ph.D., Co-Chairperson

John W. Seaman, Jr., Ph.D., Co-Chairperson

James D. Stamey, Ph.D.

Jack D. Tubbs, Ph.D.

Corey P. Carbonara, Ph.D.

Accepted by the Graduate School
May 2015

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES vi

LIST OF TABLES ix

ACKNOWLEDGMENTS x

1 Introduction 1

2 Bayesian Hurdle Models 3

2.1 Introduction to Hurdle Models 3

2.1.1 Bernoulli-Poisson Hurdle Model 5

2.1.2 Maximum Likelihood Estimation 6

2.1.3 Bayesian Approach 7

2.1.4 Application: Sleep Disorders 7

2.2 Diffuse Prior Structure 8

2.2.1 Example 9

2.2.2 Operating Characteristics 11

2.3 Informative Prior Structure 13

2.3.1 Conditional Means Priors 14

2.3.2 Conditional Means Priors for the Bernoulli-Poisson Hurdle Model 16

2.3.3 Example 21

2.4 Summary 25

3 A Coupled Prior Structure for Bayesian Hurdle Models 27

3.1 Motivation: Independent vs. Coupled Approaches 27

3.2 Relationship Between the Hurdle and the Count 27
3.3 Coupled Prior Structures 33
3.3.1 Fixed Approach 33
3.3.2 Variable Approach 40
3.3.3 Discussion 44
3.4 Simulation 46
3.5 Summary & Future Work 49

4 Sample Size Determination for Hurdle Models 51
4.1 Introduction 51
4.2 Bayesian Sample Size Determination 51
4.3 Application to the Bernoulli-Poisson Hurdle Model 52
4.4 Simulation 54
4.4.1 Design and Analysis Priors 54
4.4.2 Results 58
4.4.3 Interpretation 66
4.5 Discussion 67

A Data Generation for Sleep Study Example 70

BIBLIOGRAPHY 72

LIST OF FIGURES

2.1 Diffuse Bayesian Model Structure. 8

2.2 Histogram of Generated Weights (left) and Study Outcomes (right) for a Sample of Size $n = 100$. 9

2.3 Posterior Density Plots for the Diffuse Model. 10

2.4 Simulation Results for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ for the Diffuse Model with $n = 50$ (left) and $n = 100$ (right). 13

2.5 Simulation Results for Select Values of θ and λ for the Diffuse Model with $n = 50$ (left) and $n = 100$ (right). 14

2.6 Informative Beta Priors for θ at 200 and 400 lbs. 17

2.7 Simulated Density Plots for the Induced Priors on β_0 and β_1. 18

2.8 Informative Gamma Priors for λ at 200 and 400 lbs. 20

2.9 Simulated Density Plots for the Induced Priors on γ_0 and γ_1. 20

2.10 Bayesian Model with Conditional Means Priors. 21

2.11 Simulated Joint Prior Contours for the Regression Parameters. 21

2.12 Posterior Density Plots for the Informative Model (solid) and the Diffuse Model (dashed). 22

2.13 Posterior Density Plots for the Modified Informative Model (solid) and the Previous Informative Model (dashed). 24

3.1 Prior Information on (a) the Probability of Suffering from a Sleep Disorder and (b) the Number of Sleep Disturbances Suffered by an Afflicted Subject. 29

3.2 Combined Prior Information on θ_i and λ_i. 30

3.3 Prior Information on (a) the Probability of Suffering from a sleep disorder and (b) the number of Sleep Disturbances Suffered by an Afflicted Subject. 31

3.4 Combined Prior Information on θ_i and λ_i. 31

3.5 Bayesian Model With Coupled CMP. 34

3.6 Induced 90% Intervals for λ_i at Specific Weights. 35

3.7 Simulated Density Plots for the Induced Priors on β_0, β_1, γ_0, and γ_1 under the Fixed Coupled Model (solid) and the Independent Model (dashed). 36

3.8 Posterior Density Plots for the Fixed Coupled Model (solid) and the Independent Model (dashed). 37

3.9 Posterior Density Plots for the Fixed Coupled Model with $u = 10.27$ and $v = 0.75$ (solid) Compared to When $u = 12.72$ and $v = 0.48$ (dashed). . . 39

3.10 Informative Priors on u and v. 41

3.11 5,000 Possible Curves Relating θ_i and λ_i. 41

3.12 Bayesian Model with Variable Coupled CMP. 42

3.13 Simulated Density Plots for the Induced Priors on β_0, β_1, γ_0, and γ_1 Under the Variable Coupled Model (solid) and the Fixed Coupled Model (dashed). 42

3.14 Posterior Density Plots for the Variable Coupled Model (solid) and the Independent Model (dashed). 43

3.15 Joint Posterior Contours for β_1 and γ_1 Under the Independent Model (left) and the Variable Coupled Model (right). 45

3.16 Posterior Density for λ/θ at 300 lbs. Under the Variable Coupled Model (solid) and the Independent Model (dashed). 46

3.17 Simulation Results for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$. 48

3.18 Simulation Results for Select Values of θ and λ for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$. . . . 48

3.19 Simulation Results for λ/θ at 300 lbs. for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$. 49

4.1 Normal Distributions Fitted to Simulated Densities for β_0, β_1, γ_0, and γ_1. 55

4.2 Design Priors for Weight Covariate. 56

4.3 Posterior Samples for λ_i in Four Instances of the 200 Replications for $n = 100$ with $x_i \sim \beta(1.7, 3)$. 59

4.4 Density of 200 Interval Lengths for λ_i at 225 lbs. with $x_i \sim \beta(1.7, 3)$. . . 60

4.5 Median Interval Widths Over 200 Replications for λ_i at Various Weights with $x_i \sim Beta_{[200,400]}(1.7, 3)$. 61

4.6 Posterior Samples for λ_i in Four Instances of the 200 Replications for $n = 100$ with $x_i \sim \beta(3, 1.7)$. 62

4.7 Density of 200 interval lengths for λ_i at 225 lbs. with with $x_i \sim \beta(3, 1.7)$. 63

4.8 Median Interval Widths Over 200 Replications for λ_i at Various Weights with $x_i \sim Beta_{[200,400]}(3, 1.7)$. 63

4.9 Median Interval Widths Over 200 Replications for $\log(\lambda_i)$ at Various Weights with $x_i \sim Beta_{[200,400]}(3, 1.7)$. 64

4.10 Realistic v. Unrealistic Analysis Priors for $x_i \sim \beta(1.7, 3)$. 65

4.11 Realistic v. Unrealistic Analysis Priors for $x_i \sim \beta(3, 1.7)$. 65

4.12 Median Interval Widths Over 200 Replications for λ_i at 225 lbs. with $x_i \sim \beta(1.7, 3)$. 67

LIST OF TABLES

2.1 Posterior Means and 95% Credible Sets for the Diffuse Model 11

2.2 Simulation Results for the Diffuse Model with $n = 50$ 12

2.3 Simulation Results for the Diffuse Model with $n = 100$ 12

2.4 Expert Elicited Information on θ 17

2.5 Expert Elicited Information on λ 19

2.6 Posterior Means and 95% Credible Sets for the Informative Model 23

2.7 Posterior Means and 95% Credible Sets for the Informative Model 25

3.1 Posterior Means and 95% Credible Sets for the Fixed Coupled Model . . 37

3.2 center 39

3.3 Posterior Means and 95% Credible Sets for the Variable Coupled Model . 44

3.4 Simulation Results for the Independent CMP Model with $n = 100$ 47

3.5 Simulation Results for the Variable Coupled Model with $n = 100$ 47

4.1 Design and Analysis Priors 55

4.2 Coverage for λ_i at Various Weights Across the 200 Intervals with $x_i \sim \beta(1.7, 3)$ 60

4.3 Coverage for λ_i at Various Weights Across the 200 Intervals with $x_i \sim \beta(3, 1.7)$ 62

ACKNOWLEDGMENTS

Thank you to my advisors, David J. Kahle and John W. Seaman, Jr., for all the support and guidance throughout this process. Additional thanks go to the Baylor University Statistics Department as a whole– faculty, staff, and fellow students, both past and present– and, of course, my friends and family.

CHAPTER ONE

Introduction

In many areas of research, data containing a large number of zero outcomes are common. For example, count data outcomes are common in controlled clinical trials. These counts could be recording occurrences of symptoms, occurrences of adverse events, episodes of risky behavior, etc., over a certain time period. Count data of this sort often exhibit the additional characteristic of zero-inflation, meaning the data include a larger number of zeros than would be expected under a standard count distribution, such as the Poisson, binomial, or negative binomial. However, zero-inflation is not restricted to count data. There are also cases in which data is characterized by a large number of zeros and a positive continuous outcome, often referred to as semicontinuous data.

In the statistics literature, models dealing with excess zeros are often in a count data context, where an important distinction is made in terms of how zeros are interpreted. Generally, there are two types of zeros: structural and sampling. Neelon et al. (2010) give the following example to distinguish between the two types. Consider an outpatient service study where patients may either decline service ($y = 0$) or use services one or more times ($y > 0$). If all the patients in the study will potentially use the service, then the observed zeros are sampling zeros. However, suppose that only a subset "at-risk" group of patients will potentially use the service. The zeros observed in this case come from two sources: patients who do not use the service because they are not at-risk and patients who are at-risk' yet decline service. The zeros resulting from the patients who are not "at risk" are structural zeros.

Two commonly discussed methods of dealing with excess zeros in count data are zero-inflated models (i.e. zero-inflated Poisson, zero-inflated negative binomial) and hurdle models. Zero-inflated models are mixture models consisting of a degenerate

distribution at zero and a full count distribution, including zeros (Lambert, 1992; Neelon et al., 2010). Thus, their structure is conducive to handling two sources of zeros. On the other hand, hurdle models, whose structure will be discussed in detail in this dissertation, are designed to handle only one source of zeros. By this logic, the main distinction made in the literature, for example in Rose et al. (2006), between these two types of models is that, when a study allows for both types of zeros, it is appropriate to use a zero-inflated model, and when a study only has one type of zero, it is appropriate to use a hurdle model.

The focus of this dissertation is on Bayesian methods for hurdle models. In Chapters Two and Three, we define the general form of the hurdle model used in the literature, introduce a contextual example for when these models may be used, discuss what has been done with regards to a Bayesian approach, and propose alternative prior structures that incorporate prior information. In Chapter Four, we consider sample size determination issues for these models using the Bayesian two-priors approach.

CHAPTER TWO

Bayesian Hurdle Models

2.1 Introduction to Hurdle Models

Hurdle models are historically rooted in a model developed by Cragg (1971), which had an econometric context with regards to expenditure or consumption data; they were first developed in a count context by Mullahy (1986) and later Heilbron (1989, 1994). Hurdle models are also known as two-part models, and the econometrics literature covers an abundance of similar models that deal with excess zeros, including the double-hurdle and tobit models, as well as sample selection models, which are somewhat related (Heckman, 1979). We will not delve into these models, but Humphreys (2013) gives a good overview. Similarly, Ridout et al. (1998) provide an overview of count data models for excess zeros.

Hurdle models work under the assumption that the zero counts are generated from a different process than the positive counts (Hilbe, 2011). To define the hurdle model, we follow the general formulation used by Tutz (2012) and Congdon (2005). Suppose f_1 and f_2 are probability mass functions with support $\{0, 1, 2, ...\}$. Let C be a binary variable that determines if the count variable is zero or positive; this threshold is what is referred to as the "hurdle" in hurdle model. When $C = 1$, the hurdle is crossed and a positive outcome is observed, and when $C = 0$, the hurdle is not crossed and a zero is observed. This binary outcome is determined by f_1 such that

$$P(C = 1) = 1 - f_1(0)$$

and

$$P(C = 0) = f_1(0).$$

If the hurdle is crossed, we observe a positive outcome, which is modeled by a truncated count distribution, such that

$$P(y = r|C = 1) = \frac{f_2(r)}{(1 - f_2(0))}, \quad r = 1, 2, ...$$

Following the law of total probability,

$$P(y = r) = P(y = r|C = 0)P(C = 0) + P(y = r|C = 1)P(C = 1),$$

so the hurdle model is defined as

$$P(y = 0) = P(C = 0) = f_1(0) \tag{2.1}$$

and

$$\begin{aligned} P(y = r) &= P(y = r|C = 1)P(C = 1) \\ &= \frac{f_2(r)}{1 - f_2(0)}(1 - f_1(0)), \quad r = 1, 2, ... \end{aligned} \tag{2.2}$$

It follows that the mean is

$$E(y) = \sum_{r=1}^{\infty} r f_2(r) \frac{1 - f_1(0)}{1 - f_2(0)} = \omega \sum_{r=1}^{\infty} r f_2(r),$$

and the variance is

$$Var(y) = \omega \sum_{r=1}^{\infty} r^2 f_2(r) - \left[\omega \sum_{r=1}^{\infty} r f_2(r) \right]^2,$$

where

$$\omega = \frac{1 - f_1(0)}{1 - f_2(0)}.$$

From this, note that the model allows for both under and overdispersion. When $\omega = 1$, $f_1 = f_2$ and the model reduces to the standard count model. However, if $0 < \omega < 1$, then there are excess zeros (overdispersion), and if $\omega > 1$, then there are less zeros than expected (underdispersion).

As noted by Tutz (2012), any specific hurdle model is determined by choices of f_1 and f_2. Generally, f_1 and f_2 can be any discrete probability density, and we can link one or both of them to explanatory variables. Thus, one or both parts can be extended to generalized linear models, which may or may not share covariates.

2.1.1 Bernoulli-Poisson Hurdle Model

Consider a hurdle model where f_1 is Bernoulli and f_2 is Poisson, henceforth referred to as the Bernoulli-Poisson hurdle model. Applying the general hurdle model formula given in equations (2.1) and (2.2), this model is defined as

$$P(y_i = 0) = 1 - \theta_i \tag{2.3}$$

and

$$P(y_i = r) = \frac{\lambda_i^r e^{-\lambda_i}}{r!} \frac{\theta_i}{1 - e^{-\lambda_i}}, \quad r > 0, \tag{2.4}$$

where $0 < \theta_i < 1$ is the Bernoulli probability parameter and $\lambda_i > 0$ is the Poisson rate parameter.

The corresponding likelihood function is

$$L(\theta_i, \lambda_i | \mathbf{y}) = \prod_{i=1}^{n} \left[(1 - \theta_i)^{I_0(y)} \left(\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} \frac{\theta_i}{1 - e^{-\lambda_i}} \right)^{I_{(0,\infty)}(y)} \right], \tag{2.5}$$

where $I_A(y)$ is an indicator function defined as

$$I_A(y) = \begin{cases} 1, & y \in A \\ 0, & y \notin A. \end{cases}$$

Note that both θ_i and λ_i in (2.5) can either be constant or dependent on explanatory variables, resulting in four variations on the Bernoulli-Poisson hurdle model:

(1) Both θ_i and λ_i are constant.

(2) θ_i is constant and λ_i is dependent on explanatory variables.

(3) θ_i is dependent on explanatory variables and λ_i is constant.

(4) Both θ_i and λ_i are dependent on explanatory variables.

Throughout this dissertation, we use variation (4). Specifically, we focus on the Bernoulli-Poisson hurdle model, with likelihood given by equation (2.5), where θ_i is modeled by a logistic regression,

$$\text{logit}(\theta_i) = \mathbf{z}_i' \boldsymbol{\beta},$$

and λ_i is modeled by a Poisson regression,

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\gamma},$$

where $\mathbf{z}_i$ and $\mathbf{x}_i$ represent vectors of explanatory variables, which may have common components, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding regression coefficient vectors. Moving forward, we will often refer to the logistic regression as the "hurdle" part of the model and the Poisson regression as the "count" part of the model.

One artifact of the Bernoulli-Poisson hurdle model, and hurdle models in general, is that the probability of crossing the hurdle θ_i has strong bearing on the effective sample size for the Poisson count part of the model. If θ_i is very small, resulting in very few observations that cross the hurdle, it may result in poor estimation of λ_i. Thus it is important to be aware of the effective sample size for the Poisson part of the model in any data set considered.

2.1.2 Maximum Likelihood Estimation

Frequentist estimation of the parameters in this model is straightforward. From the likelihood function (2.5), we determine the log-likelihood is

$$l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{y_i=0} \log(1-\theta_i) + \sum_{y_i>0} \log\left(\frac{\lambda_i^{y_i}}{y_i!}\frac{\theta_i}{1-e^{-\lambda_i}}e^{-\lambda_i}\right),$$

where

$$\text{logit}(\theta_i) = \mathbf{z}_i'\boldsymbol{\beta}$$

and

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\gamma}.$$

Cameron and Trivedi (1998) and Tutz (2012) point out that the log-likelihood can easily be decomposed into two components: $l(\boldsymbol{\beta}, \boldsymbol{\gamma}) = l_1(\boldsymbol{\beta}) + l_2(\boldsymbol{\gamma})$, such that

$$l_1(\boldsymbol{\beta}) = \sum_{y_i=0} \log(1-\theta_i) + \sum_{y_i>0} \log(\theta_i)$$

and

$$l_2(\boldsymbol{\gamma}) = \sum_{y_i>0} y_i \log(\lambda_i) - \lambda_i - \log(y_i!) - \log(1-e^{-\lambda_i}).$$

This means the joint likelihood can be maximized by simply maximizing each component separately. Maximizing these two components can be done computationally using the R function `hurdle` (Zeileis et al., 2008) in the **pscl** package (Jackman, 2014).

2.1.3 Bayesian Approach

The Bayesian approach to hurdle models has been covered by Congdon (2005) and is widely used in the literature. Some applications include dental caries data (Levin et al., 2009, 2010), pupal population counts (Aldstadt et al., 2011), Verotoxigenic *Escherichia coli* (VTEC) infections data (Jalava et al., 2011), and ecological data (Kuhnert et al., 2005; Martin et al., 2005a,b). In the literature, hurdle models are often mentioned as an alternative to other zero-modified count models such as zero-inflated count distributions and sometimes the zero-altered model. There are a number of papers in the literature that compare these models, some of which do so with a Bayesian approach. For example, Neelon et al. (2010) investigate Bayesian approaches to various zero-inflated models for repeated measures.

The Bayesian hurdle models in the sources cited above all typically use diffuse priors on the model parameters. One exception being that Martin et al. (2005a) and Kuhnert et al. (2005) describe a two-component (hurdle) model with random effects for bird density data, which incorporates expert knowledge in the priors for the random effects. More recently, Neelon et al. (2013, 2014) have developed Bayesian hurdle models with spatially and temporally correlated random effects.

2.1.4 Application: Sleep Disorders

Consider a hypothetical study where we are interested in investigating the relationship between a subject's weight and the number of sleep disturbances they experience over a night.[1] Suppose the subjects of the study are an at-risk group of men classified as obese with weights between 200 and 400 lbs. In the course of the

[1] Although our scenario is hypothetical, studies of sleep disturbances have identified obesity as a potential risk factor. See, for example, Dunson (2005).

study, the subjects' weights are recorded and they are asked to recall the number of sleep disturbances they experienced over the night. From previous knowledge of sleep disorders, we believe this number may range from 4 or 5 up to around 30 disturbances per night's sleep. We also believe that these outcomes are likely to exhibit zero-inflation, as not all subjects in the study will suffer from a sleep disorder, despite being at risk for it.

Using a hurdle model would be appropriate in this scenario because all the men in the study are obese and thus "at-risk" of suffering from a sleep disorder. In this chapter we construct a Bayesian hurdle model for scenarios such as this. We use the hypothetical sleep-disturbance study to illustrate use of our model throughout the dissertation.

2.2 Diffuse Prior Structure

As mentioned in Section 2.1.3, most Bayesian hurdle models in the literature rely on diffuse priors. We demonstrate this approach with a Bayesian Bernoulli-Poisson hurdle model for our hypothetical sleep-disturbance study. To represent the absence of prior information regarding this problem, the typical prior structure places independent diffuse $Normal(0, \sigma^2)$ priors on both sets of regression coefficients. This is diagrammed in Figure 2.1.

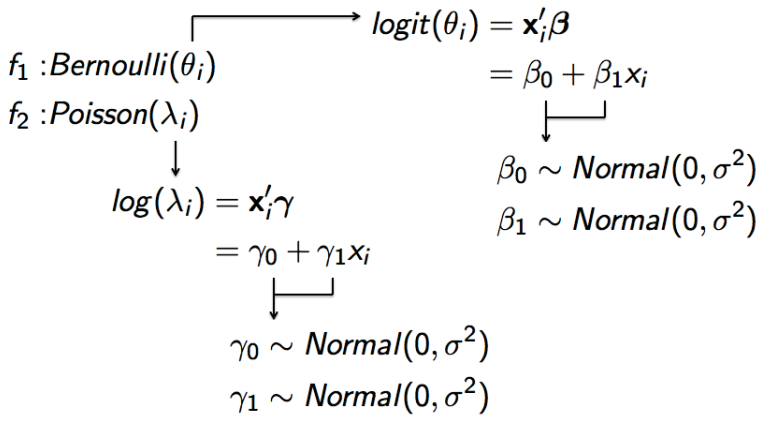

Figure 2.1: Diffuse Bayesian Model Structure.

For this example we take $\sigma^2 = 25^2$, which is very large relative to the true values of the regression parameters from which we are generating data, as described in Appendix A. We address more realistic ways to specify diffuse priors when discussing Bayesian sample size simulations in Chapter Four.

2.2.1 Example

Consider an example set of outcomes from the hypothetical sleep disorder study. We use the method described in Appendix A to generate data for a sample of size $n = 100$. Of the $n = 100$ responses, 40 positive responses made up the effective sample size for estimation on the Poisson part of the model. Histograms of the subjects' weights and outcomes are shown in Figure 2.2.

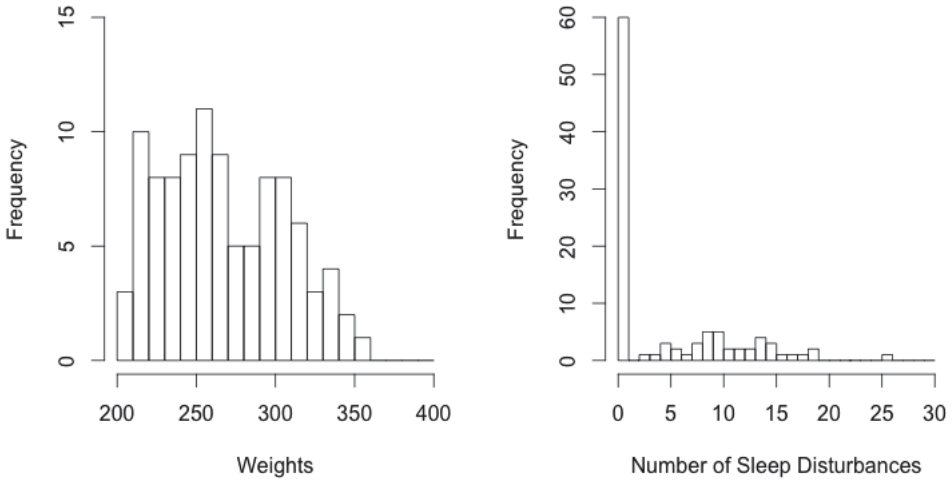

Figure 2.2: Histogram of Generated Weights (left) and Study Outcomes (right) for a Sample of Size $n = 100$.

The diffuse Bayesian Bernoulli-Poisson hurdle model can be fit to these outcomes using Markov chain Monte Carlo (MCMC) methods implemented, for example, in OpenBUGS. To do so, we follow the method used by Congdon (2005) for a sample $\mathbf{y} = (y_1, ..., y_n)$:

(1) Arrange the outcomes so that zero counts $(i = 1, ..., n_1)$ are at the beginning and positive counts $(i = n_1 + 1, ..., n)$ are at the end.

(2) For $i = 1, ..., n$ the logistic regression models the probability of crossing the hurdle ($y_i > 0$) or not ($y_i = 0$).

(3) For $i = n_1 + 1, ..., n$ the truncated Poisson regression models the number of events occurred, given a hurdle cross.

Note that OpenBUGS was chosen over WinBUGS to implement this model because of its ability to easily handle truncated distributions. Alternatively, the model can also be fit using the zeros trick in WinBUGS/OpenBUGS to define the non-standard likelihood, as done by Neelon et al. (2010), with comparable results.

The resulting posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown in Figure 2.3. Table 2.1 shows the posterior means and 95% credible sets for each of the parameters, as well as for θ and λ at certain weights. For reference Table 2.1 also includes the maximum likelihood estimates (MLEs) of these parameters and their true values. The MLEs were calculated using the previously discussed `hurdle` function (Zeileis et al., 2008) in the **pscl** package (Jackman, 2014).

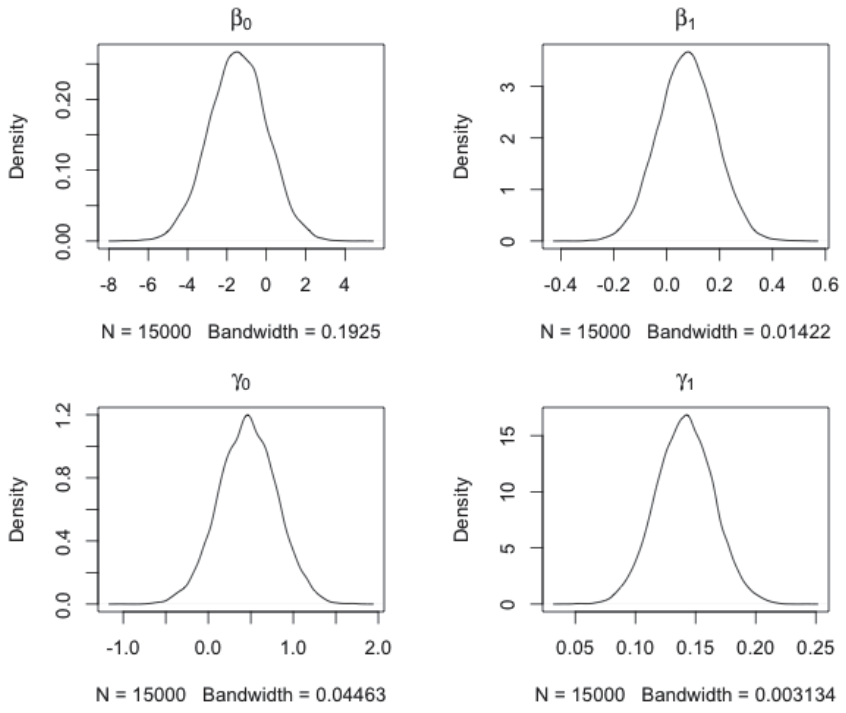

Figure 2.3: Posterior Density Plots for the Diffuse Model.

Table 2.1: Posterior Means and 95% Credible Sets for the Diffuse Model

Parameter	Truth	MLE	Mean	2.5%	97.5%	Width
β_0	-4.18	-1.40	-1.43	-4.31	1.43	5.75
β_1	0.28	0.07	0.08	-0.14	0.29	0.42
$\theta \vert x_i = 200$	0.20	0.34	0.34	0.18	0.54	0.36
$\theta \vert x_i = 300$	0.50	0.43	0.43	0.30	0.56	0.25
$\theta \vert x_i = 400$	0.80	0.52	0.52	0.20	0.82	0.62
γ_0	0.54	0.47	0.47	-0.20	1.13	1.33
γ_1	0.13	0.14	0.14	0.09	0.19	0.09
$\lambda \vert x_i = 200$	6.56	6.55	6.57	5.28	8.09	2.82
$\lambda \vert x_i = 300$	12.84	13.26	13.24	11.92	14.62	2.70
$\lambda \vert x_i = 400$	25.15	26.84	27.07	19.99	35.77	15.78

Note that the posterior means line up well with their frequentist counterpart MLEs. Also, the true values of the parameters are all contained in their respective 95% credible sets. Interval widths are difficult to interpret on the slope and intercept level, but the select values of θ and λ appear to all have reasonable interval widths, especially given this diffuse prior structure, with the exception of λ at 400 lbs. This is due in part to the inverse-link transformation and, as we shall see in the following simulation, the relatively small number of observations for the heaviest weights in this example.

2.2.2 Operating Characteristics

In order for Bayesian methods to be accepted as an alternative to their frequentist counterparts, it is often necessary to assess their operating characteristics. This is, for example, recommended in the U.S. Food and Drug Administration's guidance on the use of Bayesian methods in medical device trials (FDA, 2010). To do so, we consider a small simulation study with 100 replications for this model. Following the data generation process described in Appendix A, we generate 100 sets of subjects' weights and corresponding outcomes for sample sizes $n = 50$ and 100. For each data set, we fit the diffuse model using the same specifications as described in Section 2.2. The posterior means and 95% credible sets for $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, as well as some select values of θ and λ at various weights, were recorded for each iteration of the simulation.

Marginal results for these simulations are summarized in Tables 2.2 and 2.3, which show the true value of each parameter, the median of the 100 posterior means and 95% credible intervals, the median width of the intervals, and their coverage. For sample size $n = 50$, the lowest coverage is 0.90, and for sample size $n = 100$, the lowest coverage is 0.92. But, in both cases, the majority of the coverages are around 0.95 or higher.

Table 2.2: Simulation Results for the Diffuse Model with $n = 50$

Parameter	True	2.5%	Median	97.5%	Width	Coverage
β_0	-4.18	-8.72	-4.24	-0.12	8.92	0.95
β_0	0.28	-0.02	0.28	0.60	0.65	0.94
$\theta\|x_i = 200$	0.20	0.05	0.21	0.45	0.40	0.92
$\theta\|x_i = 300$	0.50	0.34	0.51	0.69	0.34	0.90
$\theta\|x_i = 400$	0.80	0.35	0.76	0.97	0.61	0.94
γ_0	0.54	-0.38	0.53	1.47	1.87	0.98
γ_1	0.13	0.07	0.13	0.19	0.13	0.97
$\lambda\|x_i = 200$	6.56	4.66	6.57	9.06	4.19	0.96
$\lambda\|x_i = 300$	12.84	11.16	12.68	14.32	3.20	0.98
$\lambda\|x_i = 400$	25.15	17.38	25.10	34.75	17.52	0.97

Table 2.3: Simulation Results for the Diffuse Model with $n = 100$

Parameter	True	2.5%	Median	97.5%	Width	Coverage
β_0	-4.18	-7.73	-4.56	-1.54	6.11	0.94
β_1	0.28	0.10	0.30	0.53	0.44	0.95
$\theta\|x_i = 200$	0.20	0.08	0.20	0.37	0.28	0.92
$\theta\|x_i = 300$	0.50	0.38	0.50	0.63	0.25	0.96
$\theta\|x_i = 400$	0.80	0.56	0.82	0.96	0.39	0.97
γ_0	0.54	-0.07	0.58	1.23	1.30	0.96
γ_1	0.13	0.09	0.13	0.18	0.09	0.96
$\lambda\|x_i = 200$	6.56	5.33	6.75	8.38	3.01	0.97
$\lambda\|x_i = 300$	12.84	11.83	12.97	14.15	2.32	0.97
$\lambda\|x_i = 400$	25.15	19.43	25.24	31.72	11.99	0.96

These results are also shown as box plots in Figures 2.4 and 2.5. In each box plot, the horizontal line represents the true value of the specified parameters. The center point in each box plot represents the median of the 100 posterior means. The upper and lower limits represent the average of the upper and lower bounds of the 100 posterior 95% credible sets. The grey bars represent ± 1 simulation standard

deviation above and below each respective point on the box plot, which occasionally overlap, resulting in a darker shade of grey.

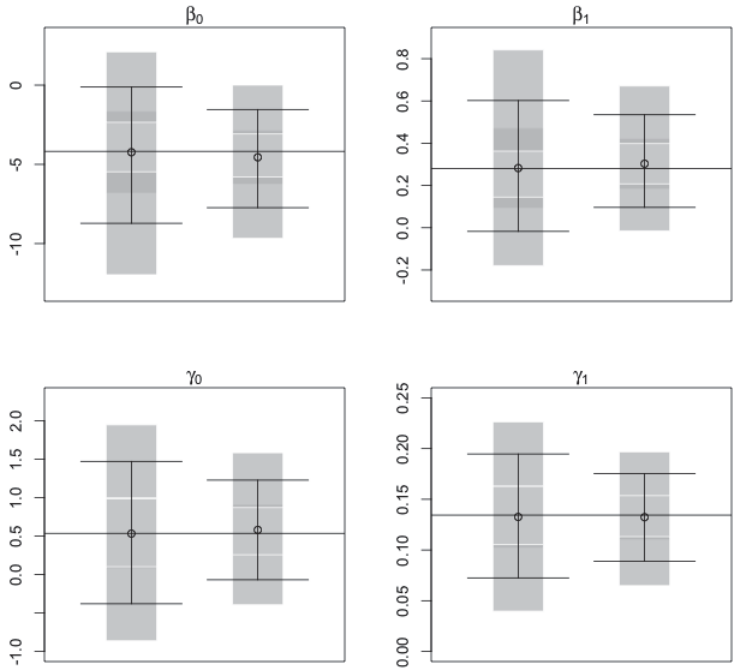

Figure 2.4: Simulation Results for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ for the Diffuse Model with $n = 50$ (left) and $n = 100$ (right).

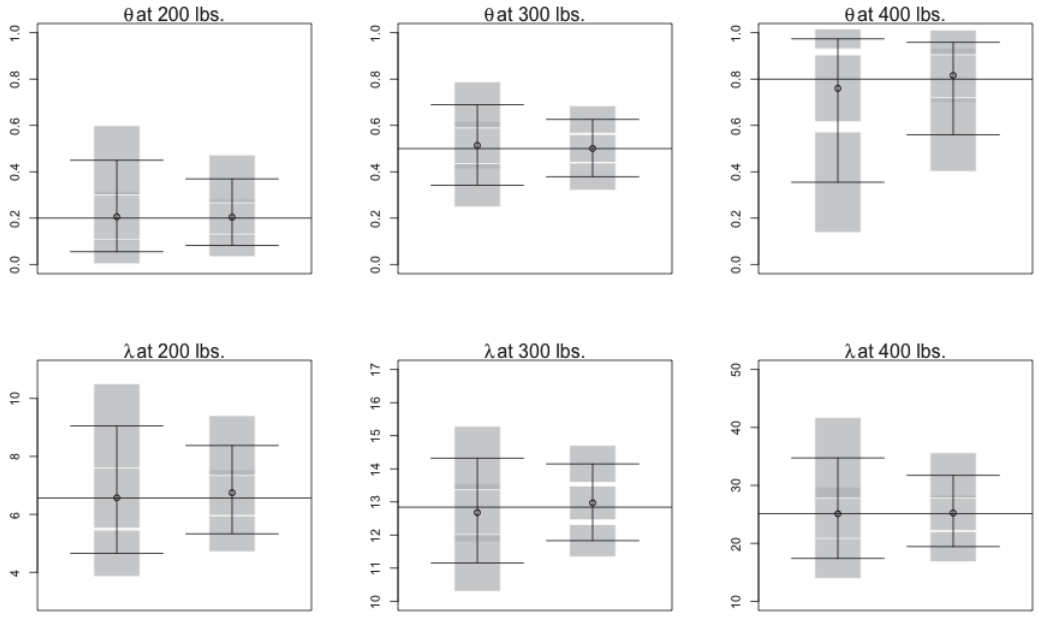

Figure 2.5: Simulation Results for Select Values of θ and λ for the Diffuse Model with $n = 50$ (left) and $n = 100$ (right).

The results show that on average the posterior means for all the parameters tend towards the truth with very little bias. In each plot, the left box plot shows the results for $n = 50$ and the right box plot shows the results for $n = 100$. There is a clear difference in average interval widths as sample size increases. Particularly, the intervals for θ and λ at extreme weight values (i.e. 400 lbs.) appear to get slightly more reasonable with a sample size increase, a trend that will likely continue to hold for any larger sample sizes considered.

2.3 Informative Prior Structure

Methods of prior construction for Bayesian hurdle models have not been covered in detail and are worth exploring. Recall that the Bernoulli-Poisson hurdle model we consider here consists of a logistic regression model for the hurdle part and a truncated Poisson regression for the count part. There are a number of prior elicitation methods for generalized linear models widely discussed in the literature that can potentially be adapted for this model of interest, including power priors, commensurate priors, informative g-priors, and conditional means priors. In this section, we focus on the latter. We discuss how prior information can be elicited from experts and incorporated into a prior structure for the Bernoulli-Poisson hurdle model using conditional means priors.

2.3.1 Conditional Means Priors

Bedrick et al. (1996) developed the idea of conditional means priors, also known as BCJ priors, which are commonly used on regression parameters in generalized linear models. It is difficult to elicit prior information on regression coefficients because it is often hard to interpret slopes and intercepts, making it challenging to elicit the appropriate information from experts. The conditional means prior approach solves this problem by instead asking experts for average response values at various covariate configurations, which is much more operational. This information is then used to induce priors on the regression coefficients.

Consider the general form of a generalized linear model. Let y_i have a density $f(y_i|\mu_i, \boldsymbol{\phi})$, with $\mu_i = h^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$, where $\mathbf{x}_i$ is a $p \times 1$ vector of covariates and $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. Here, h^{-1} represents the inverse link function and $\boldsymbol{\phi}$ represents any nuisance parameters. The goal is to induce a prior distribution on $\boldsymbol{\beta}$, based on priors elicited on μ_i, at various configurations of $\mathbf{x}_i$.

Suppose we have K covariate configurations, as specified by a $K \times p$ design matrix

$$\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1 \\ \vdots \\ \tilde{\mathbf{x}}_K \end{bmatrix},$$

where the rows represent K distinct values of $\mathbf{x}_i$. Note that in order for $\tilde{\mathbf{X}}$ to be nonsingular, K must be equal to p. For each covariate configuration, $\tilde{\mathbf{x}}_i$, a prior is elicited for the corresponding mean response value, $\tilde{\mu}_i = h^{-1}(\tilde{\mathbf{x}}_i\boldsymbol{\beta})$. This informative prior is elicited from an expert and denoted as F_i. Note that all the F_i's are assumed to be independent, meaning the covariate configurations $\tilde{\mathbf{x}}_\mathbf{i}$ are assumed to be sufficiently distinct from each other. The prior structure for the conditional means approach can thus be derived as follows. Observe that

$$\tilde{\boldsymbol{\mu}} = \begin{bmatrix} \tilde{\mu}_1 \\ \vdots \\ \tilde{\mu}_K \end{bmatrix} = \begin{bmatrix} h^{-1}(\tilde{\mathbf{x}}_1'\boldsymbol{\beta}) \\ \vdots \\ h^{-1}(\tilde{\mathbf{x}}_K'\boldsymbol{\beta}) \end{bmatrix} \equiv h^{-1}(\tilde{\mathbf{X}}\boldsymbol{\beta}).$$

Since $\tilde{\mathbf{X}}$ was designed to be invertible, $\boldsymbol{\beta} = \tilde{\mathbf{X}}^{-1}h(\tilde{\boldsymbol{\mu}})$. Thus, the induced priors on $\boldsymbol{\beta}$ are defined as

$$\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_K \end{bmatrix} \sim \tilde{\mathbf{X}}^{-1}h\left(\begin{bmatrix} F_1 \\ \vdots \\ F_K \end{bmatrix}\right),$$

which generally has no closed form.

2.3.2 Conditional Means Priors for the Bernoulli-Poisson Hurdle Model

For the Bernoulli-Poisson hurdle model, we can use conditional means priors to incorporate expert opinion in both parts of the model, leading to improved inference.

We describe this process in terms of the sleep disorder scenario introduced in Section 2.1.4.

First, consider eliciting a conditional means prior for the logistic regression used to model whether or not a subject suffers from a sleep disorder,

$$\begin{aligned} \text{logit}(\theta_i) &= \mathbf{x}_i'\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 x_i, \end{aligned}$$

where x_i represents the single covariate, weight. Since there are two regression coefficients, the design matrix $\tilde{\mathbf{X}}$ will have two covariate configurations. Bedrick et al. (1996) explain in detail how these configurations can be chosen. For our purposes, we choose the two weights to be the endpoints of the weight range, 200 and 400 lbs. Here we are assuming that this weight difference renders corresponding responses sufficiently distinct.

We use the mode-percentile method of elicitation to translate information relayed from an expert into informative prior distributions on θ at 200 and 400 lbs. At both weights, the expert is asked "What, do you think, is the probability of suffering from a sleep disorder for a x_i lb. man?" In response, they are prompted to relay a most likely probability (mode) along with an upper or lower bound (percentile) to represent uncertainty. Numerical methods are then used to translate this information into parameters for an appropriate prior distribution at each weight. Note that other methods of elicitation can be used as well.

Suppose Table 2.4 shows the resulting information collected from the expert. That is, the expert believes that for a 200 lb. subject, the probability of suffering from a sleep disorder is most likely $\theta = 0.20$ and is not likely to be more than $\theta = 0.30$. For a 400 lb. subject, he believes the probability is most likely $\theta = 0.80$ and is not likely to be less than $\theta = 0.70$. The upper and lower bounds are interpreted as 95th and 5th percentiles respectively. Note that the choice of eliciting an upper or lower bound is determined at the researcher's discretion based on how to best fit an appropriate distribution.

Table 2.4: Expert Elicited Information on θ

Weight	Mode	Percentile
200	0.20	0.30 (upper)
400	0.80	0.70 (lower)

This mode-percentile information can be translated into beta priors. This can be done, for example, with the `elicitor` function in the **glmcmp** package (Kahle et al., 2014) in R. The resulting priors at $x_i = 200$ and 400 are

$$\theta|_{x_i=200} = \theta_{200} \equiv \text{logit}^{-1}(\beta_0 + \beta_1(200)) \sim Beta(12.82, 48.28)$$

and

$$\theta|_{x_i=400} = \theta_{400} \equiv \text{logit}^{-1}(\beta_0 + \beta_1(400)) \sim Beta(48.28, 12.82).$$

Density plots for these priors are shown in Figure 2.6. In practice, these priors would be shown to the expert to verify whether or not they properly reflect their beliefs and appropriate changes would be made until acceptance. For our purposes, we assume that the expert was consulted and agreed with the appropriateness of these prior choices.

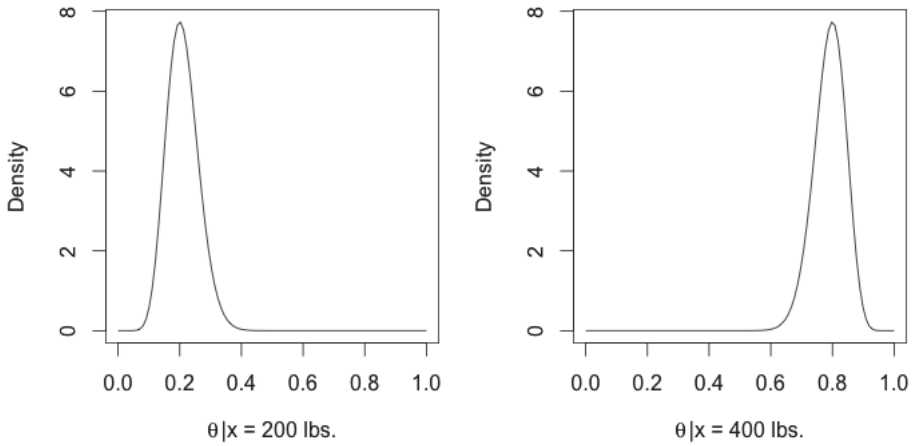

Figure 2.6: Informative Beta Priors for θ at 200 and 400 lbs.

The conditional means priors for β_0 and β_1 are the resulting induced priors

$$\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \sim \tilde{\mathbf{X}}^{-1}\text{logit} \begin{bmatrix} Beta(12.82, 48.28) \\ Beta(48.28, 12.82) \end{bmatrix},$$

where

$$\tilde{\mathbf{X}} = \begin{bmatrix} 1 & 200 \\ 1 & 400 \end{bmatrix}.$$

These priors have no closed form but can easily be simulated and the resulting densities are shown in Figure 2.7.

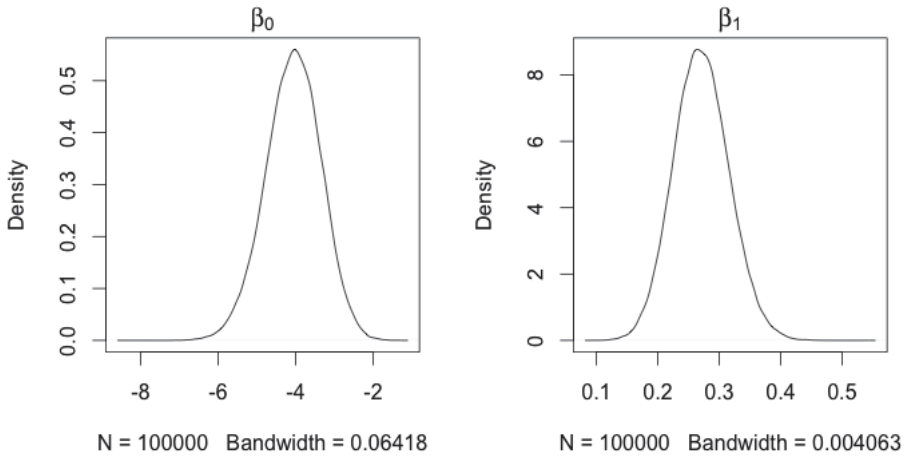

Figure 2.7: Simulated Density Plots for the Induced Priors on β_0 and β_1.

Following this, an analogous process can be used to elicit conditional means priors for the parameters of the Poisson regression used to model the number of unwanted sleep interruptions, given the subject suffers from a sleep disorder,

$$\begin{aligned} \log(\lambda_i) &= \mathbf{x}_i'\boldsymbol{\gamma} \\ &= \gamma_0 + \gamma_1 x_i, \end{aligned}$$

where x_i again represents the single covariate, weight. The same two covariate configurations of $x_i = 200$ and 400 lbs. are used, but this time the expert is asked, "For a x_i lb. subject suffering from a sleep disorder, how many unwanted sleep interruptions do you expect him to experience in a single night?"

Using mode-percentile elicitation as before, Table 2.5 shows the information collected from the expert. The expert believes a 200 lb. subject most likely has $\lambda = 6$ interruptions with a 95th percentile (upper bound) of $\lambda = 8$ and a 400 lb. subject most likely has $\lambda = 25$ interruptions with a 95th percentile of $\lambda = 33$ interruptions.

Table 2.5: Expert Elicited Information on λ

Weight	Mode	Percentile
200	6	8 (upper)
400	25	33 (lower)

We use the `elicitor` function in **glmcmp** (Kahle et al., 2014) to fit the following gamma priors for $x_i = 200$ and 400,

$$\begin{aligned}\lambda|_{x_i=200} = \lambda_{200} \equiv \exp(\gamma_0 + \gamma_1(200)) &\sim Gamma(34.65, 0.18)\\ \lambda|_{x_i=400} = \lambda_{400} \equiv \exp(\gamma_0 + \gamma_1(400)) &\sim Gamma(37.11, 0.69).\end{aligned}$$

Density plots for these priors are shown in Figure 2.8. Again, we assume that the expert agreed with these chosen priors. The conditional means priors for γ_0 and γ_1 are the resulting induced priors

$$\begin{bmatrix}\gamma_0\\ \gamma_1\end{bmatrix} \sim \tilde{\mathbf{X}}^{-1}\log\begin{bmatrix}Gamma(34.65, 0.18)\\ Gamma(37.11, 0.69)\end{bmatrix},$$

which again have no closed form but simulated density plots are shown in Figure 2.9.

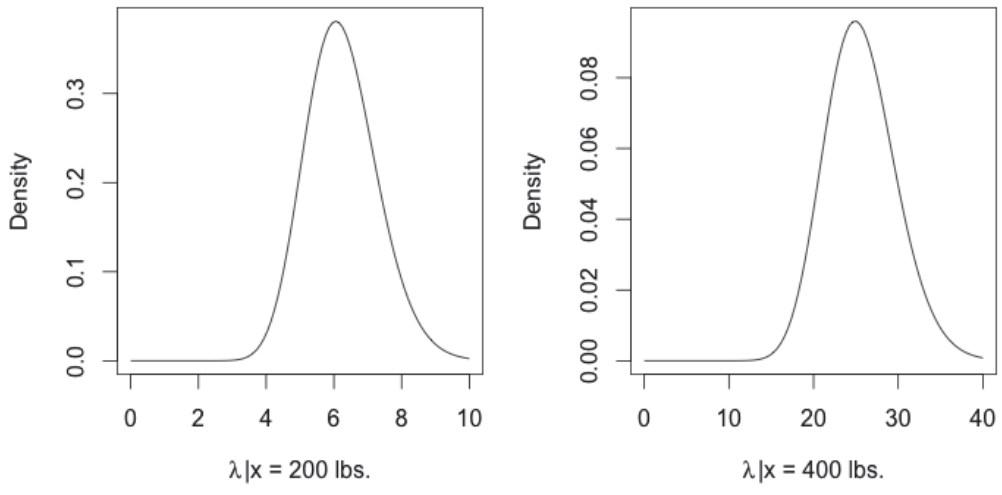

Figure 2.8: Informative Gamma Priors for λ at 200 and 400 lbs.

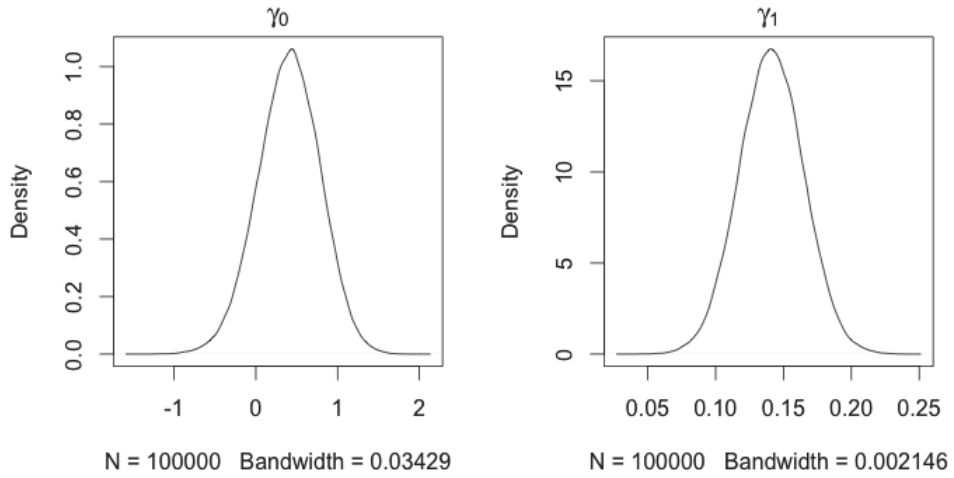

Figure 2.9: Simulated Density Plots for the Induced Priors on γ_0 and γ_1.

The Bayesian Bernoulli-Poisson hurdle model with these conditional means priors is diagrammed in Figure 2.10. Note that, as the diagram suggests, the information elicited on the hurdle part of the model is considered completely separate from the information elicited on the count part of the model. This can also be seen in the joint prior contours for all pairings of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ shown in Figure 2.11. The contours for (β_0, β_1) and (γ_0, γ_1) reflect the covariance that is known to exist between them. The lack of a connection between elicitation for the hurdle and count parts of the models is further reflected in the circular contours for (β_0, γ_0), (β_0, γ_1), (β_1, γ_0), and (β_1, γ_1).

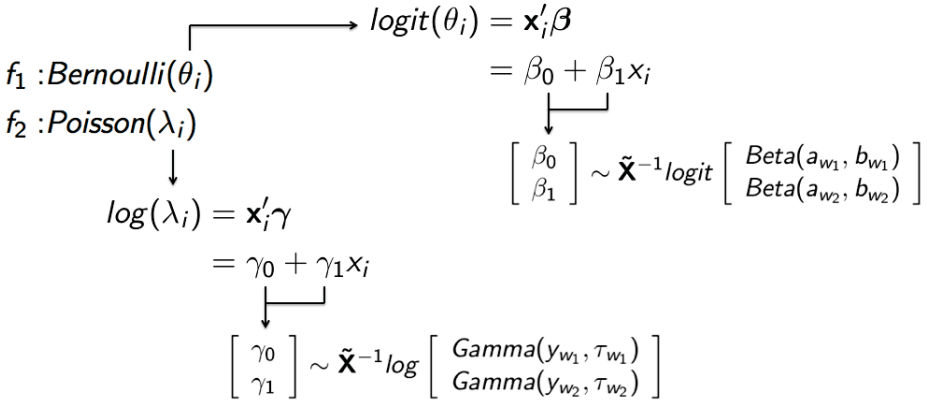

Figure 2.10: Bayesian Model with Conditional Means Priors.

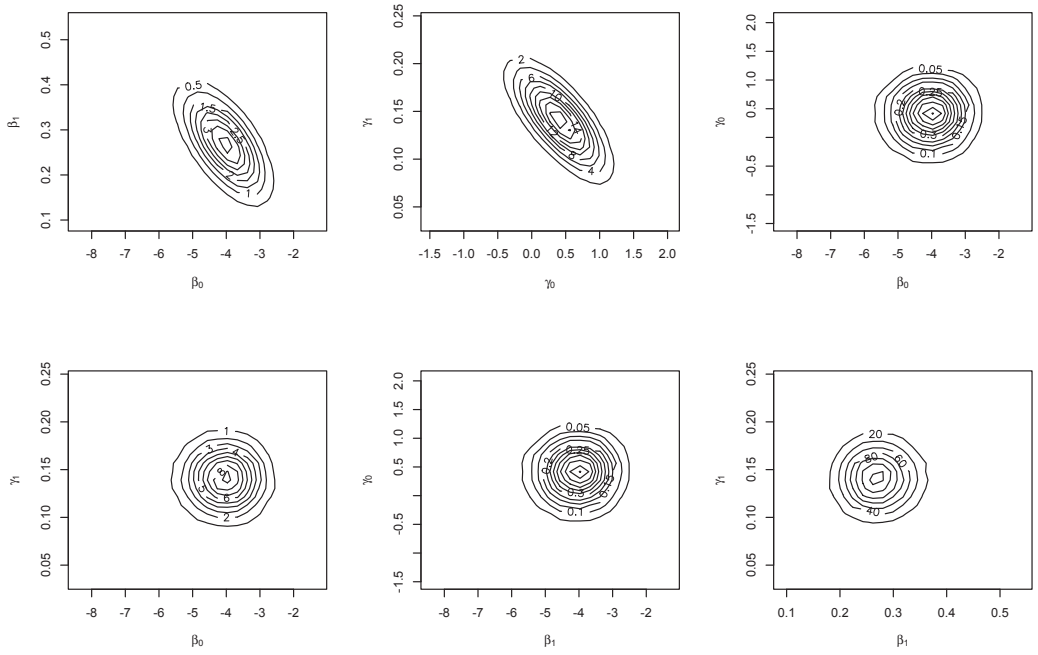

Figure 2.11: Simulated Joint Prior Contours for the Regression Parameters.

2.3.3 Example

Consider the same set of generated sleep study outcomes shown in Figure 2.2. Given the described expert elicited information, we fit the Bayesian Bernoulli-Poisson hurdle model with independent conditional means priors to this data using Markov chain Monte Carlo (MCMC) methods in OpenBUGS following the method specified by Congdon (2005). We ran a single chain with 16,000 iterations, including a burn-in of 1,000 iterations, with initial values set at the elicited modal values: $\theta_{200} = 0.2$, $\theta_{400} = 0.8$, $\lambda_{200} = 6$, and $\lambda_{400} = 25$. To reduce autocorrelation issues, we set thinning at 10. As before, we investigated convergence issues for this model using multiple chains and a variety of starting values. We found little difference in the posteriors and the Gelman-Rubin plots all converged to one.

The resulting posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown by the solid curves in Figure 2.12. For reference, the dashed curves show the posterior densities under the previously fit diffuse Bayesian model. Additionally, Table 2.6 shows the posterior means and 95% credible sets for each of the parameters for

this informative model. Note the improved agreement between posterior means and true parameter values compared to the diffuse model, as expected with the use of informative priors. Furthermore, the the true values of the parameters are all still contained in their respective 95% credible sets.

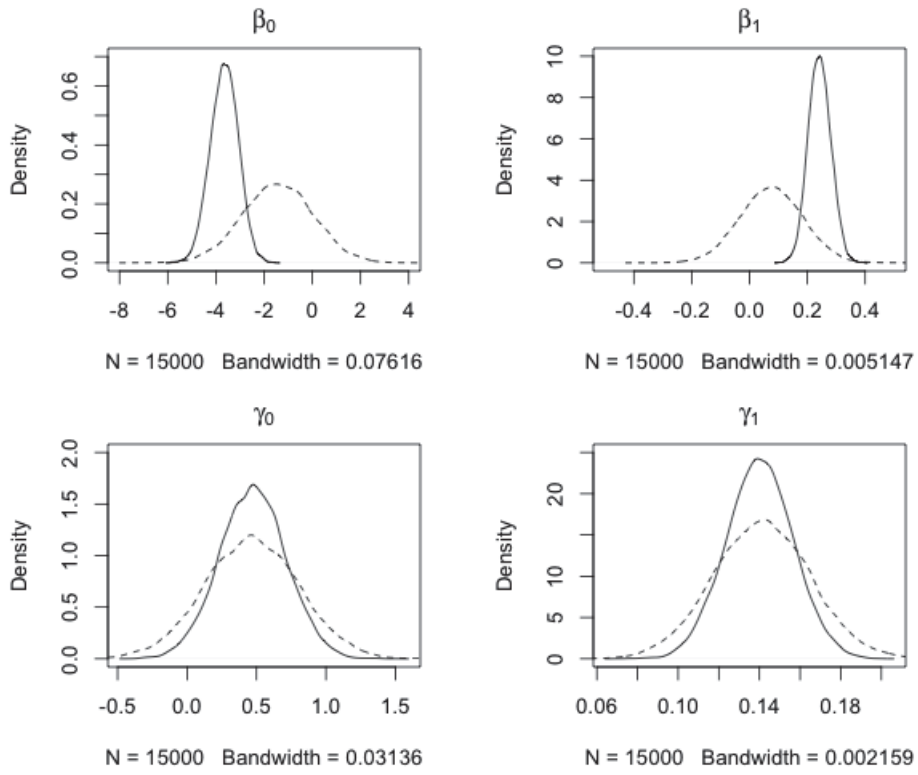

Figure 2.12: Posterior Density Plots for the Informative Model (solid) and the Diffuse Model (dashed).

Table 2.6: Posterior Means and 95% Credible Sets for the Informative Model

Parameter	Truth	Mean	2.5%	97.5%	Width
β_0	-4.18	-3.65	-4.80	-2.54	2.25
β_1	0.28	0.24	0.17	0.32	0.15
$\theta \vert x_i = 200$	0.20	0.23	0.16	0.31	0.15
$\theta \vert x_i = 300$	0.50	0.50	0.42	0.57	0.15
$\theta \vert x_i = 400$	0.80	0.76	0.66	0.85	0.19
γ_0	0.54	0.47	0.00	0.94	0.94
γ_1	0.13	0.14	0.11	0.17	0.06
$\lambda \vert x_i = 200$	6.56	6.54	5.54	7.65	2.10
$\lambda \vert x_i = 300$	12.84	13.15	12.00	14.35	2.35
$\lambda \vert x_i = 400$	25.15	26.61	21.50	32.35	10.85

As another illustration of using twin CMP structures for our hurdle model, suppose we have substantially different precision in elicitation for the two components. Recall that the independent conditional means prior is not structured to account for any relationship between the hurdle and count parts of the model. Thus, having more or less information with respect to one part of the model should have no effect on the other part. To illustrate this concept, suppose that, in contrast to what was previously described, researchers are unable to get such precise information on the hurdle part of the model. That is, for whatever reason, we do not know as much regarding the probability of suffering from a sleep disorder for these subjects compared to our previous illustration in Section 2.3.2. Specifically, the elicited percentile information is widened to account for increased uncertainty. The expert still believes that, for a 200 lb. subject, θ is expected to be 0.2, but sets the upper bound (or 95th percentile) at $\theta = 0.4$. Similarly, he or she still believes that, for a 400 lb. subject, θ is expected to be 0.8, but sets the lower bound (or 5th percentile) at $\theta = 0.6$. Again, using `glmcmp` in R for this mode-percentile elicitation results in the following beta priors,

$$\theta|_{x_i=200} = \theta_{200} \equiv \text{logit}^{-1}(\beta_0 + \beta_1(200)) \sim Beta(4.46, 14.84)$$

and

$$\theta|_{x_i=400} = \theta_{400} \equiv \text{logit}^{-1}(\beta_0 + \beta_1(400)) \sim Beta(14.84, 4.46),$$

which are then transformed to induced priors on $\boldsymbol{\beta} = (\beta_0, \beta_1)$.

With this slight modification in prior information, the model is once again fit to the given data set using Markov chain Monte Carlo (MCMC) methods in OpenBUGS under the same specifications. The resulting posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown by the solid curves in Figure 2.13. For reference, the posteriors from the previously fit informative model are shown by the dashed curves. Table 2.7 shows the posterior means and 95% credible sets for each of the parameters under the modified informative model.

For the parameters on the hurdle part of the model, the true values are still contained in their respective 95% credible sets. However, the intervals are wider,

reflecting the fact that our information is less precise than before. On the other hand, the posteriors means and 95% credible sets for the parameters on the Poisson count part of the model have not changed significantly. Any slight differences can be attributed to Monte Carlo error.

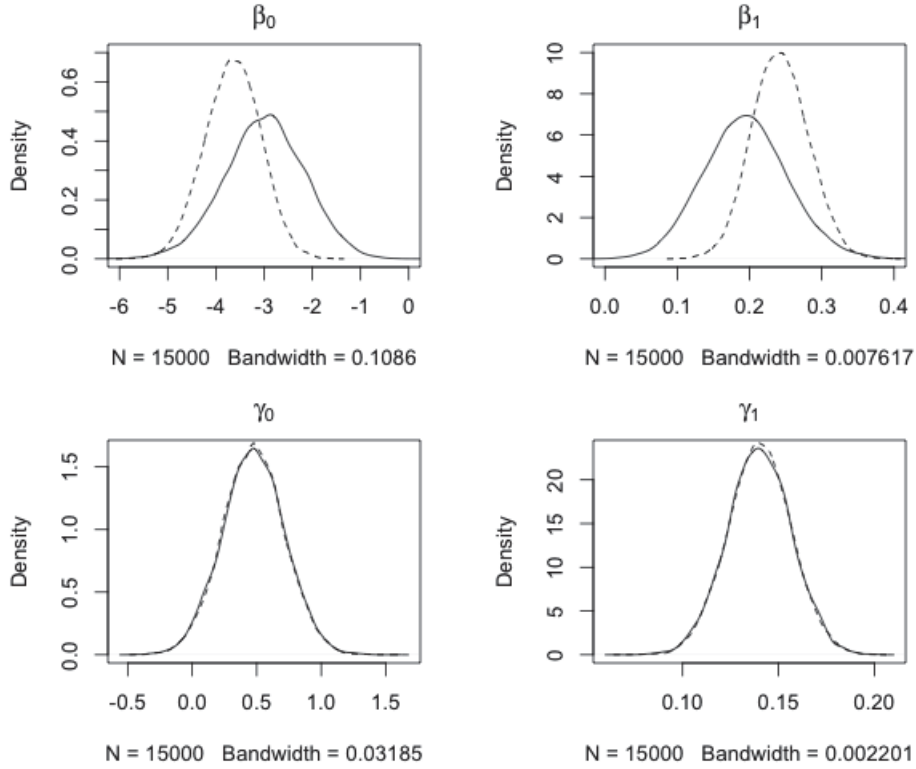

Figure 2.13: Posterior Density Plots for the Modified Informative Model (solid) and the Previous Informative Model (dashed).

Table 2.7: Posterior Means and 95% Credible Sets for the Informative Model

Parameter	Truth	Mean	2.5%	97.5%	Width
β_0	-4.18	-2.99	-4.65	-1.40	3.25
β_1	0.28	0.19	0.08	0.31	0.23
$\theta \vert x_i = 200$	0.20	0.26	0.16	0.38	0.22
$\theta \vert x_i = 300$	0.50	0.48	0.39	0.57	0.18
$\theta \vert x_i = 400$	0.80	0.70	0.53	0.85	0.31
γ_0	0.54	0.48	0.00	0.95	0.95
γ_1	0.13	0.14	0.11	0.17	0.07
$\lambda \vert x_i = 200$	6.56	6.55	5.52	7.66	2.13
$\lambda \vert x_i = 300$	12.84	13.15	11.99	14.34	2.35
$\lambda \vert x_i = 400$	25.15	26.59	21.40	32.37	10.97

2.4 Summary

In this chapter we introduced the Bernoulli-Poisson hurdle model along with a hypothetical sleep disorder application. The Bayesian approach to this model with diffuse priors has been well covered in the literature. For completeness, we considered a diffuse model with respect to data generated to reflect the sleep disorder scenario. Operating characteristics showed that the performance of the Bayesian model with an absence of prior information was comparable to frequentist maximum likelihood methods.

We then proposed an informative prior structure for the Bernoulli-Poisson hurdle model, with both parts dependent on covariates, that consisted of conditional means priors placed separately, or independently, on both parts. We discussed how information could be elicited from experts in the context of the sleep disorder example. This informative Bayesian model was fit to an example generated data set, and the resulting posteriors reflected the influence of the added information. Then we relaxed the specificity of the information provided on the hurdle part of the model and re-fit the model to the same data. Posterior results showed wider credible sets for parameters on the hurdle part and no changes on the count part, illustrating and emphasizing the independence of the approach.

When the two parts of the Bernoulli-Poison hurdle model share the same covariates, as in our sleep disorder application, a relationship may exist between parameters for which prior elicitation is needed. In this chapter, the proposed independent conditional means prior approach essentially ignores this relationship. This is an important aspect of the elicited information and we examine methods of incorporating this dependency directly into the prior structure in Chapter 3.

CHAPTER THREE

A Coupled Prior Structure for Bayesian Hurdle Models

3.1 Motivation: Independent vs. Coupled Approaches

In Chapter Two, we introduced a Bayesian Bernoulli-Poisson hurdle model consisting of a logistic regression in the hurdle part and a Poisson regression in the count part. We proposed an informative prior structure consisting of two conditional means priors independently constructed for the hurdle and count parts of the model. We described how information could be elicited from experts for this purpose, using a hypothetical sleep disorder study for context. In that elicitation process, the expert first relays their beliefs regarding the probability that a subject suffers from a sleep disorder for the hurdle part of the model. The same expert then relays their beliefs regarding the number of sleep disturbances expected, given the subject suffers from a sleep disorder, for the count part of the model.

Constructing hurdle and count prior components independently ignores any available joint knowledge. Plausibly, higher sleep disturbance probabilities may be associated with higher sleep disturbance rates. In this chapter, we propose a coupled conditional means prior structure that better reflects how the expert's opinions on the two parts of the Bernoulli-Poisson hurdle model are entangled in the case of shared covariates. Continuing with the hypothetical sleep disorder example, we explore the potential gains, as well as the shortcomings, of such an approach.

3.2 Relationship Between the Hurdle and the Count

In the special case where the covariates in both parts of the hurdle model coincide, Ridout et al. (1998) state that the general hurdle model assumption that the linear predictor for θ_i is unrelated to the linear predictor for λ_i is restrictive (and unrealistic), proposing a model analogous to the $ZIP(\tau)$ model (Lambert, 1992)

where both parts of the model are related by a scalar value τ. Applied to the Bernoulli-Poisson hurdle model (2.5) that we have been discussing, it would follow that

$$\text{logit}(\theta_i) = \tau \mathbf{x}_i' \boldsymbol{\beta}$$

and

$$\log(\lambda_i) = \mathbf{x}_i' \boldsymbol{\beta}.$$

Suppose we generalize this further and propose instead that the relationship between θ_i and λ_i can be modeled by some *coupling function*, $f(\theta_i) = \lambda_i$, such that

$$\text{logit}(\theta_i) = \mathbf{x}_i' \boldsymbol{\beta}$$

and

$$\log(\lambda_i) = \log \circ f \circ \text{logit}^{-1}(\mathbf{x}_i' \boldsymbol{\beta}), \tag{3.1}$$

which acts as a reparameterization of the original $\log(\lambda_i) = \mathbf{z}_i' \boldsymbol{\gamma}$.

To fit this model, we must specify the coupling function, f, using expert opinion. The latter is unlikely to be in explicit functional form. We must elicit information with which to construct a function which represents the expert's beliefs. To this end, we suggest an approach that begins with the use of information from the independent CMP elicitation we have already described. To illustrate how this might proceed, we return to the hypothetical sleep disturbance study, wherein the hurdle and count parts of the model share a covariate, the subject's weight.

From the independent elicitation of conditional means priors on both parts of the model, we have prior information on the probability of suffering from a sleep disorder and the expected number of sleep disturbances suffered by an afflicted subject at two specific weights, 200 and 400 lbs. This information is summarized graphically in Figure 3.1. The points represent the modal values elicited from the expert. The vertical line running through each of the points represent a central 90% interval (5th percentile lower bound, 95th percentile upper bound) for the parameter at that weight. Further, the solid portion of the vertical line represents the elicited upper or lower bound, and the dashed portion represents the calculated upper or lower bound,

based on the corresponding elicited beta or gamma prior. The hinges are at the endpoints of these central 90% intervals, representing bounds of uncertainty.

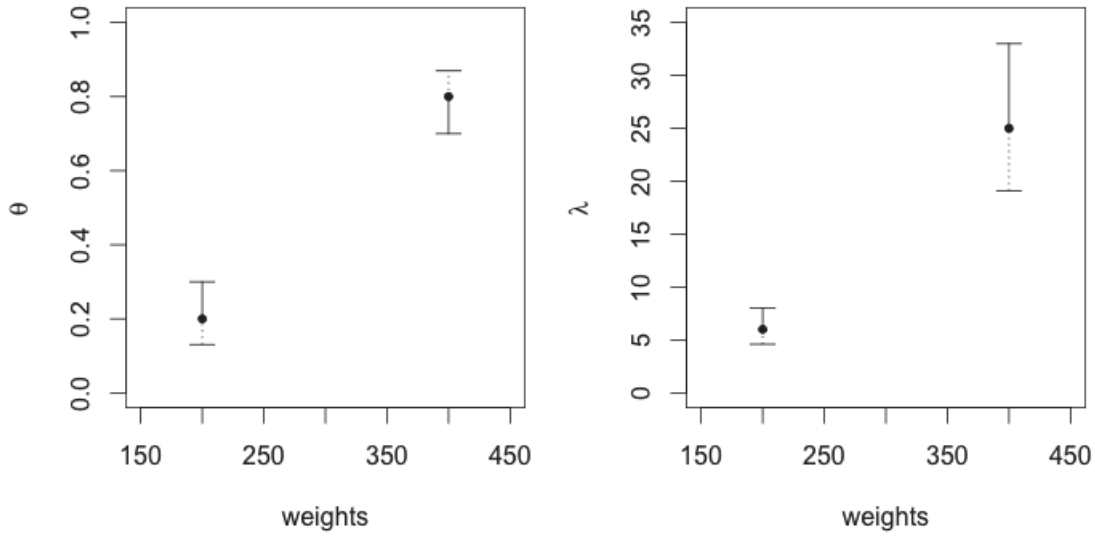

Figure 3.1: Prior Information on (a) the Probability of Suffering from a Sleep Disorder and (b) the Number of Sleep Disturbances Suffered by an Afflicted Subject.

Note that, in Figure 3.1, θ_i and λ_i are both plotted against the same weights. Suppose we collapse the weight covariate and instead plot θ_i versus λ_i directly. This allows us to better picture the relationship between the two, as shown in Figure 3.2. The box at the lower left corner summarizes information given for subjects weighing 200 lbs. and the box at the upper right corner for subjects weighing 400 lbs. The points within the boxes represent the pair of modal values (θ_i, λ_i) at each weight and the length and width of the boxes depict the now two-dimensional uncertainty, with length being the uncertainty with respect to λ_i and width the uncertainty with respect to θ_i.

To specify a coupling function, we need to determine the functional relationship between θ and λ for covariate values within the extremes. Thus, suppose that, in addition to the information at 200 and 400 lbs., we ask the expert to provide modal values for θ_i and λ_i at 200, 300, and 350 lbs. This additional information is shown by the red points in Figure 3.3, with the solid lines connecting them representing interpolated modes for the weights in between. Now, collapsing the shared weight covariate

gives us the information shown in Figure 3.4, where the functional relationship f is much clearer than before.

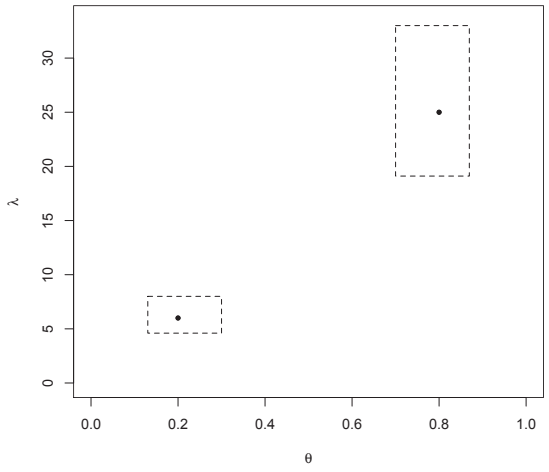

Figure 3.2: Combined Prior Information on θ_i and λ_i.

Note that it is, of course, possible to elicit full prior distributions for these additional weights by asking the expert for an appropriate upper or lower bound to correspond with each modal value, resulting in boxes around each of the three points representing 250, 300, and 400 lbs. in Figure 3.4. We explored doing this but discovered that the additional information, and incurring the costs of gathering it, was not necessary in formulating this prior. The reasons for this become clearer in the next section when we discuss our method for calibrating the elicitation.

The choice of the coupling function f is important. There are a number of functions that could model the relationship shown in Figure 3.4: linear, parabolic, cubic, exponential, etc., with some more appropriate than others. Fit is one concern; however, more importantly, we have found that the most practical coupling function is one that is 'separable' in terms of its parameters after the link function in the count part of the model is applied, as in (3.1). Further, we want it to have the same number of parameters as the generalized linear model that it is replacing.

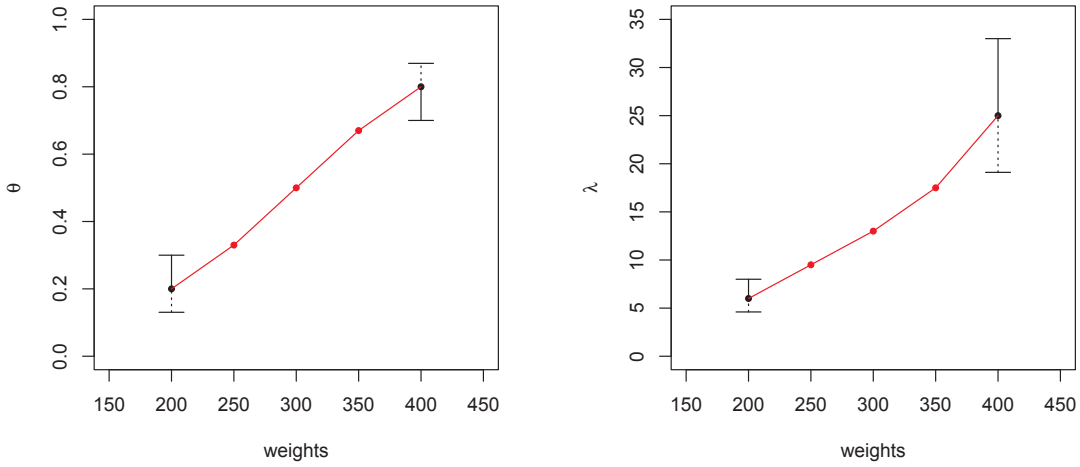

Figure 3.3: Prior Information on (a) the Probability of Suffering from a sleep disorder and (b) the number of Sleep Disturbances Suffered by an Afflicted Subject.

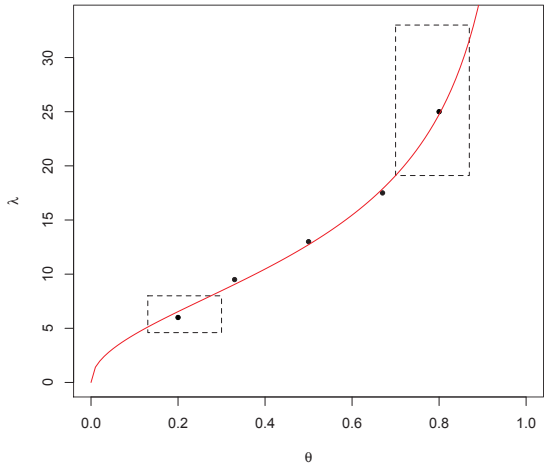

Figure 3.4: Combined Prior Information on θ_i and λ_i.

To illustrate, consider a coupling function of the form

$$\lambda_i = f(\theta_i) = u \exp(v \text{logit}(\theta_i)). \tag{3.2}$$

The red line in Figure 3.4 shows the least squares fit of this function to the modal values, which, calculated in R, is $u = 12.72$ and $v = 0.48$.

Note that the chosen coupling function has two parameters, u and v, corresponding to the two parameters, γ_0 and γ_1 in the original Poisson regression model.

Observe that when the log-link is applied,

$$\begin{aligned}\log(\lambda_i) &= \gamma_0 + \gamma_1 x_i \\ &= \log(u) + v\text{logit}(\theta_i) \\ &= \log(u) + v(\beta_0 + \beta_1 x_i) \\ &= [\log(u) + v\beta_0] + [v\beta_1]x_i \\ &= \gamma_0 + \gamma_1 x_i.\end{aligned}$$

The resulting equation is 'separable' in that the terms can be regrouped in a way that essentially reparameterizes γ_0 and γ_1 in terms of u and v, where

$$\gamma_0 = \log(u) + v\beta_0 \tag{3.3}$$

and

$$\gamma_1 = v\beta_1. \tag{3.4}$$

This proves to be useful for comparison and interpretation later on.

The convenient separability was made possible due to the choice for f to be exponential, which is the inverse of the log-link, as a function of the logit transform of θ_i, which we know to be linear. Then, by choosing a coupling function with the same number of parameters as the original regression, it was easy to reparameterize the original parameters (i.e. γ_0 and γ_1) as functions of the new parameters introduced in the coupling function (i.e. u and v).

This idea is generalizable and analogous choices of coupling functions should be possible for any chosen pair of generalized linear models in a given hurdle model with shared covariates. For example, suppose g is the link function for the hurdle part of the model, such that

$$g(\theta_i) = \mathbf{x}_i'\boldsymbol{\beta},$$

and h is the link function of the count part of the model, such that

$$h(\lambda_i) = \mathbf{x}_i'\boldsymbol{\gamma}.$$

Recall we want to choose the *coupling function* $f(\theta_i) = \lambda_i$ in order for

$$h(\lambda_i) = h \circ f(\theta_i) = h \circ f \circ g^{-1}(\mathbf{x}_i'\boldsymbol{\beta})$$

to reparameterize $h(\lambda_i) = \mathbf{x}_i'\boldsymbol{\gamma}$. Thus, generally, we should let $f(\theta_i) = h^{-1} \circ g(\theta_i)$, with the appropriate number of new parameters incorporated within it.

3.3 Coupled Prior Structures

We want to incorporate the elicited information regarding the relationship between θ_i and λ_i into a coupled prior structure for our Bernoulli-Poisson hurdle model. Given the elicited coupling function f, we propose two approaches to such a prior. The "fixed approach" fixes the parameters of the coupling function at values believed to correctly characterize the relationship. The "variable approach" relaxes this assumption, instead allowing these values to vary, giving the prior more flexibility, albeit at a price. Both approaches are rooted in conditional means priors and have advantages and disadvantages compared to the independent approach introduced in Chapter Two.

3.3.1 Fixed Approach

Consider the prior structure diagrammed in Figure 3.5. Suppose we fix the parameters of the coupling function at their least squares estimates, $u = 12.72$ and $v = 0.48$. The resulting prior is what we refer to as the "fixed coupled prior structure." Note that unlike in the independent approach of Section 2.3.2, the two parts of the model are now coupled using information on θ_i, as shown by the dashed line in the diagram.

A problem with the coupled approach as described so far is that, although we are directly using the expert elicited information on the hurdle part of the model via the conditional means prior on $\boldsymbol{\beta} = (\beta_0, \beta_1)$, the expert elicited information on the count part is not being used. Rather, the information on the hurdle part is being transformed into information on the count part via the coupling function, with u and v fixed. Clearly, we do not want to ignore elicited information about the count part.

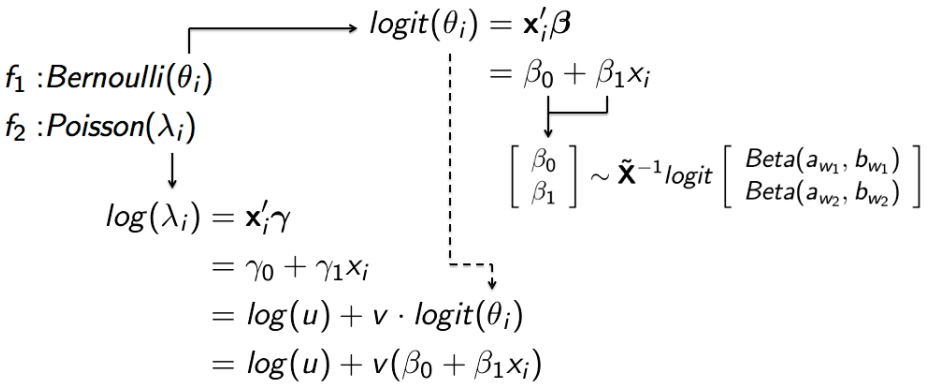

Figure 3.5: Bayesian Model With Coupled CMP.

Garthwaite et al. (2005) suggest that the last stage of prior elicitation should be to assess the accuracy of the elicitation. One way to do this for the fixed coupled prior is to determine whether or not the induced priors on λ_i are reasonable. The known information on the count part can then be used at this step as a basis of comparison. We determine the induced prior on λ_i at select weights as follows:

(1) Simulate values of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ from its conditional means prior.

(2) Use Equations (3.3) and (3.4) to transform the $\boldsymbol{\beta} = (\beta_0, \beta_1)$ values into corresponding values of $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$.

(3) The induced prior for λ_i at weight x_i is calculated as $\exp(\gamma_0 + \gamma_1 x_i)$.

Figure 3.6 plots information for λ_i at various weights. The information plotted in black represents what we previously elicited from the expert for the count side of the model. Recall that there are bounds at weights 200 and 400 lbs. from the independent conditional means prior elicitation. However, we did not elicit bounds for the three additional modal values in between. The red dots represent the 5th and 95th percentiles of the induced prior on λ_i at each weight via the fixed coupled prior.

Notice that the red dots do a reasonable job of reflecting the previously elicited information at 200 and 400 lbs. We did not elicit intervals for the three additional weights between 200 and 400 lbs; the reason being that that information has no direct use in the elicitation process. However, at this point, we can show the expert these induced intervals to determine if they believe them to be reasonable. Ideally, this

should not be a problem, given a reasonable coupling function. For our purposes, we assume that the expert relays back that these intervals are indeed reasonable and we conclude the elicitation was adequate. If this were not the case, it may be necessary to consider other possibilities for the coupling relationship, f, which would need to be explored as future work.

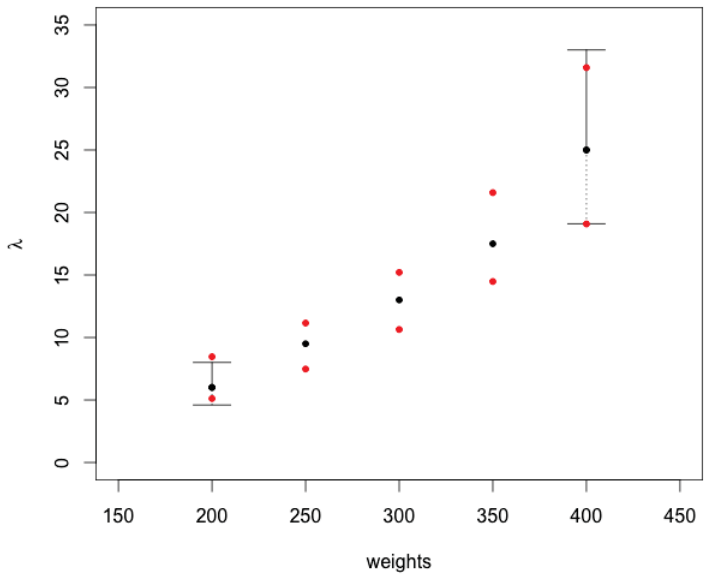

Figure 3.6: Induced 90% Intervals for λ_i at Specific Weights.

The coupling function f allows the priors on $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ to be dependent on the priors on $\boldsymbol{\beta} = (\beta_0, \beta_1)$ via the reparameterization given by equations (3.3) and (3.4). In contrast to the independent approach of Section 2.3.2, the values are now naturally "paired up" because each θ_i has a corresponding λ_i, as determined by the coupling function defined in equation (3.2). Simulated density plots for these regression parameters are shown in Figure 3.7. For comparison, the simulated densities for the independent approach are superimposed and represented by the dashed curves. Note that, due to how we structured the coupled prior, the simulated densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are exactly the same as in the independent approach. However, for $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, they are slightly shifted reflecting the effect of the coupling function.

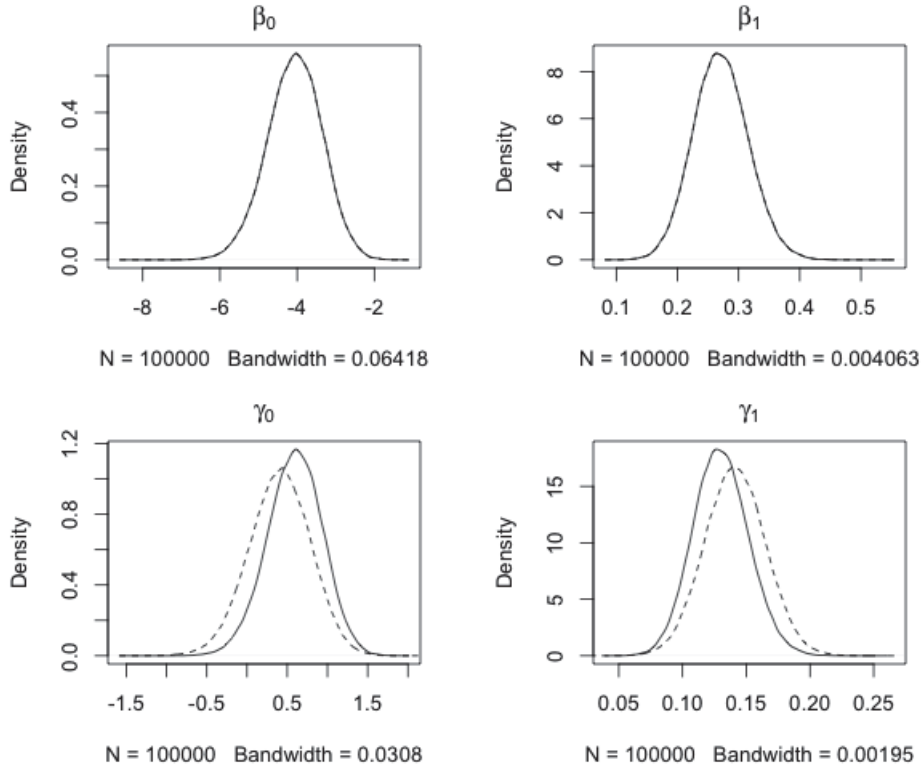

Figure 3.7: Simulated Density Plots for the Induced Priors on β_0, β_1, γ_0, and γ_1 under the Fixed Coupled Model (solid) and the Independent Model (dashed).

3.3.1.1 *Example*

We revisit the generated sleep disorder study outcomes shown in Figure 2.2. The Bayesian model is fit to this data under the fixed coupled prior structure with u and v fixed at the elicited least squares values, $u = 12.72$ and $v = 0.48$. Initial values were set at $\theta_{200} = 0.2$ and $\theta_{400} = 0.8$. To reduce autocorrelation issues, thinning was set at 10. As before, convergence issues were investigated using multiple chains and a variety of starting values resulting in little difference in the posteriors and Gelman-Rubin plots that all converged to one.

The resulting posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown by the solid curves in Figure 3.8. The corresponding posteriors under the independent conditional means priors model are shown by the dashed curves for reference. Table 3.1 summarizes the posterior means and 95% credible sets for each of the parameters, as well as for θ and λ at certain weights. For this data set, it appears that the fixed coupled model results in a slight decrease in posterior variability over the independent

model for each of the parameters, as seen graphically in Figure 3.8 and by comparing interval widths in Table 3.1. Further, the 95% credible sets still contain the true values. However, without extensive simulations we cannot be sure if this behavior always holds true or if it is just an artifact of the particular data set.

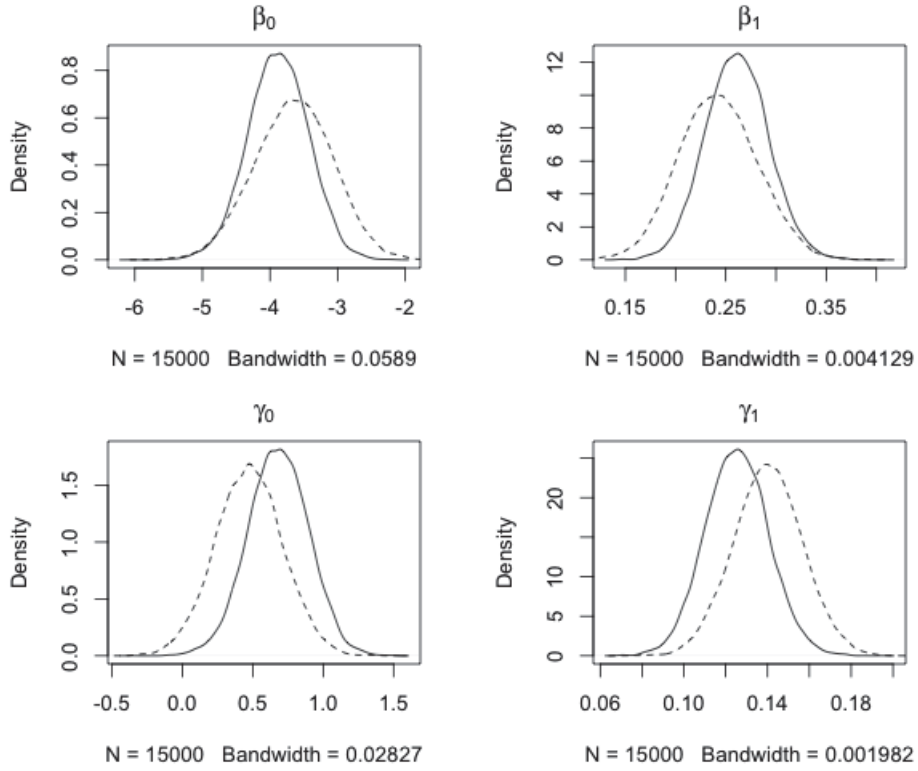

Figure 3.8: Posterior Density Plots for the Fixed Coupled Model (solid) and the Independent Model (dashed).

Table 3.1: Posterior Means and 95% Credible Sets for the Fixed Coupled Model

Parameter	Truth	Mean	2.5%	97.5%	Width
β_0	-4.18	-3.88	-4.77	-3.02	1.75
β_1	0.28	0.26	0.20	0.32	0.12
$\theta \vert x_i = 200$	0.20	0.22	0.17	0.27	0.10
$\theta \vert x_i = 300$	0.50	0.51	0.47	0.55	0.09
$\theta \vert x_i = 400$	0.80	0.79	0.72	0.85	0.13
γ_0	0.54	0.68	0.25	1.09	0.84
γ_1	0.13	0.13	0.10	0.16	0.06
$\lambda \vert x_i = 200$	6.56	6.94	6.00	7.97	1.97
$\lambda \vert x_i = 300$	12.84	12.97	11.92	14.08	2.16
$\lambda \vert x_i = 400$	25.15	24.37	20.01	29.48	9.47

One interesting observation is that the posterior intervals for the parameters on the hurdle part are narrower for the fixed coupled model, compared to the independent model, even though the priors appear to be the same in both approaches. The reason for this is that, when u and v are fixed, both parts of the model become dependent on $\boldsymbol{\beta} = (\beta_0, \beta_1)$, as $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ is simply a transformation of $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Thus, the positive counts are essentially used twice for estimation of β_0 and β_1: they are interpreted as hurdle crosses on the hurdle part and positive counts on the count part. This sort of double counting results in what appears to be improved estimation.

When we chose to fix u and v at their least squares estimates, we depicted a situation in which a highly informed expert provided information that conveniently led us to infer the exact coupling function from which the data was generated, as described in Appendix A. Consider instead that a somewhat less informed expert leads us to infer that the values should be fixed at $u = 10.27$ and $v = 0.75$. Note that these values were not chosen completely at random. Instead, this pair of (u, v) values were randomly selected from one of the 5,000 sampled (u, v) pairs that result in the curves plotted in Figure 3.11, which will be discussed in the following section.

We re-fit the Bayesian model with the fixed coupled prior structure, under the previous specifications, assuming that u and v are fixed at these newly specified values. The resulting posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown by the solid curves in Figure 3.9. This time the dashed curves in the figure show the corresponding posteriors under the previously fit fixed coupled model for reference. Posterior means and 95% credible sets are summarized in Table 3.2.

The results from this modified fixed coupled model show similarly narrow intervals as before. However, for this data set, the posterior means of the parameters do not approximate their true values as well. Particularly, there are certain parameters for which the true value is not contained within its 95% credible set. Again, extensive simulations must be done before we can determine if this is always the case. However, the varied results from this one example suggests that the performance of

the Bayesian model under the fixed coupled prior is not surprisingly affected by the choices of u and v.

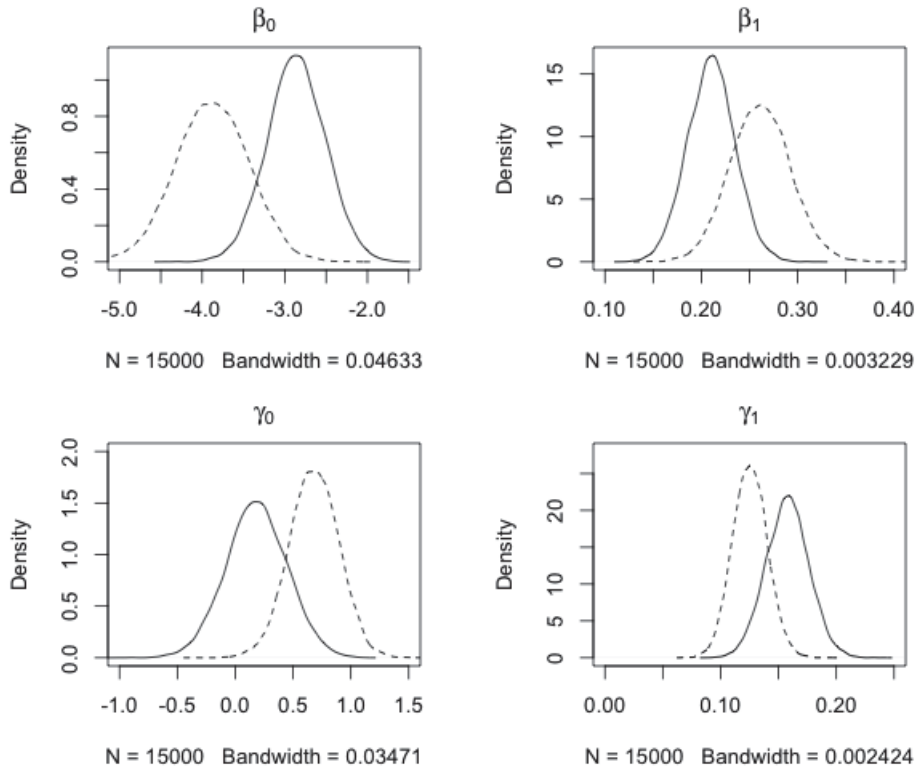

Figure 3.9: Posterior Density Plots for the Fixed Coupled Model with $u = 10.27$ and $v = 0.75$ (solid) Compared to When $u = 12.72$ and $v = 0.48$ (dashed).

Table 3.2: Posterior Means and 95% Credible Sets for the Fixed Coupled Model with $u = 10.27$ and $v = 0.75$

Parameter	Truth	Mean	2.5%	97.5%	Width
β_0	-4.18	-2.86	-3.55	-2.17	1.39
β_1	0.28	0.21	0.16	0.26	0.10
$\theta\|x_i = 200$	0.20	0.32	0.27	0.37	0.10
$\theta\|x_i = 300$	0.50	0.57	0.54	0.60	0.06
$\theta\|x_i = 400$	0.80	0.79	0.74	0.84	0.10
γ_0	0.54	0.18	-0.34	0.70	1.04
γ_1	0.13	0.16	0.12	0.19	0.07
$\lambda\|x_i = 200$	6.56	5.84	4.89	6.90	2.01
$\lambda\|x_i = 300$	12.84	12.82	11.69	14.04	2.35
$\lambda\|x_i = 400$	25.15	28.39	22.42	35.60	13.18

3.3.2 Variable Approach

The disadvantages to the fixed coupled prior include its dependency on the selection of an exact coupling function as well as its inability to directly make use of the existing expert opinion regarding the count part of the model. One way to address these concerns is to add an additional dimension to the proposed Bayesian model, with informative priors placed on u and v in order to better reflect the uncertainty of the coupling function. We will refer to this as the variable coupled prior, as the coupling function is now varying.

Informative priors for u and v can be determined as follows. Recall that applying the log-link to the coupling function conveniently reparameterized γ_0 and γ_1 as

$$\gamma_0 = \log(u) + v\beta_0$$

and

$$\gamma_1 = v\beta_1.$$

Solving for u and v, this becomes

$$u = \exp\left(\gamma_0 - \gamma_1 \frac{\beta_0}{\beta_1}\right) \tag{3.5}$$

and

$$v = \frac{\gamma_1}{\beta_1}.$$

From the independent conditional means prior elicitation described in Chapter Two, we have induced priors on both $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$. By sampling from these induced priors, a joint prior on u and v can be induced via (3.5). As this joint prior has no closed form, we approximate it with a fitted bivariate normal distribution, which acts as an informative joint prior on u and v. Figure 3.10 shows the sampled marginal densities, represented by the dashed curves, along with the fitted bivariate normal, represented by the solid curves.

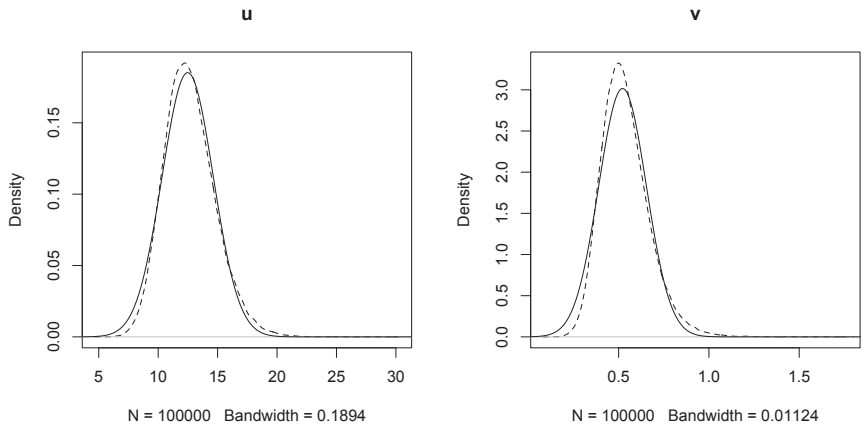

Figure 3.10: Informative Priors on u and v.

In the fixed approach, the prior elicitation on the count side of the model, for $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, was used only indirectly, serving as a checking reference in the last step of elicitation. The variable approach uses this information more directly. To see how the joint prior on u and v better reflects our uncertainty, we sample 5,000 pairs of (u, v) values and plot the resulting curves, as determined by equation (3.2), in Figure 3.11. The curves were plotted with transparency, so the darker red area reflects where the majority of the curves fall. This new Bayesian model with variable coupled priors is diagrammed in Figure 3.12.

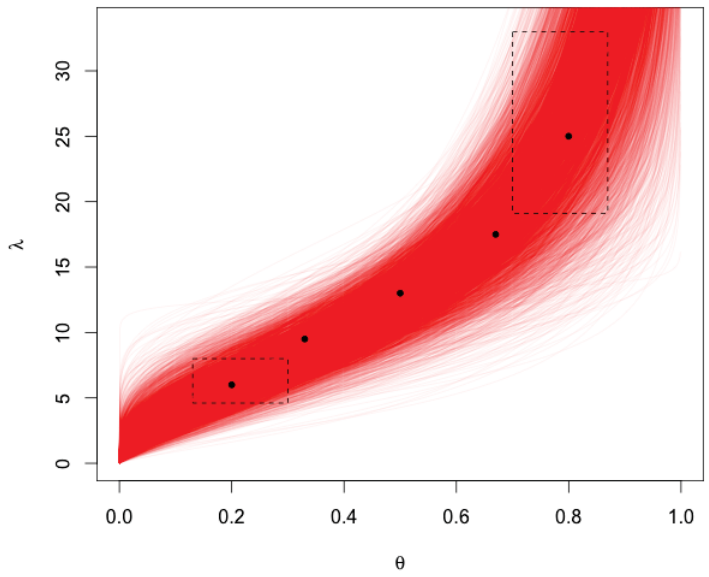

Figure 3.11: 5,000 Possible Curves Relating θ_i and λ_i.

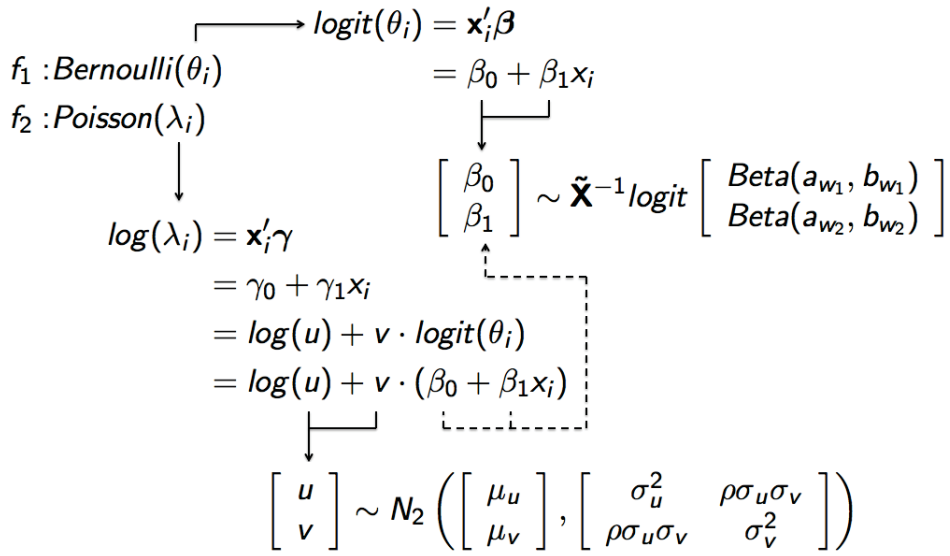

Figure 3.12: Bayesian Model with Variable Coupled CMP.

Simulated density plots for the induced priors on $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ resulting from this structure are shown by the solid curves in Figure 3.13. For reference, the dashed curves show the induced priors resulting from the previously described fixed coupled prior. Again, the priors on $\boldsymbol{\beta} = (\beta_0, \beta_1)$ remain the same. But, as a result of the priors on u and v, there is now much more variability on the priors for $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$.

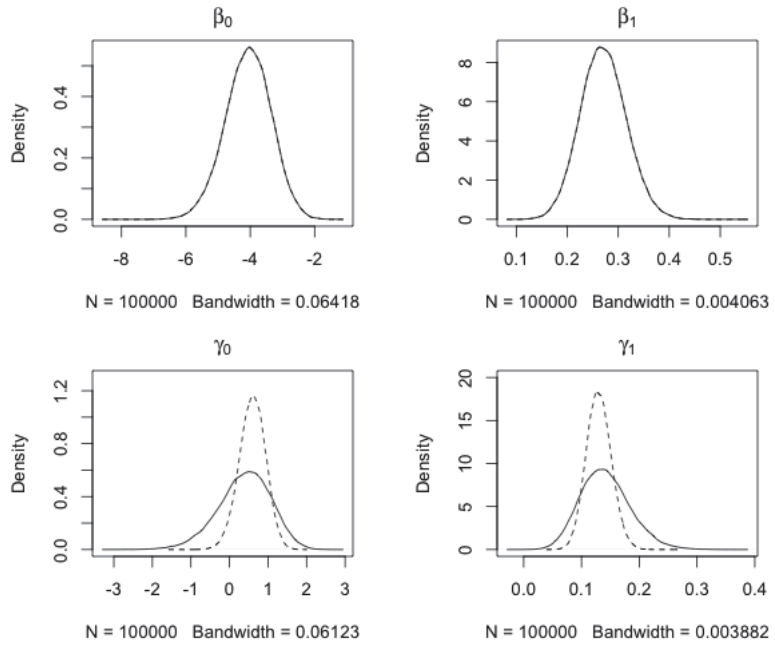

Figure 3.13: Simulated Density Plots for the Induced Priors on β_0, β_1, γ_0, and γ_1 Under the Variable Coupled Model (solid) and the Fixed Coupled Model (dashed).

3.3.2.1 *Example*

Consider again the generated sleep disorder study outcomes shown in Figure 2.2. This time the Bayesian model is fit to this data under the variable coupled prior structure using Markov chain Monte Carlo (MCMC) methods in OpenBUGS. Once again, one chain with 16,000 iterations, including a burn-in of 1,000 iterations, was used. Initial values were set at $\theta_{200} = 0.2$, $\theta_{400} = 0.8$, and $(u, v) = (0, 0)$. Thinning was set at 10 to reduce autocorrelation and, as before, convergence issues were investigated using multiple chains and a variety of starting values, finding little difference in the posteriors, and all Gelman-Rubin plots converged to one.

Posterior densities for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ are shown by the solid curves in Figure 3.14. The posteriors from the independent model are shown by the dashed curves for reference. Table 3.3 shows posterior means and 95% credible sets for each of the parameters, as well as for θ and λ at certain weights.

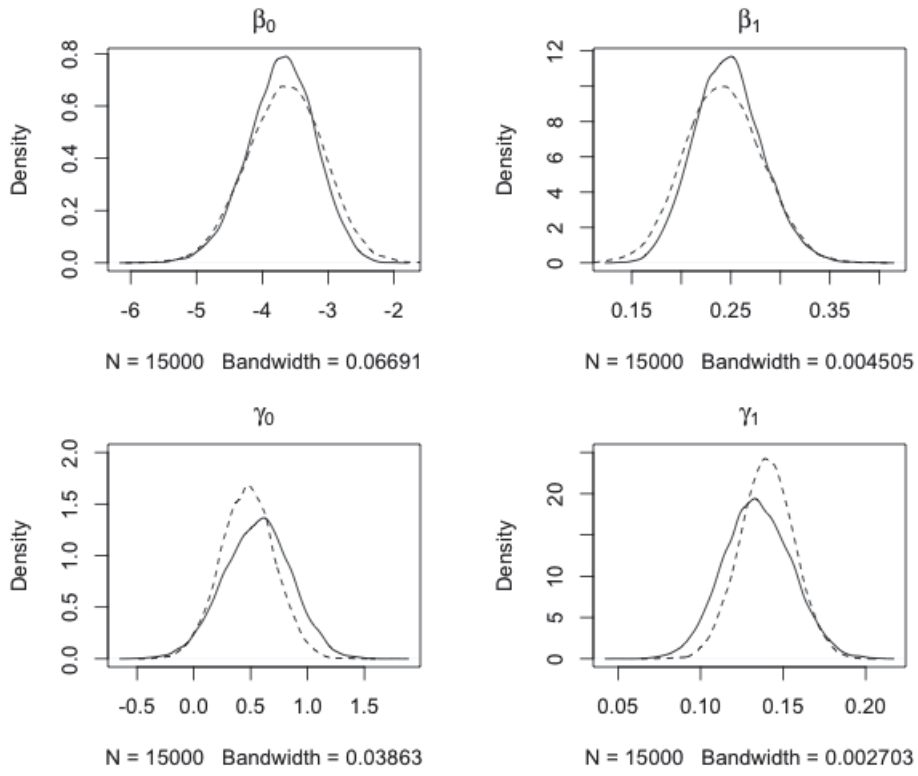

Figure 3.14: Posterior Density Plots for the Variable Coupled Model (solid) and the Independent Model (dashed).

Table 3.3: Posterior Means and 95% Credible Sets for the Variable Coupled Model

Parameter	Truth	Mean	2.5%	97.5%	Width
β_0	-4.18	-3.71	-4.76	-2.75	2.01
β_1	0.28	0.25	0.18	0.32	0.13
$\theta\|x_i = 200$	0.20	0.23	0.16	0.30	0.14
$\theta\|x_i = 300$	0.50	0.50	0.44	0.57	0.13
$\theta\|x_i = 400$	0.80	0.77	0.69	0.85	0.16
γ_0	0.54	0.56	-0.02	1.12	1.14
γ_1	0.13	0.13	0.10	0.18	0.08
$\lambda\|x_i = 200$	6.56	6.73	5.54	8.07	2.54
$\lambda\|x_i = 300$	12.84	13.15	11.91	14.42	2.51
$\lambda\|x_i = 400$	25.15	25.95	20.04	33.17	13.13

For the parameters on the hurdle part, the variable coupled model has slightly less posterior variability compared to the independent model. This suggests that we are still seeing the effect of having $\boldsymbol{\beta} = (\beta_0, \beta_1)$ involved in both parts of the model, as discussed earlier. However, on the count part, there is increased posterior variability, reflecting the increased variability in the variable coupled prior that was a result of adding priors on u and v. Still, all the 95% credible sets contain the true parameter values.

3.3.3 Discussion

Despite the increased posterior variability over the fixed approach, we suggest that the variable coupled approach results in the more practical coupled prior which has some potential gains over the independent prior described in Chapter Two. There are several reasons for this.

So far, we have only considered marginal posterior results for all the models. Suppose we are interested in simultaneous inference on β_1 and γ_1. Consider the joint posterior contours of β_1 and γ_1 for the independent conditional means prior model versus the variable coupled model based on one sample of $n = 100$ observations shown in Figure 3.15. The contours show that the posteriors for β_1 and γ_1 are independent under the independent conditional means prior model, but there is a dependence under the variable coupled conditional means prior model.

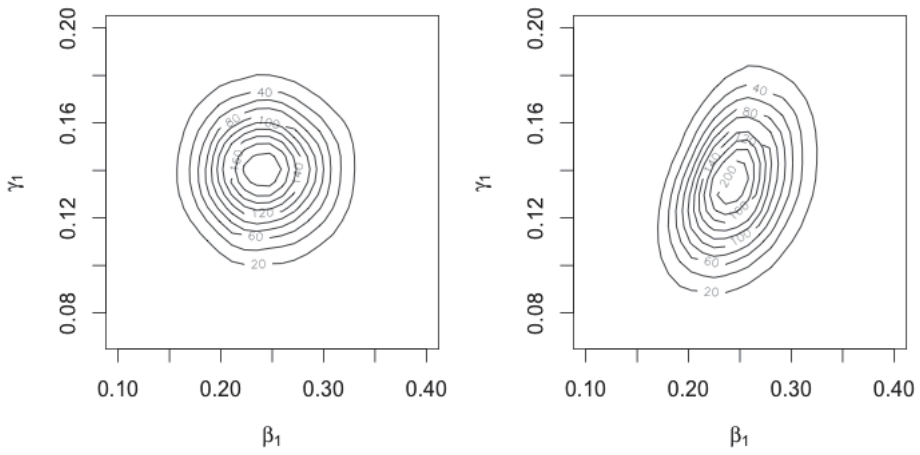

Figure 3.15: Joint Posterior Contours for β_1 and γ_1 Under the Independent Model (left) and the Variable Coupled Model (right).

Having knowledge of a joint posterior opens up the opportunity for joint hypothesis testing on this data. Additionally, there may be situations in which it is useful to know the relationship between β_1 and γ_1. For example, given a subpopulation for which we know the change in log odds, β_1, we can make inference on the corresponding change in log rate, γ_1, based on this joint posterior. This sort of scenario may occur if we were to introduce random effects into the given model such that β_1 and γ_1 will vary in subpopulations. These ideas will need to be explored further as future work.

Also, we suspect that the influence of the variable coupled approach can be more clearly seen when considering a parameter of interest that is a function of both θ and λ at specific weights. For example, say that, for whatever reason, we are interested in the ratio of θ/λ for a subject weighing 300 lbs. The posterior density for this ratio under the variable coupled model is shown by the solid curve in Figure 3.16. For reference the posterior density under the independent model is shown by the dashed curve. Note that the variable coupled model appears to result in less posterior variability for this ratio.

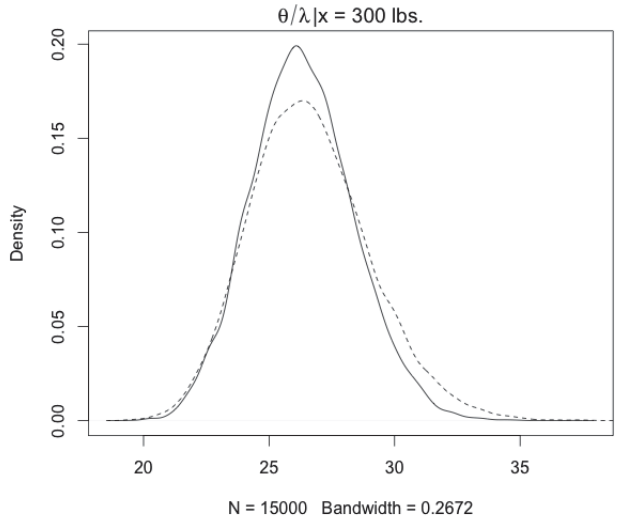

Figure 3.16: Posterior Density for λ/θ at 300 lbs. Under the Variable Coupled Model (solid) and the Independent Model (dashed).

3.4 Simulation

To better assess how the variable coupled model performs relative to the independent model described in Chapter Two, we perform a small scale simulation study. We generate 100 data sets of size $n = 100$ from the same set of parameters used to generate the sleep disorder study outcomes shown in Figure 2.2. For each data set, we fit the two Bayesian models under the previously described specifications. We record the posterior means and 95% credible sets for $\boldsymbol{\beta} = (\beta_0, \beta_1)$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, select values of θ and λ at specific weights, as well as the ratio, λ/θ, at 300 lbs., for each iteration of the simulation.

Simulation results for the independent model are summarized in Table 3.4 and results for the variable coupled model are summarized in Table 3.5. Each table shows the true value of each parameter, the median of the posterior means and 95% credible intervals, and the median width of the intervals. Figures 3.17, 3.18, and 3.19 show this information in box plots. Recall that the horizontal line represents the true value of the parameters, the center point is the median of the 100 posterior means, and the upper and lower limits represent the average of the upper and lower bounds across 100 posterior 95% credible sets. The grey bars represent ± 1 simulation standard deviation above and below each respective point on the box plot, which occasionally

overlap resulting in a darker shade of grey. Further, in each individual plot, the box plot on the left shows results under the independent model and the box plot on the right shows results under the variable coupled model.

Table 3.4: Simulation Results for the Independent CMP Model with $n = 100$

Parameter	True	2.5%	Median	97.5%	Width
β_0	-4.18	-5.29	-4.07	-2.92	2.37
β_1	0.28	0.20	0.27	0.35	0.16
$\theta \vert x_i = 200$	0.20	0.14	0.21	0.29	0.15
$\theta \vert x_i = 300$	0.50	0.43	0.50	0.58	0.15
$\theta \vert x_i = 400$	0.80	0.70	0.79	0.87	0.17
γ_0	0.54	0.05	0.52	0.99	0.94
γ_1	0.13	0.10	0.14	0.17	0.06
$\lambda \vert x_i = 200$	6.56	5.48	6.56	7.73	2.25
$\lambda \vert x_i = 300$	12.84	11.93	12.95	14.02	2.12
$\lambda \vert x_i = 400$	25.15	21.23	25.68	30.57	9.15
$\lambda/\theta \vert x_i = 300$	25.69	21.82	25.97	31.10	9.01

Table 3.5: Simulation Results for the Variable Coupled Model with $n = 100$

Parameter	True	2.5%	Median	97.5%	Width
β_0	-4.18	-5.07	-3.97	-2.96	2.14
β_1	0.28	0.20	0.27	0.34	0.14
$\theta \vert x_i = 200$	0.20	0.15	0.21	0.29	0.14
$\theta \vert x_i = 300$	0.50	0.44	0.50	0.57	0.14
$\theta \vert x_i = 400$	0.80	0.71	0.79	0.87	0.16
γ_0	0.54	0.00	0.57	1.14	1.15
γ_1	0.13	0.09	0.13	0.17	0.08
$\lambda \vert x_i = 200$	6.56	5.43	6.73	8.11	2.68
$\lambda \vert x_i = 300$	12.84	11.87	12.95	14.12	2.23
$\lambda \vert x_i = 400$	25.15	19.99	25.07	31.03	10.83
$\lambda/\theta \vert x_i = 300$	25.69	22.12	25.87	30.15	7.87

Overall, the results from this simulation mirror what we observed from the single example we previously considered. For all parameters considered, there is no indication that any one model is resulting in more or less bias. For the parameters on the hurdle part, the variable coupled model results in narrower posterior intervals, on average, than the independent model. The opposite is true for the parameters on the count part. However, when considering the ratio λ/θ at 300 lbs., as suspected,

the posterior intervals are narrower under the variable coupled model compared to the independent model.

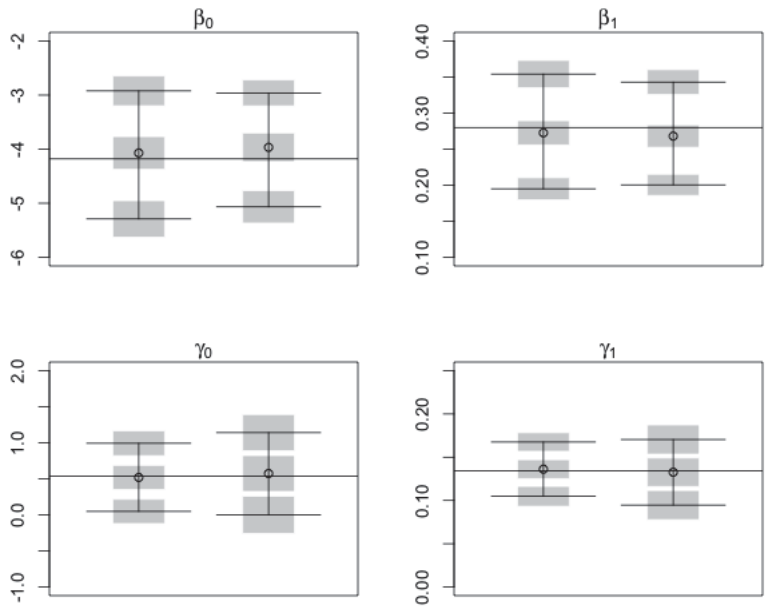

Figure 3.17: Simulation Results for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$.

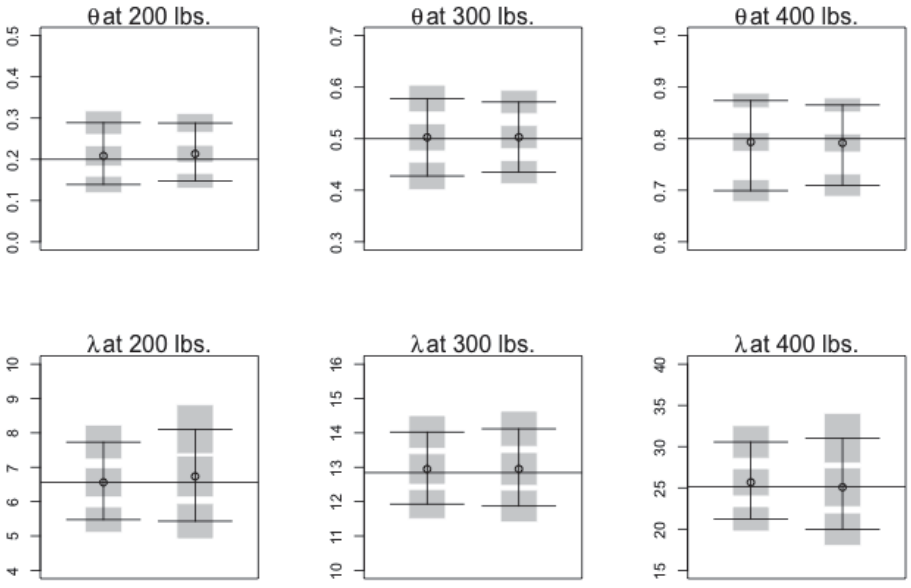

Figure 3.18: Simulation Results for Select Values of θ and λ for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$.

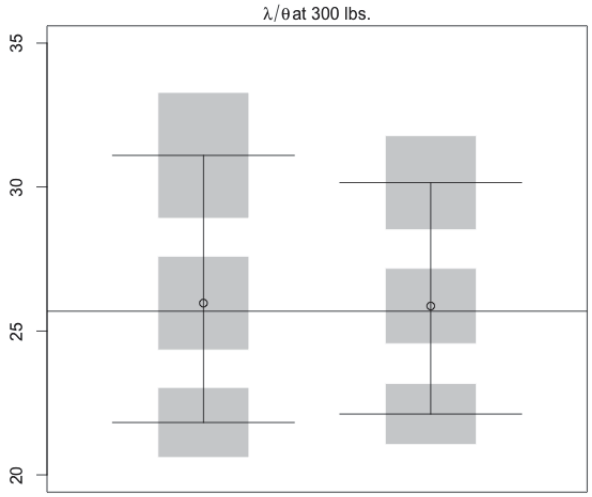

Figure 3.19: Simulation Results for λ/θ at 300 lbs. for the Independent CMP Model (left) and the Variable Coupled Model (right) with $n = 100$.

3.5 Summary & Future Work

In this chapter, we sought to develop a prior structure that was able to take advantage of the dependency inherent between the two parts of the Bernoulli-Poisson hurdle model in the case of shared covariates. We proposed a method that relied on eliciting a functional relationship between θ_i and λ_i, thus coupling the information in the hurdle and count parts of the model. We described two such prior structures to take advantage of the relationship: the fixed coupled prior and the variable coupled prior.

The two proposed approaches both have their advantages and disadvantages. In a sense, they represent two extremes. The fixed coupled prior is straightforward to implement but is very dependent on the chosen values of u and v. Fixing these values effectively eliminates all parameters from the model except for $\boldsymbol{\beta} = (\beta_0, \beta_1)$. The priors on these parameters become the model's only source of variability and there is no direct use for the information known about the count part of the model. The variable coupled prior addresses some of these issues by placing informative priors on u and v rather than fixing them. We proposed a method of inducing priors on u and v which made use of the information the count part of the model. However, this method may have led us to double count some variability as the priors for $\boldsymbol{\beta}$ were

reused when inducing the joint prior on (u, v). Indeed the resulting prior structure has inflated variability on the count part of the model. However, the variable coupled model has the added benefit of allowing for joint posterior inference. Thus, we believe the variable coupled approach is more practical than the fixed coupled approach.

We further evaluated how the variable coupled Bayesian model performed on generated sleep disorder study outcomes relative to the independent model described in Chapter Two. The results from a short simulation study suggest that, in this scenario, there is not much to gain marginally from using the coupled prior. However, we found that the influence of the coupled model may be more apparent when interest is in some function of parameters from both parts of the model. Still, we only considered how these models perform for a single set of parameters designed to emulate this sleep disorder scenario. For future work, a more extensive simulation study exploring other scenarios with zero-inflated count data should be performed to fully understand the performance of these prior structures.

We believe a more practical coupled prior would be a medium between the fixed and variable approaches. One possibility is to consider how the induced prior on (u, v) can be made more informative, so another avenue for future work would be to explore how this can be done. Additionally, we believe the idea of a coupled prior, with a family of coupling functions, can be extended to hurdle models with more than one shared covariate and generalized to other similar applications.

CHAPTER FOUR

Sample Size Determination for Hurdle Models

4.1 Introduction

Sample size determination is an important aspect of experimental design for clinical trials. This is not a new problem in the realm of zero-inflated count data and has been addressed in the literature. Williamson et al. (2007) perform power calculations for zero-inflated count models and suggest that a reverse approach can be used to calculate sample size. Lachenbruch (2001) investigate power for two-part models for semicontinuous data, and derive sample size calculations for these models, noting that their procedures can be modified for the discrete case. Channouf et al. (2014) modify a method by Shieh (2001) for sample size determination for Poisson regression models and extend their methodology to the zero-inflated Poisson regression model.

However, the cited methodologies are all based on the frequentist approach to sample size determination. There are advantages to a Bayesian approach to this problem, which are worth exploring. In this chapter, we give an overview of the Bayesian two-prior approach to sample size determination, describe how it can be applied in the context of hurdle models for count data, and show its results when applied to the hypothetical sleep disorder study first introduced in Section 2.1.4.

4.2 Bayesian Sample Size Determination

One approach to Bayesian sample size determination is the "two-priors" approach, which makes use of "design" and "analysis" priors. This method has been well covered by Adcock (1997), Joseph et al. (1997), Wang and Gelfand (2002), and Brutti et al. (2008). Suppose φ is the parameter of interest and, for a sample of size n, the vector of observations $\mathbf{y}_n = (y_1, y_2, ..., y_n)$ has density $f(\mathbf{y}_n|\varphi)$. In frequentist

sample size determination, it is sometimes necessary to fix the parameter of interest at a certain value, known as a planning estimate. In the Bayesian approach, the design prior replaces this fixed value with a range of values, which allows pre-experimental uncertainty to be incorporated in the process. Additionally, each parameter in the model, not just the parameter of interest, is given such a design prior. Elicitation for these design priors is based on either certain regulatory requirements, previous studies, or expert opinion.

This approach to sample size determination is typically carried out via simulation. The design priors are used to generate values for all the necessary parameters and covariates in the model. These parameter and covariate values are then used to generate a data set for the model. Given a generated data set, the analysis prior is used in the actual fitting of the Bayesian model. Typically analysis priors are chosen to be rather diffuse, in contrast to relatively informative design priors.

The Bayesian sample size determination algorithm repeats this data generation and model fitting process for a number of iterations across various sample sizes. Then, the optimal sample size n is chosen to satisfy a certain criterion. There are many such criterion, all based on optimizing certain aspects of the posterior distribution of φ, such as average variance, average power, or average interval length.

4.3 Application to the Bernoulli-Poisson Hurdle Model

In this section, we describe how Bayesian sample size determination can be implemented on the Bernoulli-Poisson hurdle model discussed in Chapters Two and Three, whose likelihood is given by equation (2.5), with

$$\text{logit}(\theta_i) = \mathbf{z}_i'\boldsymbol{\beta} \tag{4.1}$$

and

$$\log(\lambda_i) = \mathbf{x}_i'\boldsymbol{\gamma}, \tag{4.2}$$

where $\mathbf{z}_i$ and $\mathbf{x}_i$ represent vectors of explanatory variables, which may have common components, and $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are the corresponding regression coefficient vectors. Recall

that in Chapters Two and Three, we discussed informative prior structures for this model in the special case where both parts of the model shared the same covariate. However, here we focus more generally on sample size determination for a model where both sides do not necessarily share covariates.

As discussed in Section 2.3.1, one method for eliciting informative priors for regression parameters in generalized linear models is conditional means priors, as developed by Bedrick et al. (1996). Recall that conditional means priors allow for information to be elicited at an operational level, inducing informative priors on the regression coefficients. The resulting priors typically do not have closed form, but we can fit distributions to them, which become the design priors for each parameter.

Since there are covariates on both parts of the model, appropriate design priors for these covariates, $\mathbf{z}_i$ and $\mathbf{x}_i$, must also be specified based on the population at hand. The choice of distribution depends on what the covariate is meant to represent. Generally, binary treatment covariates are generated from Bernoulli distributions and continuous covariates are generated from uniform or normal distributions. Also, if the covariate is expected to be skewed, this can be taken into account through use of a skewed distribution. In contrast, analysis priors are meant to be appropriately diffuse. The typical diffuse prior used on regression coefficients is $Normal(0, \sigma^2)$ for a sufficiently large σ chosen to fit the scenario.

Suppose we are interested in the average length criterion (ALC), which looks for the necessary sample size to achieve a desired average interval length. With the design and analysis priors and the parameter of interest specified, following Stamey et al. (2013), the steps for sample size determination for this model are shown in Algorithm 1.

for *a range of sample sizes,* n **do**

  for $b = 1, ..., B$ *Monte Carlo iterations at sample size* n **do**

    Generate values of parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ from their design priors;

    Generate values for the covariates $\mathbf{z}_i$ and $\mathbf{x}_i$ from specified distributions;

    Compute θ_i and λ_i according to equations (4.1) and (4.2);

    Generate response data y_i using the values generated. In this step we follow the data generation method described in Appendix A;

    Fit the Bayesian Bernoulli-Poisson hurdle model using the analysis priors. A Gibbs sampler such as WinBUGS/OpenBUGS is usually used at this step;

    Record the length of the $1 - \alpha$ credible interval for the specified parameter of interest;

  Average B values to get the average length, l, at sample size n;

Plot and fit a curve through the points (n, l);

Algorithm 1: Sample Size Determination for the Bernoulli-Poisson Hurdle Model.

4.4 Simulation

To illustrate this process, we use Algorithm 1 to determine the appropriate sample size under the average length criterion (ALC) for the hypothetical sleep disorder study. It should be noted that in the sleep disorder scenario, the recorded weights are a shared covariate on both parts of the model. Thus, it would likely be possible to adapt the coupled prior described in Chapter Three when defining design priors for this scenario. However, the general method for sample size determination described in Section 4.3 can still be implemented, so we will not demonstrate that approach.

4.4.1 Design and Analysis Priors

The design and analysis priors used in this simulation are summarized in Table 4.1. We let the conditional means priors from Section 2.3.2 act as the design priors

for the two sets of regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$. Since these priors do not have closed forms, we fit normal distributions to them. Figure 4.1 shows the normal fit (solid curve) superimposed upon the simulated densities (dashed curve) originally shown in Figures 2.7 and 2.9. Thus, we confirm that these normal distributions are an adequate approximation.

Table 4.1: Design and Analysis Priors

Parameter	Design Prior	Analysis Prior
β_0	$N(-4.049, 0.713)$	$N(-4.049, 2.646)$
		$N(0, 25)$
β_1	$N(0.270, 0.045)$	$N(0.270, 0.227)$
		$N(0, 25)$
γ_0	$N(0.413, 0.381)$	$N(0.413, 1.282)$
		$N(0, 25)$
γ_1	$N(0.141, 0.024)$	$N(0.141, 0.120)$
		$N(0, 25)$
X	$Beta_{[200,400]}(1.7, 3)$	NA
	$Beta_{[200,400]}(3, 1.7)$	

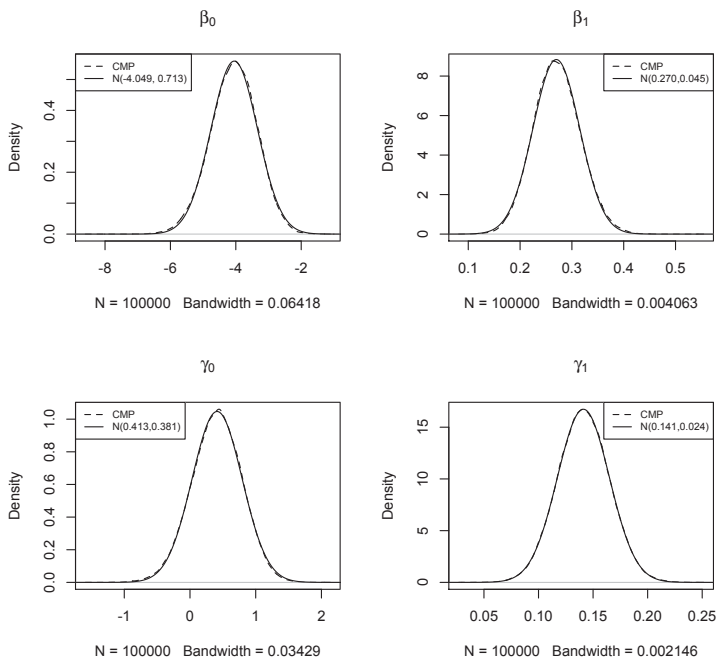

Figure 4.1: Normal Distributions Fitted to Simulated Densities for β_0, β_1, γ_0, and γ_1.

It is also necessary to specify a design prior for the weight covariate. For this simulation study, we consider two such specifications to better understand the sample size problem. Suppose we assume that the overweight subjects in the study all have weights in the range of 200 to 400 lbs. For one specification, we further assume that the majority weigh closer to 200 than 400 lbs. For the other, we assume the opposite, that the majority weigh closer to 400 than 200 lbs. For both cases, we choose appropriate shifted beta distributions to represent these beliefs, whose densities are plotted in Figure 4.2.

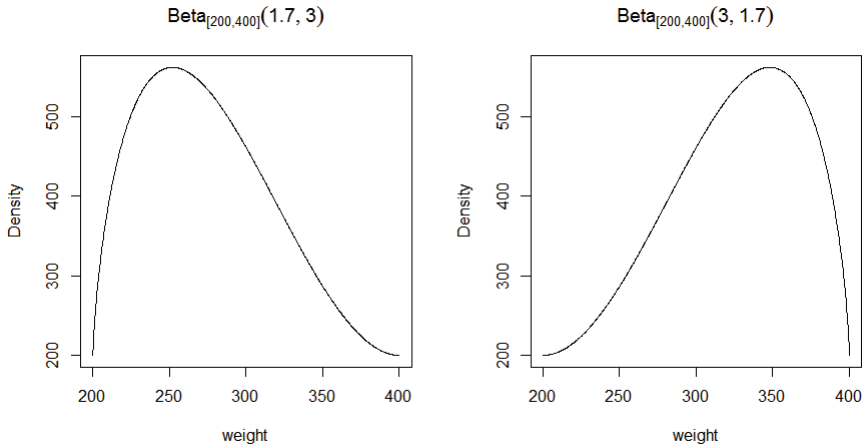

Figure 4.2: Design Priors for Weight Covariate.

Recall that in contrast to the design priors, analysis priors are typically meant to be rather diffuse. However, at the same time, they should not be unrealistically so. Suppose one way to specify appropriately diffuse analysis priors is to center them at the means of the design priors but inflate the standard deviation. The question then becomes, how large should the standard deviation be.

Consider the following reasoning for β_0. We know that the probability of suffering from a sleep disorder should not fall below 10% for 200 lb. subjects or above 90% for 400 lb. subjects. This means the log odds, $\beta_0 + \beta_1 x_i$, should stay between -2.197 and 2.197 for all values of the covariate x_i. From its design prior, we know that β_1 mainly stays between $0.270 \pm 2(0.045) = (0.180, 0.360)$. It follows that, for a

subject weighing 200 lbs., β_0 should be in the following range:

$$\begin{aligned} \beta_0 + 0.180(200/20) &= -2.197 \Rightarrow \beta_0 = -3.997 \\ \beta_0 + 0.360(200/20) &= -2.197 \Rightarrow \beta_0 = -5.797. \end{aligned}$$

Similarly, for a subject weighing 400 lbs.,

$$\begin{aligned} \beta_0 + 0.180(400/20) &= 2.197 \Rightarrow \beta_0 = -1.403 \\ \beta_0 + 0.360(400/20) &= 2.197 \Rightarrow \beta_0 = -5.003. \end{aligned}$$

Note that the weights are divided by 20 because, in our data generation scheme described in Appendix A, we chose to work with 20 lb. units of weight. With this logic, we determine that all reasonable values for β_0 should fall between -5.797 and -1.403. Note that the mean of the design prior for β_0 is -4.049, which falls within this range, but not directly in the middle. In fact, the mean has a distance of $|-4.049-(-5.797)| = 1.748$ from the lower bound and a distance of $|-4.049-(-1.403)| = 2.646$ from the upper bound. Following the empirical rule that 95% of the values in a normal distribution lie within two standard deviations of the mean, we could set the standard deviation of the analysis prior to be half of the larger difference. However, we make the more conservative choice and let the entire difference, 2.646, be the standard deviation of analysis prior for β_0.

We can repeat this reasoning to determine the dispersion of the analysis prior for β_1. From its design prior, we know that β_0 mainly stays between $-4.049 \pm 2(0.713) = (-5.475, -2.623)$. As before, the log odds should be between -2.197 and 2.197 for all values of the covariate x_i. Thus, all reasonable values for β_1 should be between $(-2.197-(-2.623))/(200/20) = 0.043$ and $(2.197-(-5.475))/(400/20) = 0.384$. Again, the mean of the design prior does not fall exactly in the middle of this range. It has a distance of $|0.27-0.043| = 0.227$ from the lower bound and a distance of $|0.27-0.384| = 0.114|$ from the upper bound. Following previous reasoning, we set the standard deviation of the analysis prior of β_1 to be 0.227.

The same reasoning extends to the analysis priors for γ_0 and γ_1. Suppose we believe the expected number of sleep interruptions for a subject suffering from a sleep disorder should be greater than 4 for 200 lb. subjects and less than 35 for 400 lb. subjects. Thus, $\log(\lambda_i) = \gamma_0 + \gamma_1 x_i$ should stay between 1.386 and 3.555 for all values of the covariate x_i. Similar calculations and reasoning as before justify the analysis priors for γ_0 and γ_1 shown in Table 4.1.

In addition to "realistically" diffuse analysis priors, we also consider "unrealistically" diffuse analysis priors centered at zero with a large standard deviation $\sigma = 25$ to investigate the affect of the choice of analysis priors on sample size recommendations.

4.4.2 Results

Following the algorithm described in Section 4.3, we consider samples of size $n = 50$ through 200 in increments of 25. For each sample size, we generate 200 replications of parameter and covariate values and their corresponding data sets. We first consider the design prior for the weight covariate x_i that is skewed towards the lower weights. We fit the Bayesian model using OpenBUGS with realistically diffuse analysis priors under the same specifications as the previous chapters: we ran 16,000 iterations, discarding the first 1,000 as burn-in, with thinning set at 10.

For each replication, we record the posterior samples for $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$. Suppose the parameter of interest is the average amount of sleep interruptions experienced by subjects of certain weights who suffer from sleep disorders, defined as $\lambda_i = \exp(\gamma_0 + \gamma_1 x_i)$. Since λ_i is a function of the parameters $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$, we can transform the recorded posterior samples into a posterior sample of λ_i at various weights. Figure 4.3 summarizes resulting posterior samples for λ_i at various weights between 200 and 400 lbs. for four select instances of the 200 replications for sample size $n = 100$. The solid line represents the posterior median and the dashed bounds represent 95% credible intervals. Note that the scale of the y axis changes from each instance, reflecting the changing true values of $\boldsymbol{\beta} = (\beta_0, \beta_1)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1)$ in the

simulation. But, generally, we see that the interval widths narrow at around 300 lbs. and then begin to spread out as weight increases to 400 lbs.

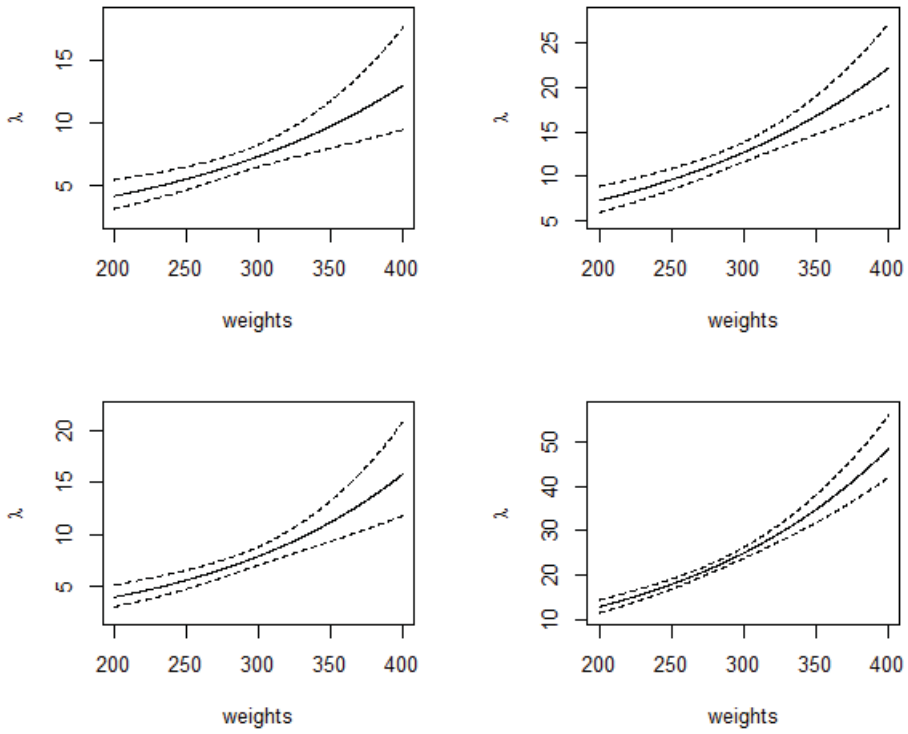

Figure 4.3: Posterior Samples for λ_i in Four Instances of the 200 Replications for $n = 100$ with $x_i \sim \beta(1.7, 3)$.

In this manner, we record the 95% interval lengths of λ_i at certain covariate specifications across all 200 replications for a range of sample sizes. Table 4.2 shows the coverage of these 200 intervals for λ_i at various weights across the range of sample sizes. The lowest coverage was 0.915. Figure 4.4 shows the density of the 200 interval lengths, at sample size $n = 100$,of λ_i for a subject weighing 225 lbs. Similar density plots for the other weights and sample sizes exhibited similar behavior. Typically, and according to the algorithm, we record the mean of the 200 lengths to summarize these results. However, due to the skewness of this distribution, we use the median interval length instead.

Table 4.2: Coverage for λ_i at Various Weights Across the 200 Intervals with $x_i \sim \beta(1.7, 3)$

n	225 lbs.	275 lbs.	325 lbs.	375 lbs.
50	0.950	0.945	0.975	0.995
75	0.920	0.915	0.945	0.945
100	0.960	0.960	0.970	0.980
125	0.970	0.930	0.955	0.985
150	0.950	0.965	0.935	0.920
175	0.945	0.935	0.940	0.955
200	0.960	0.945	0.955	0.965

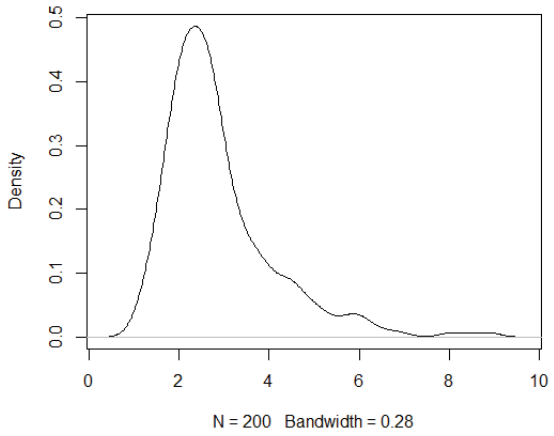

Figure 4.4: Density of 200 Interval Lengths for λ_i at 225 lbs. with $x_i \sim \beta(1.7, 3)$.

Figure 4.5 plots the median interval widths across the sample sizes for λ_i at $x_i = 225, 275, 325$, and 375. We see that λ_i at 375 lbs. results in by far the largest interval widths over all the sample sizes followed by λ_i at 325, 225, and 275 lbs. These results are logical as the design prior on weight shows that the majority of the generated weights are be between 225 and 275 lbs. with significantly fewer weights on the upper end of the scale. That is, the necessary sample size for a study where the parameter of interest is λ_i at larger weights is understandably larger to ensure that the sample will have enough subjects with larger weights.

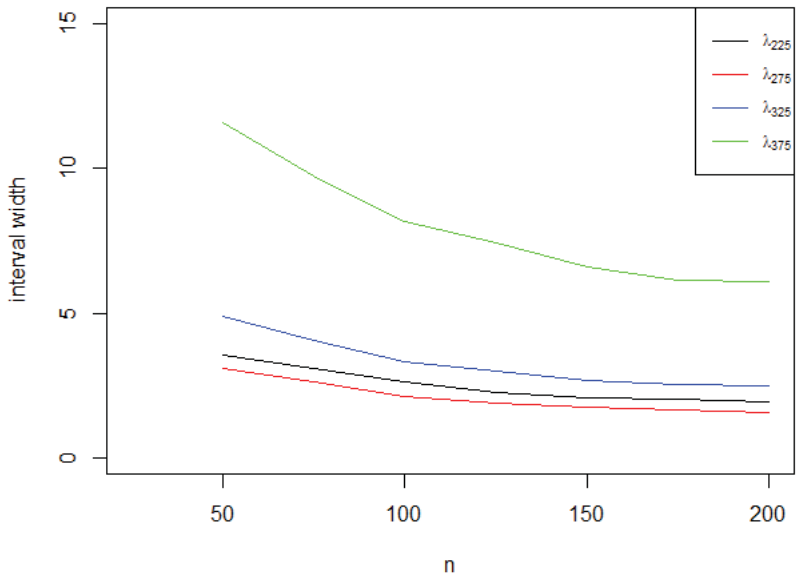

Figure 4.5: Median Interval Widths Over 200 Replications for λ_i at Various Weights with $x_i \sim Beta_{[200,400]}(1.7, 3)$.

To determine whether our beliefs concerning the effect of the covariate design prior on sample size are justified, we redo the sample size simulation, replacing the current covariate design prior with the alternative that is skewed towards the larger weights. As before, we look at the resulting posterior samples for λ_i at various weights for four select instances of the 200 replications at sample size $n = 100$, shown in Figure 4.6. Again, the solid line represents the posterior median and the dashed bounds represent 95% credible intervals. This time, we see that the interval widths narrow around 350 lbs., but there is no significant difference in widths at the ends of the weight scale.

We also record 95% interval lengths of λ_i at various weights across all 200 replications fo a range of sample sizes, with their coverage shown in Table 4.3. The lowest coverage here is 0.935. The density of the 200 interval lengths, at sample size $n = 100$ of λ_i for a subjet weighing 225 lbs. is shown Figure 4.7. Thus, we continue to record median interval lengths.

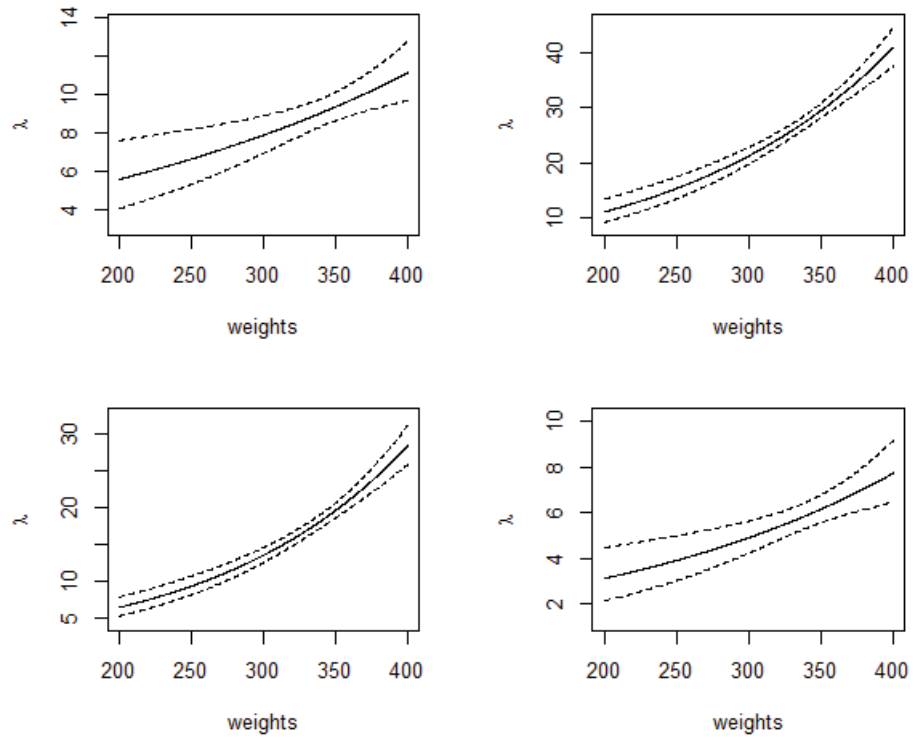

Figure 4.6: Posterior Samples for λ_i in Four Instances of the 200 Replications for $n = 100$ with $x_i \sim \beta(3, 1.7)$.

Table 4.3: Coverage for λ_i at Various Weights Across the 200 Intervals with $x_i \sim \beta(3, 1.7)$

n	225 lbs.	275 lbs.	325 lbs.	375 lbs.
50	0.960	0.960	0.950	0.955
75	0.970	0.970	0.960	0.965
100	0.955	0.955	0.935	0.960
125	0.970	0.970	0.960	0.940
150	0.970	0.965	0.965	0.985
175	0.950	0.945	0.935	0.960
200	0.945	0.965	0.965	0.945

Figure 4.8 plots the median interval widths across the sample sizes for λ_i at $x_i = 225, 275, 325$, and 375 for this case. We see that λ_i at 375 lbs. again results in the largest interval widths, though not as large as before, over all the sample sizes, followed by λ_i at 225, 275, and 325 lbs. In this case, we know from the design prior that the majority of the generated weights should be between 325 and 375 lbs. with significantly fewer weights on the lower end of the scale. However, the results still

show that λ_i at 375 lbs. results in the largest interval widths at each sample size. We suspect that this result is due to the exponential affect of the transformation.

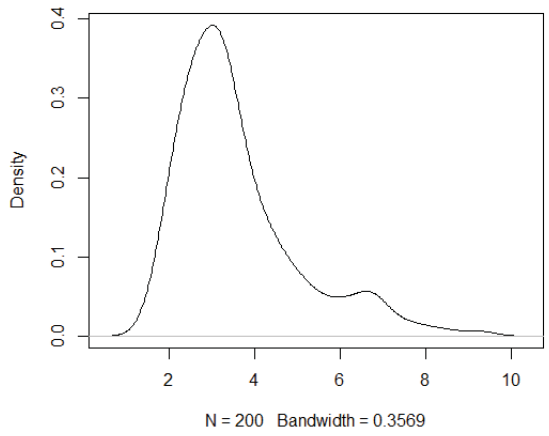

Figure 4.7: Density of 200 interval lengths for λ_i at 225 lbs. with with $x_i \sim \beta(3, 1.7)$.

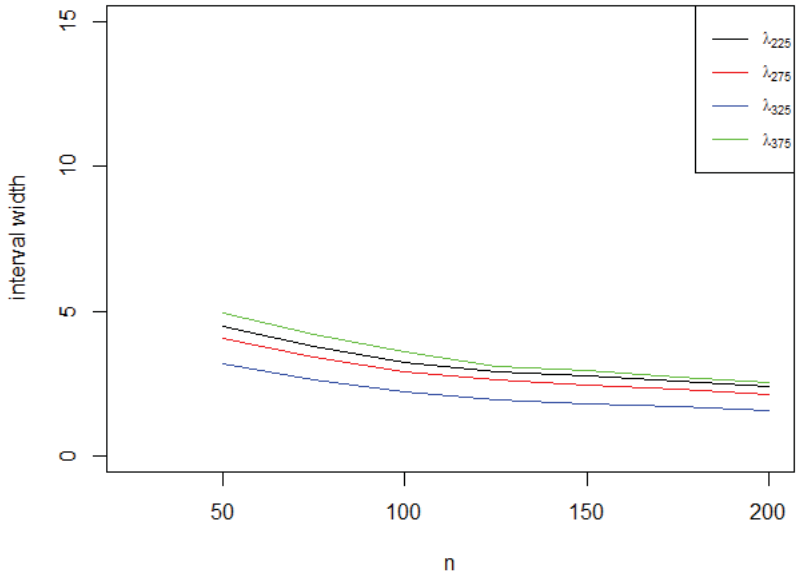

Figure 4.8: Median Interval Widths Over 200 Replications for λ_i at Various Weights with $x_i \sim Beta_{[200,400]}(3, 1.7)$.

Consider instead that the parameter of interest is the log transform of the expected number of sleep interruptions, or $\log(\lambda_i) = \gamma_0 + \gamma_1 x_i$. Figure 4.9 shows interval width curves for $\log(\lambda_i)$ at the same weights as before. Here the curve for $\log(\lambda_i)$ at 375 lbs. shows smaller interval widths, confirming the suspicion that the exponential transform is causing the inflated widths seen in Figure 4.8.

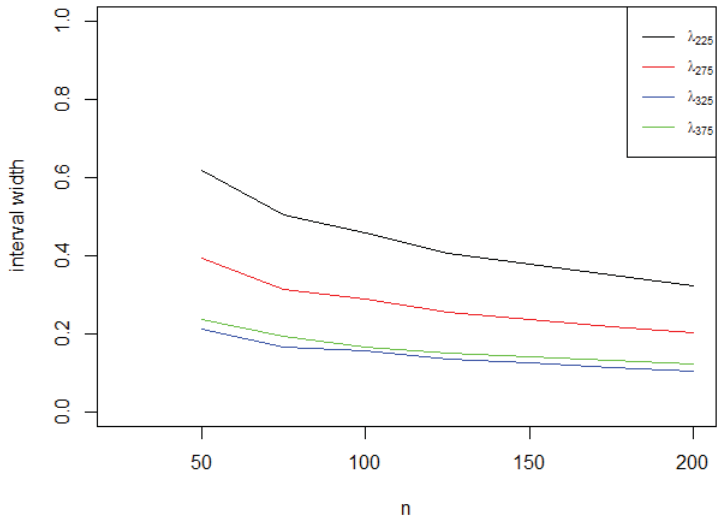

Figure 4.9: Median Interval Widths Over 200 Replications for $\log(\lambda_i)$ at Various Weights with $x_i \sim Beta_{[200,400]}(3, 1.7)$.

Suppose we repeat the simulation again, this time generating data sets and fitting them with unrealistically diffuse analysis priors. Figure 4.10 shows the resulting median interval width curves for λ_i at four weights of interest under both sets of analysis priors in the case where the design prior for the weight covariate is skewed towards the lower weights. Similarly Figure 4.11 shows the same for the case where the covariate design prior is skewed towards the higher weights. In both cases, it appears that there is little difference in results between the two sets of analysis priors, with the most noticeable differences occurring at the smaller sample sizes. It is possible that what we are observing is, as sample size increases, the increasing number of observations overwhelms the effective sample size of our chosen "realistically" diffuse analysis priors. Thus, to really see a clear difference, we must be less conservative in defining these priors.

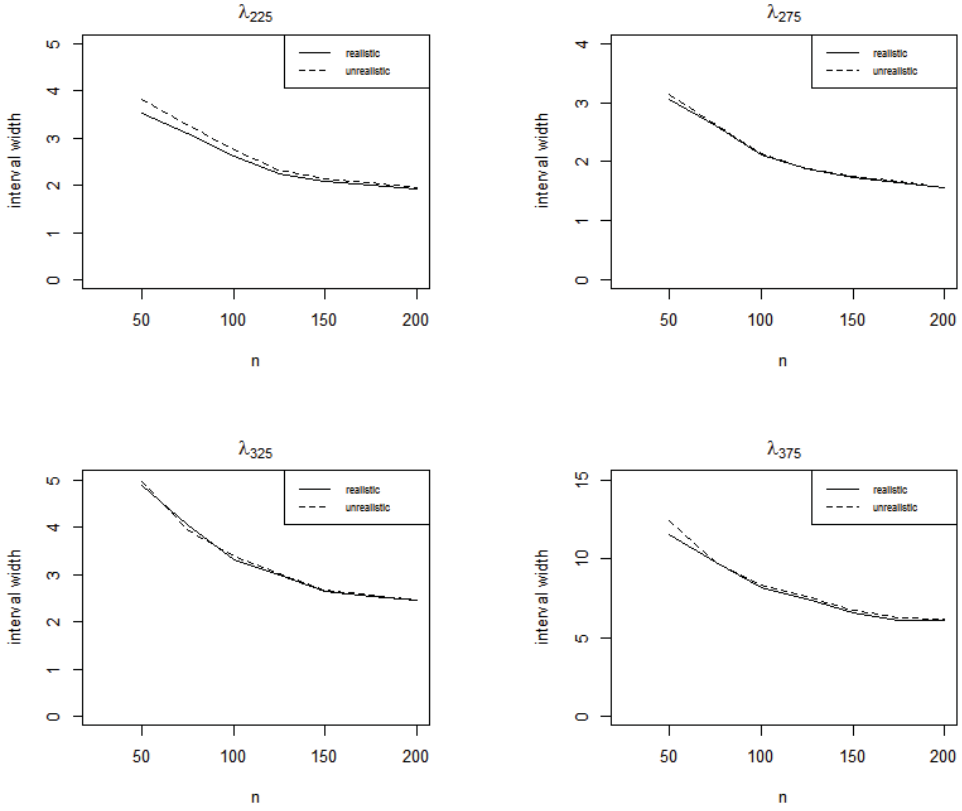

Figure 4.10: Realistic v. Unrealistic Analysis Priors for $x_i \sim \beta(1.7, 3)$.

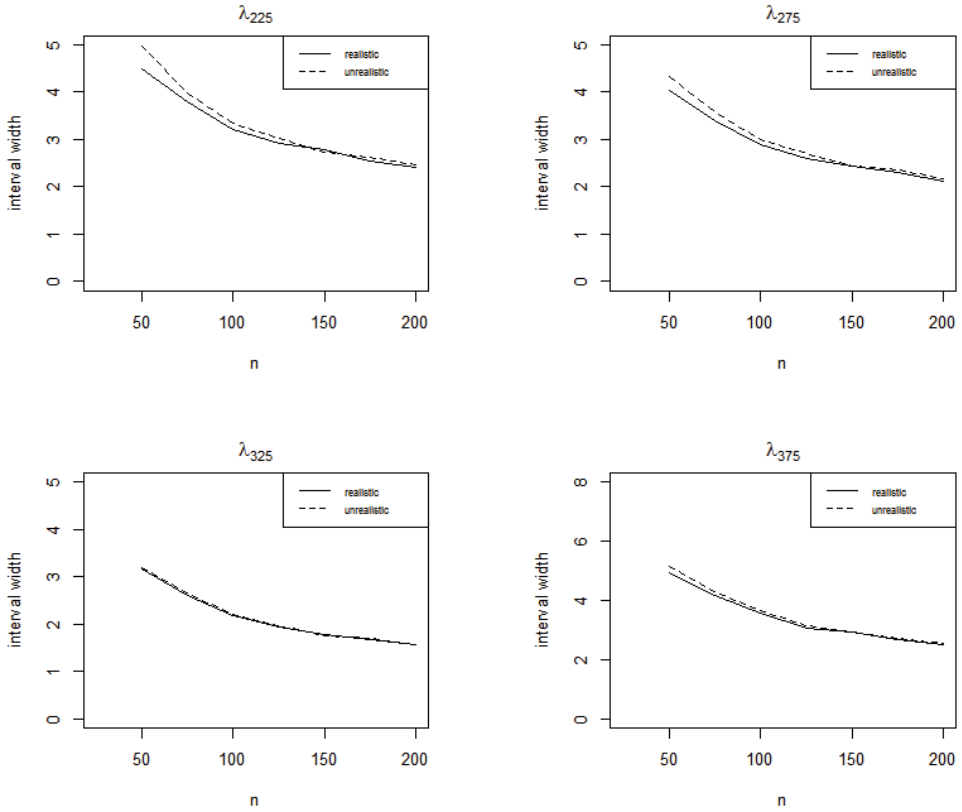

Figure 4.11: Realistic v. Unrealistic Analysis Priors for $x_i \sim \beta(3, 1.7)$.

4.4.3 Interpretation

The goal of the these simulations is, of course, to use these interval width curves to make sample size recommendations to researchers designing this sleep disorder study. The researcher will likely specify a required interval length for the parameter of interest. In our case, the parameter of interest is the expected number of sleep disturbances for an afflicted subject, λ_i, at various weights. Thus, the required interval length may change depending on the subject's weight. For example, the researcher may believe the range of expected number of disturbances for a 225 lb. subject to be much smaller than that of a 375 lb. subject, and that has a bearing on interval length requirements.

Consider the simulation where the design prior for weight, x_i, was skewed towards the lower weights, whose results are shown in Figure 4.5. Suppose the researcher requires a length of 2 for λ_i for subjects weighing 225 lbs. In Figure 4.12, we reproduce the interval width curve for λ_i at 225 lbs. The horizontal dashed line is at an interval width of 2. Thus, the appropriate sample size for this scenario occurs where the curve intersects with the horizontal line, which appears to happen around $n = 180$. Note that the final step of the algorithm is to fit a functional curve to the one shown in Figure 4.12, in which case one could solve more definitively for n. This same process can be repeated given any requirement for any parameter of interest.

We have described how to determine the appropriate sample size for a sleep disorder study in which the parameter of interest is the expected number of sleep interruptions, λ_i, suffered by afflicted subjects of a particular weight. In practice, it is unlikely that researchers would want to design a study in which they are only specifically interested in subjects of one weight. However, they may be particularly interested in subjects of a certain weight group. For example, perhaps they are most interested in subjects weighing between 300 and 350 lbs. One way to use this method to determine the required sample size in this case would be to consider various weights within this interval. As demonstrated, each weight will produce a curve via the sample

size determination algorithm, leading to a sample size recommendation. We propose that researchers can simply use the maximum of these sample sizes for their study.

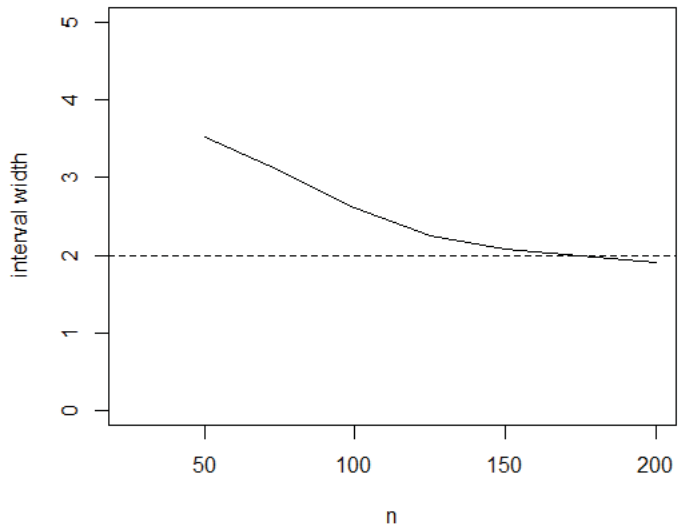

Figure 4.12: Median Interval Widths Over 200 Replications for λ_i at 225 lbs. with $x_i \sim \beta(1.7, 3)$.

4.5 Discussion

In this chapter, we considered the sample size determination problem for the proposed sleep disorder study discussed throughout the dissertation. While sample size issues for zero-inflated data have been discussed in the literature, the majority of it is based in the frequentist approach. We discussed the "two-priors" Bayesian approach to sample size determination and the advantages it has. Then, we described how this method could be applied to hurdle models and performed the simulation to determine sample size requirements for the sleep disorder study. We found that sample size requirements in this problem are dependent on specific parameters of interest, as well as the nature of the population from which the sample is taken with respect to covariate design priors.

We also found that specifying realistically diffuse analysis priors had little effect over using exceedingly diffuse ones. We hypothesize that the reason for this may be that the effective sample size of the realistically diffuse analysis priors was over-

whelmed by the sample size of the observations. Thus, one possible solution would be to consider even less conservative analysis priors.

In the traditional Bayesian two-priors approach to sample size determination, diffuse analysis priors are typically used. However, the results we found lead us to believe that, in order for there to be a significant reduction in sample size, these analysis priors need to incorporate more information. The difficulty in doing so is that, in the simulation algorithm, the "true" parameter values change change at each iteration. For future work it would be useful to explore how the coupled prior structure described in Chapter Three can be adapted into informative analysis priors with potentially larger effective sample sizes than the ones considered in this analysis.

The effective sample size of these priors can be explored using a method developed by Morita et al. (2008, 2012). The idea is that the total necessary sample size can potentially be reduced by the effective sample size of the more informative analysis prior. Better understanding of this relationship would prove extremely useful in the design of experiments overall.

APPENDIX

APPENDIX A

Data Generation for Sleep Study Example

The process we use to generate outcomes for the hypothetical sleep disorder study described in Section 2.1.4 is described here. We begin by discussing how zero-inflated outcomes appropriate for a hurdle model can be generated. Recall that, in the Bernoulli-Poisson hurdle model, the parameter θ represents the probability that the subjects will cross the hurdle. The `rbinom` function in R generates a specified number of random values from a specified binomial distribution. Since the Bernoulli distribution is essentially binomial with one fixed trial, we can use `rbinom`, with specified parameters n (sample size) and θ, to generate 0's and 1's, which determine whether or not subjects cross the hurdle. Once we determine the subjects who cross the hurdle, we want to generate positive counts for them from a truncated count distribution, which for our purposes is the Poisson. The **VGAM** package (Yee, 2014) in R has a function `rpospois` for generating random values from a truncated Poisson distribution. Thus, for all subjects assigned a 1 by `rbinom`, we use `rpospois`, with specified parameters n and λ, to generate corresponding positive counts.

In the sleep disorder example, we suppose that weight acts as a shared covariate on both parts of the model. Assume that the overweight men all weigh between 200 and 400 lbs., with slightly more weighing closer to 200 than 400 lbs. To reflect this, we generate the subjects' weights from a shifted $Beta_{[200,400]}(1.7, 3)$ distribution, rounded to the nearest whole pound. Note that when these weights are used for data generation, we consider them in 20 lb. units. This was an arbitrary decision made for interpretability reasons as well as to help combat autocorrelation issues we saw preliminarily in fitting the Bayesian model.

With the added covariate on both parts of the model in the sleep disorder context, the described generation process is modified slightly. Instead of constant

values of θ and λ, these values change depending on the subject's weight, where

$$\theta_i = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}$$

and

$$\lambda_i = \exp(\gamma_0 + \gamma_1 x_i).$$

To best assess the proposed prior structures, we want the generated data to accurately reflect the expert opinion shown in Figures 3.3 and 3.4. We consider a simple way to do this, working backwards and employing least squares methods. Recall that, in the model,

$$\text{logit}(\theta_i) = \beta_0 + \beta_1 x_i.$$

Using R, we get the least squares linear fit based on the logit transform of the elicited modal values of θ_i shown in Figure 3.3(a), finding that $\beta_0 = -4.18$ and $\beta_1 = 0.28$. Then, for the positive counts, we transform the θ_i values to λ_i values using the elicited the functional relationship (3.2) with fixed values $u = 12.72$ and $v = 0.48$

$$\lambda_i = f(\theta_i) = 12.72 \exp(0.48 \cdot \text{logit}(\theta_\text{i}))$$

so that the data reflects this relationship. Under the reparameterization (3.3), this is equivalent to having $\gamma_0 = 0.54$ and $\gamma_1 = 0.13$.

BIBLIOGRAPHY

Adcock, C. J. (1997), "Sample size determination: A review," *The Statistician*, 46, 261–83.

Aldstadt, J., Koenraadt, C. J. M., Fansiri, T., Kijchalao, U., Richardson, J., Jones, J. W., and Scott, T. W. (2011), "Ecological Modeling of Aedes aegypti (L.) Pupal Production in Rural Kamphaeng Phet, Thailand," *PLOS Neglected Tropical Diseases*, 5, e940.

Bedrick, E. J., Christensen, R., and Johnson, W. (1996), "A New Perspective on Priors for Generalized Linear Models," *Journal of the American Statistical Association*, 91, 1450–1460.

Brutti, P., De Santis, F., and Gubbiotti, S. (2008), "Robust Bayesian sample size determination in clinical trials," *Statistics in Medicine*, 27, 2290–2306.

Cameron, A. C. and Trivedi, P. K. (1998), *Regression Analysis of Count Data*, Cambridge University Press.

Channouf, N., Fredette, M., and MacGibbon, B. (2014), "Power and sample size calculations for Poisson and zero-inflated Poisson regression models," *Computational Statistics and Data Analysis*, 72, 241–251.

Congdon, P. (2005), *Bayesian Models for Categorical Data*, Wiley.

Cragg, J. (1971), "Some statistical models for limited dependent variables with application to the demand for durable goods," *Econometrica*, 39, 829–44.

Dunson, D. B. (2005), "Bayesian Semiparametric Isotonic Regression for Count Data," *Journal of the American Statistical Association*, 100, 618–27.

FDA (2010), *Guidance for the use of Bayesian statistics in medical device clinical trials*.

Garthwaite, P. H., Kadane, J. B., and O'Hagan, A. (2005), "Statistical Methods for Elicitiing Probability Distributions," *Journal of the American Statistical Association*, 100, 680–701.

Heckman, J. (1979), "Sample selection bias as a specification error," *Econometrica*, 47, 153–61.

Heilbron, D. (1989), "Generalized linear models for altered zero probabilities and over dispersion in count data," *SIMS Technical Report 9, Department of Epidemiology and Biostatistics, University of California, San Francisco*.

— (1994), "Zero-altered and other regression models for count data with added zeros," *Biometrical Journal*, 36, 531–547.

Hilbe, J. M. (2011), *Negative Binomial Regression*, Cambridge University Press, 2nd ed.

Humphreys, B. R. (2013), "Dealing With Zeros in Economic Data," `http://www.ualberta.ca/~bhumphre/class/zeros_v1.pdf`.

Jackman, S. (2014), *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory, Stanford University*, Department of Political Science, Stanford University, Stanford, California, r package version 1.4.6.

Jalava, K., Ollgren, J., Eklund, M., Siitonen, A., and Kuus, M. (2011), "Agricultural, socioeconomic and environmental variables as risks for human verotoxigenic Escherichia coli (VTEC) infection in Finland," *BMC Infectious Diseases*, 11.

Joseph, L., du Berger, R., and Belisle, P. (1997), "Bayesian and mixed Bayesian/likelihood criteria for sample size determination," *Statistics in Medicine*, 16, 769–81.

Kahle, D., Stamey, J., and Seaman, J. (2014), *glmcmp: A package for prior elicitation in generalized linear models*.

Kuhnert, P. M., Martin, T. G., Mengersen, K., and Possingham, H. P. (2005), "Assessing the impacts of grazing levels on bird density in woodland habitat: a Bayesian approach using expert opinion," *Environmentrics*, 16, 717–747.

Lachenbruch, P. A. (2001), "Power and sample size requirements for two-part models," *Statistics in Medicine*, 20, 1235–1238.

Lambert, D. (1992), "Zero-inflated Poisson regression, with an application to defects in manufacturing," *Technometrics*, 34, 1–14.

Levin, K. A., Davies, C. A., Douglas, G. V. A., and Pitte, N. B. (2010), "Urban-rural differences in dental caries of 5-year old children in Scotland," *Social Science and Medicine*, 71, 2020–2027.

Levin, K. A., Davies, C. A., Topping, G. V. A., Assaf, A. V., and Pitts, N. B. (2009), "Inequalities in dental caries of 5-year-old children in Scotland, 19932003," *European Journal of Public Health*, 19, 337–42.

Martin, T. G., Kuhnert, P. M., Mengersen, K., and Possingham, H. P. (2005a), "The Power of Expert opinion in Ecological Models Using Bayesian Methods: Impact of Grazing on Birds," *Ecological Applications*, 15, 266–280.

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., and Possingham, H. P. (2005b), "Zero tolerance ecology: improving ecological inference by modelling the source of zero observations," *Ecology Letters*, 8, 1235–1246.

Morita, S., Thall, P. F., and Mller, P. (2008), "Determining the Effective Sample Size of a Parametric Prior," *Biometrics*, 64, 595–602.

— (2012), "Prior Effective Sample Size in Conditionally Independent Hierarchical Models," *Bayesian Analysis*, 7, 591–614.

Mullahy, J. (1986), "Specification and testing of some modified count data models," *Journal of Econometrics*, 33, 341–65.

Neelon, B., Chang, H., and Hastings, S. (2014), "Spatiotemporal hurdle models for zero-inflated count data: exploring trends in emergency department visits," *Statistical Methods in Medical Research*.

Neelon, B., Ghosh, P., and Loebs, P. F. (2013), "A spatial Poisson hurdle model for exploring geographic variation in emergency department visits," *Journal of the Royal Statistical Society*, 176, 380–413.

Neelon, B. H., O'Malley, A. J., and Normand, S. T. (2010), "A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use," *Statistical Modelling*, 10, 421–39.

Ridout, M., Demetrio, C. G. B., and Hinde, J. (1998), "Models for count data with many zeros," *Proceedings of the XIXth International Biometric Conference*, 19, 179–192.

Rose, C. E., Martin, S. W., Wannemuehler, K. A., and Plikaytis, B. D. (2006), "On the Use of Zero-Inflated and Hurdle Models for Modeling Vaccine Adverse Event Count Data," *Journal of Biopharmaceutical Statistics*, 16, 463–481.

Shieh, G. (2001), "Sample size calculations for logistic and Poisson regression models," *Biometrika*, 88, 1193–1199.

Stamey, J. D., Natanegara, F., and Seaman, J. W. (2013), "Bayesian Sample Size Determination for a Clinical Trial with Correlated Continuous and Binary Outcomes," *Journal of Biopharmaceutical Statistics*, 23, 790–803.

Tutz, G. (2012), *Regression for Categorical Data*, Cambridge University Press.

Wang, F. and Gelfand, A. (2002), "A simulation-based approach to Bayesian sample size determination for performance under a given model and for separating models," *Statistical Science*, 17, 193–208.

Williamson, J. M., Lin, H.-M., Lyles, R. H., and Hightower, A. W. (2007), "Power Calculations for ZIP and ZINB Models," *Journal of Data Science*, 5, 519–34.

Yee, T. W. (2014), *VGAM: Vector Generalized Linear and Additive Models*, r package version 0.9-6.

Zeileis, A., Kleiber, C., and Jackman, S. (2008), "Regression Models for Count Data in R," *Journal of Statistical Software*, 27.