ABSTRACT

A Bivariate Regression Model with Correlated Mixed Responses

Ross A. Bray, Ph.D.

Chairpersons: John W. Seaman, Jr., Ph.D., and James D. Stamey, Ph.D.

In the dissertation we consider a bivariate model for associated binary and continuous responses such as those in a clinical trial where both safety and efficacy are observed. We designate a marginal and conditional model that allows for the association between the responses by including the marginal response as an additional predictor of the conditional response. We use a Bayesian approach to model the bivariate regression model using a hierarchical prior structure. Simulation studies indicate that the model provides good point and interval estimates of regression parameters across a variety of parameter configurations, with smaller binary event probabilities offering particular challenges. For example, as the probability of an adverse event decreases, we find that the marginal posterior variances increase for the binary safety response regression coefficients, but not for the conditional efficacy response coefficients. Potential problems with induced priors are briefly considered. We implement an asymptotic higher order approximation in order to obtain parameter estimates and confidence intervals via a simulation study. In comparison, the frequentist intervals are slightly more narrow than the Bayesian intervals (using vague priors), but the latter have far superior coverage. Finally, we implement a Bayesian sample size determination method while controlling an operating characteristic of the model, the family-wise error rate. We find that there is a savings in

power afforded by use of the multiplicity adjustment when simultaneously testing multiple hypotheses. Simulation results indicate that multiplicity adjustments improve the power of the model when compared to the overly conservative Bonferroni adjustment. We also see an improvement in power through the effective use of prior information.

A Bivariate Regression Model with Correlated Mixed Responses

by

Ross A. Bray, B.S., 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

_____

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

_____

James D. Stamey, Ph.D., Co-Chairperson

_____

Jack D. Tubbs, Ph.D.

_____

Renée Umstattd Meyer, Ph.D.

_____

Dean M. Young, Ph.D.

_____

Accepted by the Graduate School
May 2013

_____

J. Larry Lyon, Ph.D., Dean

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

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First of all I would like to thank God for blessing me with the ability and perseverance to make it through all the challenges that have come along in pursuit of this goal.

I would like to thank my advisor, Dr. John Seaman, for all of your patient advice and insight while guiding me through this dissertation. Thank you for being available any time I had a question and for constantly encouraging me as I inevitably made a mess of my first attempt at each problem.

Thank you, Dr. James Stamey for answering my questions, even when the timing wasn't convenient. Thank you also to Dr. Dean Young, for your constant encouragement when I wasn't sure I should even be here. Thank you to Dr. Tom Bratcher (1942 – 2012). Your passion for your students and for statistics was inspiring. Thank you also to the faculty, staff and students of the statistics department. I have enjoyed being welcomed into a community in times of celebration, collaboration and even sorrow.

I want to thank my parents Doug and Tammy Bray, for your love and support. Thank you for raising me to give my full effort when I have committed to something and to not consider quitting as an option.

Finally, I would like to thank my wife Mandy. You have always been there, even as I doubt myself, and your unending love and support have carried me through the most difficult times of this process. I don't deserve to have someone like you in my life.

DEDICATION

To Mandy, My Love

CHAPTER ONE

Introduction

Correlation between responses is common in many areas of statistics. One place this is seen is in the trade-off between efficacy and safety in the context of a clinical trial. When both responses are binary Thall and Cook (2004) discuss the efficacy and safety trade-off in the context of a dose-finding study using Bayesian methods and Conaway and Petroni (1996) mention various designs for clinical trials that allow for this trade-off. In addition, Thall and Cheng (1999) propose a two-dimensional treatment comparison when both efficacy and safety are real valued.

In this dissertation, we are not interested in the case when the responses are similar, both discrete or continuous, but rather when they are mixed outcome types. McCulloch (2008) proposes joint modeling of mixed outcome types using a latent random effects model. Alternatively, Fitzmaurice and Laird (1995) introduce a bivariate regression based on a marginal and conditional formulation to model the mixed outcomes. They incorporate an association between the responses by including the marginal response as an additional predictor of the conditional response. We will focus on the latter approach.

An important question in bivariate modeling is whether incorporating joint outcomes indicates an improvement over modeling each outcome independently. The principle of parsimony, or Ockham's razor, tells us that the simpler model is better if all else is equal. Stamey et al. (2013) develop a Bayesian sample size determination on the model proposed by Fitzmaurice and Laird (1995). Their results indicate that, assuming there is some correlation between the mixed outcomes, there is a savings in sample size by implementing the bivariate model, rather than the independence model.

The dissertation is organized as follows. In Chapter Two we introduce a Bayesian approach to the bivariate regression model from Fitzmaurice and Laird (1995). We utilize an hierarchical prior structure in the context of an hypothetical weight loss scenario. We investigate the performance of our model in the weight loss scenario through a simulation study and discuss our recommendations for using the model.

In Chapter Three we detail an asymptotic frequentist approach to inference for this model called the r^* method. This method is intended to provide improved approximations over those utilizing Wald and score statistics. The r^* method has its basis in normal theory and utilizes an adjustment to the likelihood root statistic. We continue to use the weight loss example from Chapter Two and perform a simulation study to analyze the use of the r^* method on the bivariate regression model. We conclude Chapter Three with a comparison of the Bayesian approach to the r^* method.

In Chapter Four we extend the Bayesian sample size determination from Stamey et al. (2013). We are interested in exploring an operating characteristic of the Bayesian model, the family-wise error rate (FWER), in testing multiple endpoints concurrently. We propose that utilizing multiplicity adjustments to control the FWER will improve the power of the model. We also discuss the merits of using frequentist methods such as multiplicity adjustments in a Bayesian context. We look at two multiplicity adjustment methods, the fixed-sequence and fallback procedures, and use a Bayesian sample size simulation to compare their results to the overly conservative Bonferroni adjustment method. We conclude the dissertation with a discussion in Chapter Five.

CHAPTER TWO

Bayesian Application of a Bivariate Mixed Response Model

2.1 Introduction

Bivariate responses are common in many areas of research. For example, a computer manufacturer may wish to model processor speed and temperature as a function of programming application (image editing vs data streaming, for example). In drug development, a researcher may need to consider both efficacy and safety as a function of treatment choice, gender, and other covariates. In general, the responses may be both continuous, both discrete, or one of each. We are concerned with the latter combination.

A general bivariate data model can, of course, be written in marginal and conditional form:

$$\begin{aligned} Y_1 &\sim \mathcal{D}_1(\mu_1(\mathbf{x}), \mathbf{V}_1) \\ Y_2|Y_1 &\sim \mathcal{D}_2(\mu_2(\mathbf{x}), \mathbf{V}_2), \end{aligned} \tag{2.1}$$

where $\mathcal{D}_i(\mu_i(\mathbf{x}), \mathbf{V}_i)$ is a specified distribution with mean function μ_i, covariate vector $\mathbf{x}$, and variance function $\mathbf{V}_i, i = 1, 2$. In this dissertation, we focus on the situation where one response is discrete and the other is continuous. Either can be chosen as Y_1 although, especially in the binary case, it is often convenient to condition on the discrete random variable.

Olkin and Tate (1961) proposed one of the earliest methods for analyzing mixed discrete and continuous outcomes with a general "location model" which has a multinomial distribution for discrete outcomes and multivariate normal distribution for the continuous outcome. In their case $\mathcal{D}_1$ is the discrete outcome and $\mathcal{D}_2$ is the continuous outcome. Lauritzen and Wermuth (1989) provide another example, using graphical models. In contrast, Cox (1972) takes $\mathcal{D}_1$ to be normal and the conditional

distribution of the binary response given the continuous response is Bernoulli with a logistic link function.

Of particular interest to us is the case of a binary random variable, Y_1, and a continuous random variable, Y_2, conditional on Y_1. In the biopharmaceutical context, this might take the form of a binary safety variable and a continuous efficacy variable. This type of model has been considered, for example, by Fitzmaurice and Laird (1995), Gueorguieva and Agresti (2001), and Teixeira-Pinto and Normand (2009). The model used by Fitzmaurice and Laird (1995) takes $\mathcal{D}_1$ to be Bernoulli and $\mathcal{D}_2$ to be normal. They fit the resulting model using maximum likelihood. We offer a Bayesian approach in this chapter. In Chapter Three, we provide frequentist interval estimates using the r^* method of Brazzale et al. (2007) to extend the results in Fitzmaurice and Laird (1995). Stamey et al. (2013) implement a Bayesian sample size determination on the model used by Fitzmaurice and Laird (1995). In Chapter Four, we extend their approach in an attempt to improve operating characteristics of our Bayesian model.

McCulloch (2008) looked at joint modeling of mixed outcomes but approached it as a latent random effects model. Wei (2012) extended the model by McCulloch (2008) using the Bayesian approach while also incorporating misclassification and zero inflation. McCulloch (2008) is concerned with being able to model the correlation between the mixed outcome types, while Fitzmaurice and Laird (1995) treat the association between the responses as a nuisance parameter. A direct comparison of these two models has not been done.

Our motivating problem stems from Robertson and Allison (2009) who study the effects of various drugs on suicidal ideations and suicide attempts. We consider two safety situations where weight loss is our efficacy variable.

In this chapter we present our Bayesian model. We detail the basic data model, common to both the Bayesian and maximum likelihood approach, in Section 2.2. We

provide two hypothetical examples in Section 2.3. A simulation study is presented in Section 2.4 and we conclude with a discussion in Section 2.5.

2.2 Data Model

In this section we present the bivariate data model. Following Fitzmaurice and Laird (1995) we assume that we have both a binary and a continuous response for each subject. In what follows, we sometimes refer to the binary variable as the "safety" variable, and the continuous response as the "efficacy" variable, reflecting our chief motivation. Of course, the model is more generally applicable.

Let Y_{1i} be the binary response for the safety variable, where $Y_{1i} = 1$ for an adverse event and $Y_{1i} = 0$ otherwise, for subject i, $i = 1, ..., n$. Also let Y_{2i} be the continuous response of the efficacy variable, for subject i. We assume both Y_{1i} and Y_{2i} are dependent on covariate vector $\mathbf{x}_i \equiv (x_{1i}, x_{2i}, ..., x_{pi})'$. We also account for potential dependence between Y_{1i} and Y_{2i} by including Y_{1i} as an additional predictor of Y_{2i}. The model is given by

$$Y_{1i} \sim Bernoulli(\pi_i)$$

with

$$logit(\pi_i) = \gamma_0 + \gamma_1 x_{1i} + \sum_{j=2}^{p} \gamma_j x_{ji}, \tag{2.2}$$

$$logit(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right), \quad 0 < \pi_i < 1,$$

and

$$Y_{2i}|Y_{1i} \sim N(\mu_i, \sigma^2),$$

where

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \sum_{j=1}^{p} \beta_j x_{ji} + \lambda(y_{1i} - \pi_i). \tag{2.3}$$

Notice that λ is included in the model to allow for the relationship between Y_{1i} and Y_{2i}. Also notice that any number of covariates can be included additively in the link functions of the model.

The logit link in (2.2) is symmetric around zero with $\lim_{\pi \to 0} logit(\pi_i) = -\infty$ and $\lim_{\pi \to 1} logit(\pi_i) = \infty$. We will be mainly interested in π_i between 0 and 0.5. For this case the intercept coefficient, γ_0, will be negative and a positive covariate coefficient corresponds to a larger probability while a negative covariate coefficient corresponds to a smaller probability.

The effect of λ in the linear link (2.2) depends on whether the binary event of interest occurred. If so, then the continuous response mean, μ_i, is adjusted by $\lambda(1 - \pi_i) \in (-\lambda, \lambda)$. Note that, for rare adverse events, this adjustment will be approximately equal to λ in magnitude.

We derive the marginal expectations and variances of Y_{1i} and Y_{2i}, as well as the covariance between Y_{1i} and Y_{2i}. The marginal expectation of Y_{1i} is straightforward:

$$E[Y_{1i}] = \pi_i = \frac{\exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)}{1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)}.$$

The marginal expectation of Y_{2i} is

$$\begin{aligned} E[Y_{2i}] &= E[E(Y_{2i}|Y_{1i})] \\ &= \beta_0 + \beta_1 x_{1i} + \sum_{j=1}^{p} \beta_j x_{ji}. \end{aligned}$$

The marginal variance components are

$$\begin{aligned} Var[Y_{1i}] &= \pi_i(1 - \pi_i) \\ &= \frac{\exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)}{\left[1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)\right]^2}, \end{aligned}$$

and

$$\begin{aligned} Var[Y_{2i}] &= E[Var(Y_{2i}|Y_{1i})] + Var[E(Y_{2i}|Y_{1i})] \\ &= \sigma^2 + \frac{\lambda \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)}{\left[1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)\right]^2}. \end{aligned}$$

Finally, the covariance is

$$\begin{aligned} Cov[Y_{1i}, Y_{2i}] &= E[y_{1i}y_{2i}] - E[y_{1i}]E[y_{2i}] \\ &= \frac{\lambda \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)}{\left[1 + \exp\left(\gamma_0 + \gamma_1 x_{1i} + \sum_{j=1}^{p} \gamma_j x_{ji}\right)\right]^2}. \end{aligned}$$

Note that the covariance is a linear function of λ and not a function of the regression parameters for the continuous response.

2.3 Examples

2.3.1 Weight Loss/Suicide Attempt

Robertson and Allison (2009) discuss drugs that are associated with suicidal ideations and attempts. Among these is a treatment for obesity. They consider the number of suicidal 'ideations' that occurred while patients were on the drug and the corresponding number of suicide attempts. Using their study in an hypothetical context, we suppose that, in addition to suicidal ideation, percent body weight change compared to baseline is available as an efficacy response. Our safety endpoint is $y_{1i} = 1$ for a suicide ideation and 0 otherwise, and the efficacy endpoint, $y_{2i}|y_{1i} \in \mathbb{R}$, is continuous and represents the percent body weight change compared to baseline. A negative value of $y_{2i}|y_{1i}$ indicates a weight gain. Let n be the number of patients enrolled in the study and let age and gender also be recorded as covariates for each subject. Here, x_{1i} = age is continuous and x_{2i} = gender is binary, with 1 indicating male and 0 female. The data model is then

$$Y_{1i} \sim Bernoulli(\pi_i),$$

with

$$logit(\pi_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i},$$

and

$$Y_{2i}|Y_{1i} \sim N(\mu_i, \sigma^2),$$

where, for $Y_{1i} = y_{1i}$,

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \lambda(y_{1i} - \pi_i). \tag{2.4}$$

The likelihood is

$$L(y_{1i}, y_{2i}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \sigma) \propto \prod_{i=1}^{n} \pi_i^{y_{1i}}(1-\pi_i)^{1-y_{1i}}(2\pi\sigma^2)^{-1/2} \exp\left\{-\frac{1}{2\sigma^2}(y_{2i} - \mu_i)^2\right\}, \tag{2.5}$$

where $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)'$, $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2)'$,

$$\pi_i = \frac{\exp[\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}]}{1 + \exp[\gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i}]}, \tag{2.6}$$

and μ_i is defined in (2.4).

Prior information might be incorporated using conditional means priors, Zellner's g-priors, or other methods for prior construction in generalized linear models. See, for example, Robert (2007) or Christensen et al. (2011) for overviews of such methods. In the absence of prior information, diffuse priors on the regression coefficients can be constructed by using normal distributions, centered at zero, and with large variances. We take the latter approach, assigning $\beta_0 \sim N(0, 1000)$, $\beta_1, \beta_2 \sim N(0, 100)$, $\gamma_i \sim N(0, 100)$, $\lambda \sim N(0, 1000)$, and $\sigma \sim U(0, 100)$, where $i = 0, 1, 2$, and the joint prior is denoted $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \sigma)$. The specified hierarchical model is pictured in Figure 2.1. The upper bound on the uniform prior is chosen to be large enough to have little influence on posterior quantities of interest. We discuss this more thoroughly in Section 2.4.

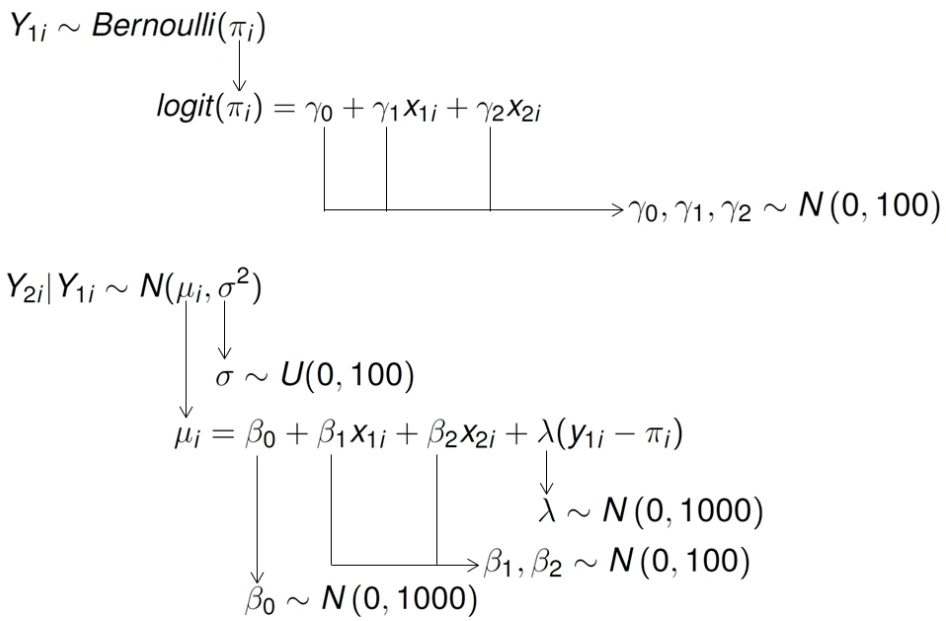

Figure 2.1: General Bayesian heirarchical model.

The joint posterior is

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \sigma | y_{1i}, y_{2i}) \propto \prod_{i=1}^{n} \left[\pi_i^{y_{1i}} (1-\pi_i)^{1-y_{1i}} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma^2} (y_{2i} - \mu_i)^2 \right\} \right]$$
$$\times p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \sigma). \tag{2.7}$$

No closed-form posteriors are available for our model, so we use Markov chain Monte Carlo (MCMC) methods to find the posterior distributions. We will use OpenBUGS (Thomas et al., 2006) to simulate from the posterior distributions of the parameters via the BRugs package (Thomas et al., 2006) in *R*.

We want to simulate a sample data set to illustrate the suicide ideation under obesity treatment scenario described above. We simulate age as $N(40, 16)$ to represent a wide range of ages and then transform so that x_{1i} is standard normal. We set gender as $x_{2i} \sim Bernoulli(0.5)$ with an equal probability for men and women. We want to target an average of $\pi_i = 0.02$ and $\mu_i = 10$ which yields parameter values $\boldsymbol{\beta} = (13, 1.67, -6)$ and $\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$. We set $\lambda = 10$ and $\sigma = 5$. With these values the marginal expectations are $E[Y_{1i}] \approx 0.017$ and $E[Y_{2i}] = 10$. Note that the negative value of β_2 corresponds to women losing more weight on average than men,

while the positive value of γ_2 results in a higher probability of an adverse event for men.

We generated 600 observations from the data model using R. The counts for y_1 are included in Table 2.1 and a histogram of y_2 is shown in Figure 2.2. Notice that the histogram appears fairly symmetric and centered around our target value of $\mu_i = 10$. We perform the MCMC analysis using OpenBUGS by running two chains with a burn-in of 5,000 iterations and then saving next 10,000 iterations. Thinning was set at 5 because chains that were not thinned exhibited autocorrelation. The initial values for the first chain were set at $\boldsymbol{\beta} = (-10, -2, -3)$, $\boldsymbol{\gamma} = (-10, -2, -4)$, $\lambda = -10$, and $\sigma = 1$. Initial values for the second chain were $\boldsymbol{\beta} = (10, 2, 3)$, $\boldsymbol{\gamma} = (10, 2, 4)$, $\lambda = 10$, and $\sigma = 7$.

Table 2.1: Frequencies of y_1 from the weight loss/suicide ideation generated data.

y_1	Frequency
Adverse event	19
No Adverse event	581

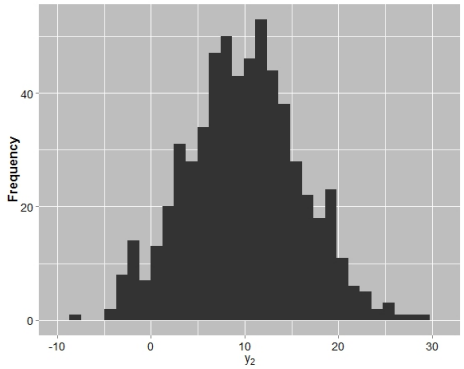

Figure 2.2: Histogram of y_2 for the weight loss/suicide ideation generated data.

To assess convergence, we consider autocorrelation plots, densities and Gelman-Rubin plots. Autocorrelation plots indicated no problems after thinning. In Figure

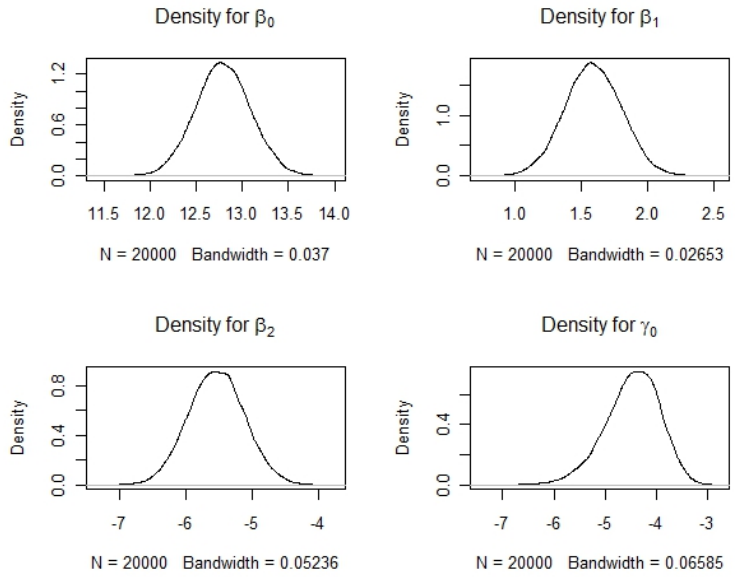

Figure 2.3: Density plots for β_0, β_1, β_2, and γ_0.

2.3, we see that the densities for β_0, β_1, β_2, and γ_0 appear smooth and unimodal. The densities for γ_1, γ_2, λ, and σ appear similar to Figure 2.3. Finally, the Gelman-Rubin plots all converge to one quickly and thus, the chains appear to converge.

The results of the MCMC simulation are included in Table 2.2. The eight parameters are aligned with their true value along with their posterior mean and the 95% credible interval. Notice that each of the credible intervals contain their respective true values. Estimation of γ_1 and γ_2 is inherently difficult given their small values and the fact that we observed "only" 19 adverse events out of a sample of 600. In fact, adverse events like suicide ideation would likely be even more rare, depending on the population.

2.3.2 Weight Loss/Systolic Blood Pressure

Consider the same weight loss scenario above where, rather than monitor suicidal ideations, we dichotomize measurements of systolic blood pressure where an increase of, say, 10 mm/Hg or greater is considered an adverse event. The binary response is now $y_{1i} = 1$ for an increase of at least 10 mm/Hg, and 0 otherwise. The

Table 2.2: Posterior means and 95% credible intervals for the weight loss/suicide ideations example.

Parameter	True	2.5%	Mean	97.5%
β_0	13.00	12.21	12.79	13.38
β_1	1.67	1.17	1.59	2.01
β_2	-6.00	-6.36	-5.54	-4.71
γ_0	-4.60	-5.64	-4.48	-3.55
γ_1	0.24	-0.33	0.26	0.84
γ_2	1.10	-0.32	0.87	2.18
λ	10.00	6.44	9.23	12.02
σ	5.00	4.72	5.00	5.29

continuous response is still weight loss and we will continue to have age and gender as our covariates. We use the prior distributions discussed in the weight loss/suicidal ideations example which results in the joint posterior given in (2.7).

The data simulation will be similar to the previous example. We keep the covariate distributions at $x_{1i} \sim N(0, 1)$, and $x_{2i} \sim Bernoulli(0.5)$. We select $\pi_i = 0.05$ and $\mu_i = 5$, so that an increase in systolic blood pressure is not considered as rare an event as a suicidal ideation. The resulting parameter values are $\beta_0 = 3$, $\beta_1 = 1.33$, $\beta_2 = 4$, $\gamma_0 = -2.44$, $\gamma_1 = 0.41$, and $\gamma_2 = -1.45$, and we set $\lambda = -5$ and $\sigma = 1$. The marginal expectations using these parameter values are $E[Y_{1i}] = 0.041$ and $E[Y_{2i}] = 5$.

The MCMC procedure was again performed using OpenBUGS via *R*. We saved 10,000 iterations with a burn-in of 5,000 for two chains with thinning set at 5, and initial values as in Section 2.3.1. Similar to the previous example, there were no convergence issues in the autocorrelation, density, or Gelman-Rubin plots.

The results of the MCMC simulation are included in Table 2.3. Again, the true value of the parameters aligned with their posterior means and 95% credible intervals. Notice that, with the exception of γ_2, the posterior means are fairly close to their true values.

Table 2.3: Posterior means and 95% credible intervals for the weight loss/systolic blood pressure example.

Parameter	True	2.5%	Mean	97.5%
β_0	3.00	2.82	3.01	3.19
β_1	1.33	1.22	1.32	1.43
β_2	4.00	3.75	3.97	4.20
γ_0	-2.44	-2.80	-2.38	-1.99
γ_1	0.41	0.12	0.42	0.71
γ_2	-1.45	-3.29	-2.11	-1.14
λ	-5.00	-5.05	-4.68	-4.31
σ	1.00	0.97	1.03	1.09

2.4 Simulation

We now assess how our model performs through a simulation study. We would like to explore the performance of the model over a wide variety of π_i and μ_i values. We continue to use the weight loss drug as the hypothetical basis for our simulation. For efficacy we will look at values of μ_i between 0 and 20, corresponding to body weight percentage lost between 0 and 20 percent compared to baseline. We set the range of π_i for the safety variable between 0 and 0.20. Larger values correspond to a safety variable with a higher probability of an adverse event such as headaches or skin irritation. We target three values each of π_i and μ_i. They are $\pi_i = 0.02, 0.05,$ and 0.20, and $\mu_i = 5, 10,$ and 20. We will also choose $\lambda = -5$ and 10 and $\sigma = 1$ and 5. Therefore, combining every possible combination of π_i, μ_i, λ, and σ gives us 36 possible design points.

The complete data model is

$$Y_{1i} \sim Bernoulli(\pi_i),$$

with

$$logit(\pi_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i},$$

and

$$Y_{2i}|Y_{1i} \sim N(\mu_i, \sigma^2),$$

where, for $Y_{1i} = y_{1i}$,

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \lambda(y_{1i} - \pi_i).$$

Let $x_{1i} \sim N(0,1)$ and $x_{2i} \sim Bernoulli(0.5)$ continue to be age and gender respectively.

By specifying values of π_i and μ_i, we can infer values of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$. Suppose we take $\pi_i = 0.02$. This means that, on average, individuals in this study have a probability of 0.02 of experiencing an adverse event. The logit of 0.02 is approximately -3.89. We must now choose how much of an effect age will have on the probability of an adverse event. We arbitrarily choose that the probability of an adverse event will fluctuate plus or minus two percent based on plus or minus three standard deviations of the standard normal distribution. Thus, for three standard deviations above the mean, we would have a probability of 0.04, and the logit of 0.04 is approximately -3.178. We can then calculate

$$\begin{aligned}
\gamma_1 &= \frac{logit(0.04) - logit(0.02)}{3} \\
&= \frac{-3.178 - (-3.89)}{3} \\
&\approx 0.24.
\end{aligned}$$

Note that this does not yield an exact zero probability for three standard deviations below the average age because of the asymmetric properties of the logit transformation. Next we must decide what effect gender will have on the probability of an adverse event. We arbitrarily choose that males, at the average age, will have a probability of 0.03 while females, at the average age, will have a probability of 0.01. The logit of 0.01 is approximately -4.6 and the logit of 0.03 is approximately -3.5.

Therefore,

$$\begin{aligned}\gamma_2 &= logit(0.03) - logit(0.01) \\ &= -3.5 - (-4.6) \\ &= 1.1.\end{aligned}$$

Finally, γ_0 is the probability of an adverse event for a female at the average age which, from before, yields $\gamma_0 = logit(0.01) \approx -4.6$.

Similarly, we can calculate values of $\boldsymbol{\beta}$ for a target value of μ_i. Suppose the baseline value of an average percent body weight lost is $\mu_i = 5$. We arbitrarily choose that the percent body weight lost will be plus or minus four percent for plus or minus three standard deviations of age. Thus, we see that $\beta_1 = 4/3 \approx 1.33$. Furthermore, we choose to have males at the average age lose seven percent of their body weight, while females at the average age lose 3 percent of their body weight. Therefore, $\beta_2 = 7 - 3 = 4$. Finally, β_0 corresponds to a female at the average age, and thus $\beta_0 = 3$.

Each target value of π_i corresponds to a different set of true $\boldsymbol{\gamma}$ values such that setting $\pi_i = 0.02$ yields $\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)'$, $\pi_i = 0.05$ yields $\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)'$, and $\pi_i = 0.20$ yields $\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)'$. Similarly, setting $\mu_i = 5$ gives $\boldsymbol{\beta} = (3, 1.33, 4)'$, $\mu_i = 10$ gives $\boldsymbol{\beta} = (13, 1.67, -6)'$, and $\mu_i = 20$ gives $\boldsymbol{\beta} = (16, 1, 8)'$. Each of the design points is listed in Table 2.4. We will refer to the design point with $\pi_i = 0.02$ and $\mu_i = 5$ as design point 1 (DP1). Then call the case when $\pi_i = 0.05$ and $\mu_i = 5$ design point 2 (DP2). Continuing from left to right in Table 2.4 and moving from top to bottom we have design points 3 through 9 (DP3 — DP9). Each of the nine design points will also be run with the four combinations of λ and σ discussed earlier.

For each of the simulations we ran 50 replications of *OpenBUGS* using *R*. The first 5,000 iterations were discarded as a burn-in and the next 10,000 iterations were

Table 2.4: The design points over which the simulation will be run.

Efficacy	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.20$
$\mu = 5$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (3, 1.33, 4)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (3, 1.33, 4)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (3, 1.33, 4)$
$\mu = 10$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (13, 1.67, -6)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (13, 1.67, -6)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (13, 1.67, -6)$
$\mu = 20$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (16, 1, 8)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (16, 1, 8)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (16, 1, 8)$

recorded with thinning set at 5. We checked a random 4% of each set of simulation replications for convergence and did not find any problems. We generated samples with $n = 600$ for each of the 50 replications and we continue to use the prior structure from the examples where $\beta_0, \lambda \sim N(0, 1000)$, $\beta_1, \beta_2, \gamma_0, \gamma_1, \gamma_2 \sim N(0, 100)$ and $\sigma \sim U(0, 100)$. Initial values were set at 0 for all parameters except $\sigma = 1$.

The upper bound of 100 in the prior on σ allows for any body weight change, from 0 to 100%. This is obviously conservative. Note that if we did have a bounded parameter space for σ we could calculate an exact upper bound for the uniform prior.

The sample size was chosen so that the 95% credible sets would not include zero for some values of the parameters. Stamey et al. (2013) give a detailed discussion of a Bayesian sample size determination for this model. They look at both the average length and power criterion for evaluating the model, while also looking at various cases of interest. One case is the situation where β_1 is the parameter of interest for some treatment covariate, x_1. They also look at sample size in relation to λ, the strength of the relationship between the binary and continuous components. Finally, they look at the case where β_1 and γ_1 are of interest for the binary treatment covariate, x_1. Sample size issues will also be considered in Chapter Four.

The results for DP2 with $\lambda = 10$ and $\sigma = 5$ are included in Table 2.5. This corresponds to our hypothetical weight loss drug example where an increase in sys-

Table 2.5: Simulation results for DP2 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	3.00	2.29	2.92	3.57	1.28	0.98
β_1	1.33	0.96	1.40	1.83	0.87	0.96
β_2	4.00	3.23	4.11	4.98	1.74	0.92
γ_0	-2.44	-2.96	-2.51	-2.09	0.88	0.98
γ_1	0.41	0.09	0.45	0.80	0.71	0.92
γ_2	-1.45	-2.64	-1.60	-0.70	1.94	0.96
λ	10.00	8.11	9.93	11.75	3.64	0.92
σ	5.00	4.69	4.97	5.26	0.56	0.96

tolic blood pressure is the safety variable of interest. This scenario assumes that the drug results in an average weight loss of 5% with a probability of an adverse event of 0.05. Table 2.5 shows the true value of each parameter, the average of the posterior means and 95% credible intervals, as well as the average width and coverage of the intervals. Notice that the average of the posterior means is within, at most, 0.15, for all parameters. The lowest coverage for this simulation is 92% and this is representative of most of the simulations. We do have a few simulations where the coverage drops below 90% for one or two parameters with the lowest coverage occurring at 84%.

Figure 2.4 shows box plots of the four combinations of λ and σ for DP2 for β_0. The horizontal center line is the true value of $\beta_0 = 3$. The center point for each box plot is the average of the posterior means for the 50 replications of that simulation. The upper and lower limits of the box plots are the average of the upper and lower bounds, respectively, on the 95% posterior credible sets for the 50 replications. The three gray bars in each box plot are centered at the average posterior means and the average lower and upper bounds for the 95% credible intervals. The gray bars extend plus and minus one simulation standard deviation above and below their respective points on the box plot in order to give an estimate of the variability of

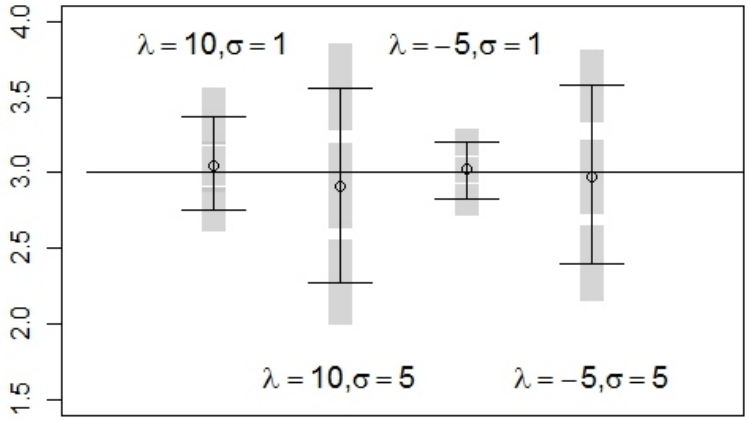

Figure 2.4: Spread of posterior estimates for β_0 in DP2.

the simulation. These bars occasionally overlap, and when this is the case we see a darker shade of gray.

There are several things to note in Figure 2.4. First, on average, the posteriors are fairly close to the true value of β_0. Also, the width of the intervals appears reasonable. These two trends follow in the estimation of the three efficacy regression parameters across all 36 of the simulations. Notice that the width of the 95% credible intervals is larger when σ is larger. This is also true for the simulation variability. Again, these observations follow across all simulations for all three efficacy regression parameters. Also notice a slight increase in the width of the credible sets when λ has a magnitude of 10, as opposed to when λ has a magnitude of 5. This also follows in the estimation of the efficacy regression parameters for each of the simulations, although the disparity is more difficult to see in β_2 and as π_i gets smaller.

Estimation for the three safety regression parameters is more complicated. Figures 2.5 and 2.6 show results that are similar to what we saw with the efficacy regression parameters. One obvious difference is that the width of the intervals for γ_0 are little changed when σ is increased from 1 to 5. Estimation of γ_2 is similar to γ_0 in this sense. The effect of σ on the estimation of γ_1 in Figure 2.6 is similar to its effect on the efficacy regression parameters but with smaller magnitude when our

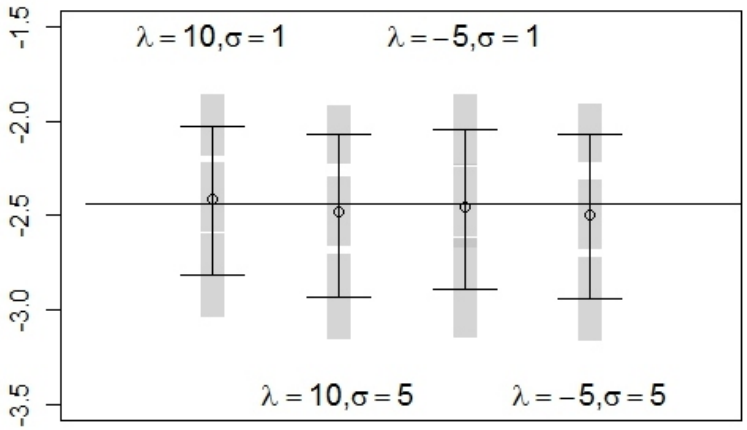

Figure 2.5: Spread of posterior estimates for γ_0 in DP2.

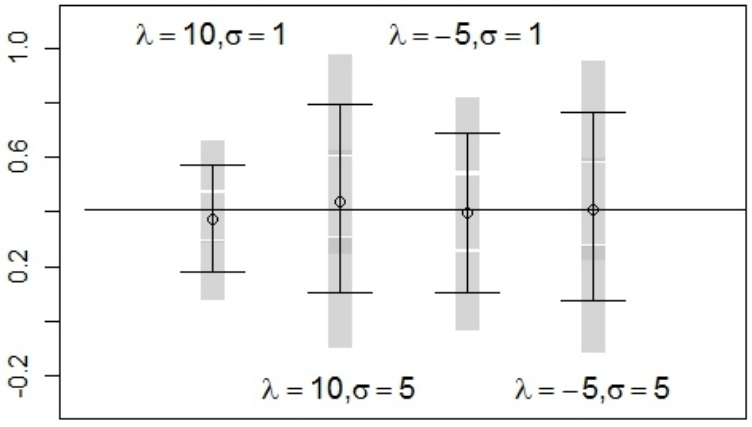

Figure 2.6: Spread of posterior estimates for γ_1 in DP2.

probability of an adverse event is larger. This shows up in design points 3, 6, and 9 where $\pi_i = 0.20$ and can be seen in Figure 2.7. This is the only observable difference between how the model estimates γ_1 and how it estimates the efficacy regression parameters.

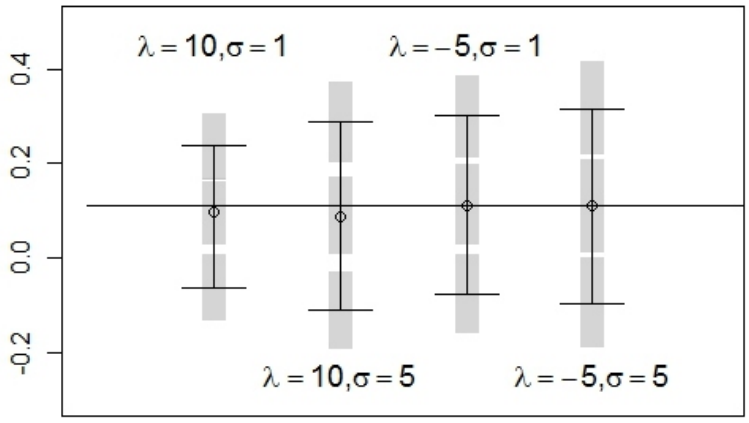

Figure 2.7: Spread of posterior estimates for γ_1 in DP3.

The model has a more difficult time estimating the parameters for the binary safety variable when the probability of an adverse event approaches 0, specifically in the estimation of γ_0 and γ_2. This is evident in our design points 1, 4, and 7. Figure 2.8 is similar to the previous figures, however in order to be readable we have moved the gray bar for the average of the lower bound of the 95% credible sets to the left of the box plots and moved the gray bar for the average of the upper bound of the 95% credible sets to the right of the box plot. Note that the simulation variation on the lower bounds is extremely high. This occurs because the small probability of an adverse event means we only have a few events out of our total sample of $n = 600$. When this happens the model tries to estimate an extremely small value for a probability of an adverse event which results in estimates of, mainly, γ_0 and γ_2 that are far from their true values. This only occurs in a small percentage of the simulation replications, as can be seen in the high coverage percentages, however

those few replications result in large simulation variability. When this is the case it may be more interesting to use an alternative variability measurement to the standard deviation, such as the median absolute deviation. Figure 2.9 is the same as Figure 2.8 but with one median absolute deviation used to show the variability rather than one standard deviation. Figure 2.9 is a much better representation of how the simulation affects the estimation of γ_0 in DP1.

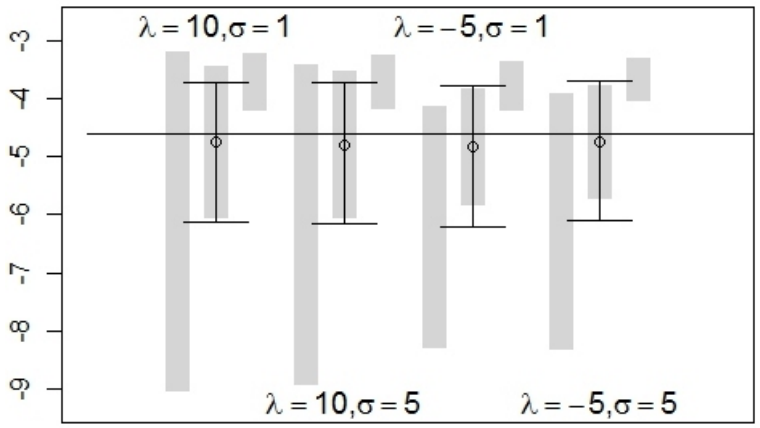

Figure 2.8: Spread of posterior estimates for γ_0 in DP1.

Estimation of λ is similar to what we saw with the efficacy regression parameters. The width of the intervals is dependent on the value of σ as well as the probability of an adverse event. Thus, the interval width decreases as σ decreases, and the interval width increases as π_i goes towards 0.

Estimation of σ seems to only depend on the value of σ itself. When $\sigma = 1$ we see interval widths of approximately 0.11 for every simulation, while when $\sigma = 5$ the interval width increases to approximately 0.56 for each simulation. Additional simulation results are included in Appendix A.

2.5 Discussion

The results of the simulation suggest that the performance of the model is heavily dependent on the probability of an adverse event. As π_i approaches 0 the

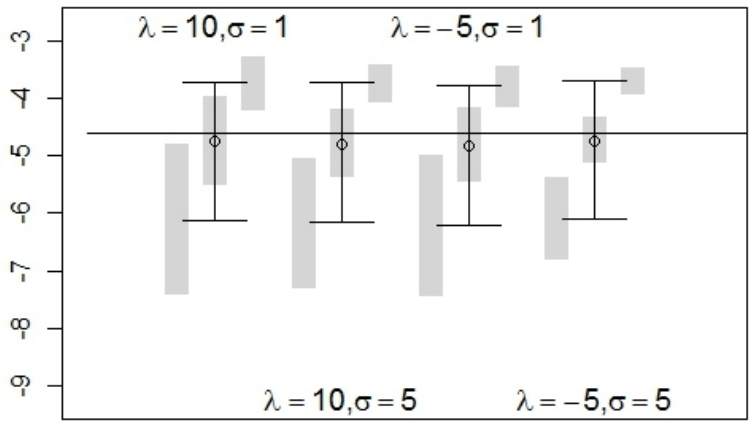

Figure 2.9: Spread of posterior estimates for γ_0 in DP1 using the median absolute deviation.

model behaves erratically, especially in estimating γ_0 and γ_1. However, as π_i increases, the model seems to perform better. We see this in Figures 2.5 and 2.8. One reason this could be happening is the induced prior on π_i. We see in Figure 2.10 that for relatively non-informative normal priors on the three safety regression parameters we induce a "bathtub" shaped prior on π_i. This means that as the true value of π_i approaches zero, the prior may start to overwhelm the likelihood in the posterior distribution. As we put more informative priors on the safety regression parameters, the induced prior becomes less of a problem. Thus, if we have information about π_i that can be translated into information about the safety regression parameters we need to include it in our prior specification. One way to do this is through conditional means priors. Seaman III et al. (2012) discuss the use of conditional means priors and also give a more complete discussion of the induced prior problem. Another way to improve the results is by increasing the sample size. This would result in a larger number of adverse events. With more information in the data, the prior would not have as much of an impact on the posterior. Note that the induced prior on μ_i appears to be, like the priors for the efficacy regression parameters and the prior for λ, a relatively non-informative normal prior.

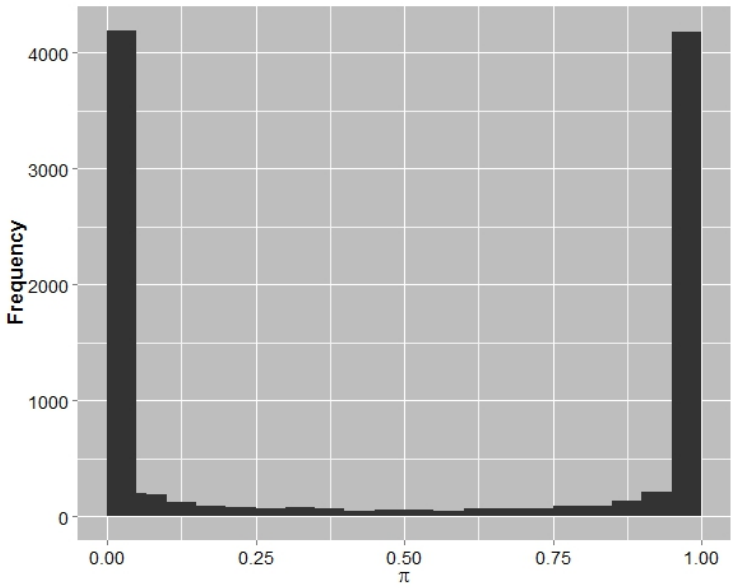

Figure 2.10: Induced prior on π when generating 10,000 values from the covariate distributions and the diffuse normal priors on the safety regression parameters.

CHAPTER THREE

Higher Order Approximation to the Mixed Response Model

3.1 Introduction

Normal-theory approximations are often used in the construction of interval estimates and hypothesis testing. One of these approximations is based on the Wald statistic. This approach tends to perform well when we have a large sample size resulting in an accurate normal approximation. We do not, however, know when the sample is large enough to facilitate an accurate approximation. For complex models, the regularity conditions necessary for the use of such approximations may be suspect and difficult to check. Higher order approximations offer an alternative. These include Edgeworth and saddlepoint approximations (Butler, 2007) the likelihood based p-formula, and Barndorff-Nielsen's p^*-formula (Barndorff-Nielsen, 1983; Pawitan, 2001), and r^*, the modified likelihood root formula (Brazzale et al., 2007).

In this chapter we consider methods of obtaining approximate likelihoods in order to perform inference for the bivariate discrete and continuous model mentioned in Chapter Two, repeated here for convenience:

$$Y_{1i} \sim \text{Bernoulli}(\pi_i), \tag{3.1}$$

and

$$Y_{2i}|Y_{1i} \sim N(\mu_i, \sigma^2),$$

with link functions

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i},$$

and

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \lambda(y_{1i} - \pi_i).$$

In Section 3.2 we consider the commonly used p^* method. As we shall see, this approach is not suitable for our model. In Section 3.3, we consider a higher-order approximation, the r^* method. We apply this approach to a simple linear regression model in Section 3.4 for illustration purposes. In Section 3.6, we use the r^* method for our bivariate discrete and continuous model in (3.1). We conclude with a discussion of our conclusions in Sections 3.8 and 3.10 respectively.

3.2 The p^* Method

We consider the p^*-formula because the r^* method was originally derived from it. The p^*-formula often yields a good approximation to the sampling density and is in wide use in the literature (Pace et al., 2011; Jiang and Wong, 2012; Cortese and Ventura, 2012; Barreto et al., 2013). Efron (1998) even goes so far as to refer to the p^*-formula as a 'magic formula' because of how well the approximation performs in many cases.

Our overview of the p^* method follows that of Pawitan (2001, pp. 247-250). Suppose we have observations from a distribution indexed by a real-valued parameter, θ. Let $\hat{\theta}$ be a maximum likelihood estimator (MLE) of θ. From normal theory we have,

$$\hat{\theta} \sim N\left(\theta, j(\hat{\theta})^{-1}\right),$$

where $j(\hat{\theta})$ is the observed information. Pawitan gives an in depth discussion of the use of the observed versus the expected information. The approximate sampling density of $\hat{\theta}$ is

$$p_\theta(\hat{\theta}) \approx (2\pi)^{-1/2}|j(\hat{\theta})|^{1/2} \exp\left\{-\frac{j(\hat{\theta})}{2}(\hat{\theta}-\theta)^2\right\}. \tag{3.2}$$

It can also be shown that an approximation of the log likelihood ratio statistic is

$$\log\frac{L(\theta)}{L(\hat{\theta})} \approx -\frac{j(\hat{\theta})}{2}(\hat{\theta}-\theta)^2.$$

Substituting, we have

$$p_\theta(\hat{\theta}) \approx (2\pi)^{-1/2}|j(\hat{\theta})|^{1/2}\frac{L(\theta)}{L(\hat{\theta})}. \tag{3.3}$$

This is referred to as the likelihood-based p-formula, and Pawitan (2001) states that it is 'much more accurate' than the normal-based formula in (3.2). We can improve (3.3) by including a normalizing constant to make the density integrate to one. The density is then

$$p^*_\theta(\hat{\theta}) \approx c(\theta)(2\pi)^{-1/2}|j(\hat{\theta})|^{1/2}\frac{L(\theta)}{L(\hat{\theta})}, \tag{3.4}$$

and is called Barndorff-Nielsen's p^*-formula (Barndorff-Nielsen, 1983).

In order to use the p^*-formula for inference, we find a p-value for testing $H_0 : \theta = \theta_0$. If we are interested in a confidence interval for θ at some specified significance level, α, we need to find the θ_0's such that the p-values are $\alpha/2$ and $(1 - \alpha/2)$. The difficulty with this method comes when the parameter is vector-valued — call it $\boldsymbol{\theta}$ — with dimension k. We are often only interested in a sub-vector of $\boldsymbol{\theta}$ but the p^* method does not allow for easy elimination of nuisance parameters. This forces the use of k-dimensional hypotheses or corresponding confidence regions, which is often not practical. For this reason, we have chosen to employ another higher-order approximation method that does allow for nuisance parameters. We turn to that approach now.

3.3 The r^* Method

The r^* approximation was originally derived from the p^*-formula in Barndorff-Nielsen (1986) and Barndorff-Nielsen (1990). The results are also summarized in Barndorff-Nielsen and Cox (1994, Chapter 6). This approach affords a higher-order approximation with the flexibility to handle nuisance parameters. We begin with a detailed look at the basic r^* method, applied to scalar parameters, following Brazzale et al. (2007, pp. 5 ff.).

Let $z_1, ..., z_n$ be independent identically distributed random variables with probability density function $f(z;\theta)$, where θ is an unknown scalar parameter. Let $\ell(\theta) = \sum \log f(z_i;\theta)$ be the log likelihood, $\hat{\theta}$ the MLE, and $j(\theta) = -\partial^2\ell(\theta)/\partial\theta^2$ the observed information function. We denote the likelihood root by

$$r(\theta) = \text{sign}(\hat{\theta} - \theta)\left[2\left\{\ell(\hat{\theta}) - \ell(\theta)\right\}\right]^{1/2}. \tag{3.5}$$

The score statistic is

$$s(\theta) = j(\hat{\theta})^{-1/2}\partial\ell(\theta)/\partial\theta, \tag{3.6}$$

and the Wald statistic is

$$t(\theta) = j(\hat{\theta})^{1/2}(\hat{\theta} - \theta). \tag{3.7}$$

Note that (3.5) is a transformation of the likelihood ratio statistic,

$$w(\theta) = r(\theta)^2 = 2\left\{\ell(\hat{\theta}) - \ell(\theta)\right\}. \tag{3.8}$$

Equations (3.5), (3.6), and (3.7) all have asymptotic standard normal distributions under $f(z;\theta)$ and assuming certain conditions. Similarly, (3.8) has an asymptotic chi-squared distribution with one degree of freedom.

Brazzale et al. (2007, Chapter 8) discuss higher order asymptotics. They describe the modified likelihood root, which is

$$r^*(\theta) = r(\theta) + \frac{1}{r(\theta)}\log\left\{\frac{q(\theta)}{r(\theta)}\right\}, \tag{3.9}$$

where $q(\theta)$ can be the score statistic, Wald statistic, or other functions depending on the situation. For the univariate case, if the model is a member of the exponential family, the Wald statistic is appropriate and $q(\theta) = t(\theta)$. As with equations (3.5)–(3.7), $r^*(\theta)$ has an asymptotic standard normal distribution. Brazzale et al. (2007) state that the "normal approximation to the distribution of $r^*(\theta)$ can provide almost exact inferences for θ, when these are available."

Confidence intervals for θ using $r^*(\theta)$ can be calculated by testing the null hypothesis that $\theta = \theta_0$ for some specified θ_0 against the alternative, $\theta \neq \theta_0$. Thus,

we use a grid of θ_0 values and find the values such that $\Phi\left(r^*(\theta_0)\right) = \alpha/2$, and $\Phi\left(r^*(\theta_0)\right) = 1 - \alpha/2$ for some specified confidence level α, where Φ denotes the standard normal distribution function.

In general, we are more interested in models with more than one variable. Let $\boldsymbol{\theta} = (\boldsymbol{\psi}, \boldsymbol{\lambda})$ be the $d \times 1$ vector of parameters where $\boldsymbol{\psi}$ is the, possibly vector-valued, parameter of interest and $\boldsymbol{\lambda}$ is the vector of nuisance parameters. In order to perform inference on $\boldsymbol{\psi}$ we must eliminate the nuisance parameters, $\boldsymbol{\lambda}$. We will do this using the profile log likelihood, by substituting $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ for $\boldsymbol{\lambda}$, where $\hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}$ is the constrained MLE found by maximizing $\ell(\boldsymbol{\psi}, \boldsymbol{\lambda})$ with respect to $\boldsymbol{\lambda}$ for fixed $\boldsymbol{\psi}$. The profile log likelihood is

$$\ell_p(\boldsymbol{\psi}) = \max_{\boldsymbol{\lambda}} \ell(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \ell(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}). \tag{3.10}$$

The observed information for the profile log likelihood is

$$\begin{aligned} j_p(\boldsymbol{\psi}) &= -\frac{\partial^2 \ell_p(\boldsymbol{\psi})}{\partial \boldsymbol{\psi} \partial \boldsymbol{\psi}^T} \\ &= \left\{ j^{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}) \right\}^{-1}. \end{aligned} \tag{3.11}$$

Partition the observed information matrix into

$$j(\boldsymbol{\psi}, \boldsymbol{\lambda}) = \begin{pmatrix} j_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}, \boldsymbol{\lambda}) & j_{\boldsymbol{\psi}\boldsymbol{\lambda}}(\boldsymbol{\psi}, \boldsymbol{\lambda}) \\ j_{\boldsymbol{\lambda}\boldsymbol{\psi}}(\boldsymbol{\psi}, \boldsymbol{\lambda}) & j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\boldsymbol{\psi}, \boldsymbol{\lambda}) \end{pmatrix}$$

such that $j^{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}) = \left(j_{\boldsymbol{\psi}\boldsymbol{\psi}}(\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}}) \right)^{-1}$. If the parameter of interest is a scalar, ψ, then rather than (3.11), we can define

$$j_p(\psi) = \frac{|j(\psi, \hat{\boldsymbol{\lambda}}_{\psi})|}{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\psi, \hat{\boldsymbol{\lambda}}_{\psi})|},$$

where $|\cdot|$ is the determinant.

We can now define a modified likelihood root, r^*, that is similar to the univariate version in (3.9), but includes an improved approximation and adjusts for eliminating the nuisance parameters. This modified likelihood root is

$$r^*(\boldsymbol{\psi}) = r(\boldsymbol{\psi}) + \frac{1}{r(\boldsymbol{\psi})} \log \left\{ \frac{q(\boldsymbol{\psi})}{r(\boldsymbol{\psi})} \right\}, \tag{3.12}$$

where

$$r(\boldsymbol{\psi}) = \text{sign}(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}) \left[2 \left\{ \ell_p(\hat{\boldsymbol{\psi}}) - \ell_p(\boldsymbol{\psi}) \right\} \right]^{1/2}, \tag{3.13}$$

$$q(\boldsymbol{\psi}) = t(\boldsymbol{\psi}) \rho(\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}), \tag{3.14}$$

$$t(\boldsymbol{\psi}) = j_p^{1/2}(\hat{\boldsymbol{\psi}})(\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}), \tag{3.15}$$

and

$$\rho(\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}) = \left\{ \frac{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}})|}{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\hat{\boldsymbol{\theta}}_{\boldsymbol{\psi}})|} \right\}^{1/2}, \tag{3.16}$$

where $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\lambda}})$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\psi}} = (\boldsymbol{\psi}, \hat{\boldsymbol{\lambda}}_{\boldsymbol{\psi}})$. The form of q has been modified from the univariate case. As before, r^* can be approximated by the standard normal distribution.

Brazzale et al. (2007, p. 14) discuss that the most difficult part of using the r^* approximation is the choice of the function q. When the parameter of interest is a component of the canonical parameter in an exponential family, then the calculation of (3.14) is fairly straightforward and has a good approximation to the exact conditional distribution. Brazzale et al. (2007) also consider other methods for choosing the q function.

Note that in order to use the r^* method we need to have MLE's for each of our parameters. In general, we are interested in the r^* approximation in cases when there are no closed form solutions for the MLE's. A numerical algorithm will have to be implemented. We have used the `mle` function found in the `stats4` package in `R`. This function calls the `optim` function from the `stats` package, and we have used the default method by Nelder and Mead (1965). We now illustrate the r^* method for the simple linear regression case.

3.4 Example: Simple Linear Regression

Consider the simple linear regression case where we let $y_1, ..., y_n$ be iid observations from a $N(\mu_i, \sigma^2)$ distribution where $\mu_i = \beta_0 + \beta_1 x_i$, and x_i is some covariate of

interest. We are interested in using the r^* approach to construct confidence intervals for β_0, β_1, and σ^2. The likelihood is

$$L(\mu_i, \sigma^2) = \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left\{ -\frac{1}{2\sigma_i^2}(y_i - \mu_i)^2 \right\},$$

and the log likelihood is

$$\ell(\mu_i, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu_i)^2.$$

Substituting for μ_i we can calculate the MLE's as the solution to the system of equations

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x},$$
$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\beta}_0 n\bar{x}}{\sum_{i=1}^{n} x_i^2},$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n}.$$

We will also implement our numerical algorithm for the MLE's, but will be able to compare their results to the exact values.

Because we are interested in inference on all three parameters for this model, we could let $\boldsymbol{\theta} = \boldsymbol{\psi} = (\beta_0, \beta_1, \sigma^2)$, using the terminology from Section 3.3. This approach would be analogous to the p^* method from Section 3.2, and we are not interested in this because we would be creating a three dimensional confidence region rather than three individual intervals. Thus, we will be implementing the r^* method three times with each of the three parameters as the one of interest.

Let β_0 be the parameter of interest in the first iteration through the method. We can then define

$$\begin{aligned}\ell_p(\beta_0) &= \max_{\boldsymbol{\lambda}} \ell(\beta_0, \boldsymbol{\lambda}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\hat{\sigma}^2 - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}(y_i - \beta_0 - \hat{\beta}_1 x_i)^2, \qquad (3.17)\end{aligned}$$

where $\boldsymbol{\lambda} = (\beta_1, \sigma^2)$. Next, using (3.17), we have

$$\begin{aligned} r(\beta_0) &= \text{sign}(\hat{\beta}_0 - \beta_0) \left[2 \left\{ \ell_p(\hat{\beta}_0) - \ell_p(\beta_0) \right\} \right]^{1/2} \\ &= \text{sign}(\hat{\beta}_0 - \beta_0) \left\{ -\frac{1}{\hat{\sigma}^2} \sum_{i=1}^{n} \left[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 - (y_i - \beta_0 - \hat{\beta}_1 x_i)^2 \right] \right\}. \end{aligned} \tag{3.18}$$

It can be shown that

$$j(\beta_0, \beta_1, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{\sum_{i=1}^n x_i}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^n h_i \\ \frac{\sum_{i=1}^n x_i}{\sigma^2} & \frac{\sum_{i=1}^n x_i^2}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^n x_i h_i \\ \frac{1}{\sigma^4} \sum_{i=1}^n h_i & \frac{1}{\sigma^4} \sum_{i=1}^n x_i h_i & \frac{1}{\sigma^6} \sum_{i=1}^n h_i^2 - \frac{n}{2\sigma^4} \end{pmatrix}, \tag{3.19}$$

where $h_i = y_i - \beta_0 - \beta_1 x_i$. Thus, using (3.19),

$$j_p(\beta_0) = \left\{ j^{\beta_0 \beta_0}(\beta_0, \hat{\boldsymbol{\lambda}}_{\beta_0}) \right\}^{-1}. \tag{3.20}$$

and from (3.20)

$$t(\beta_0) = j_p^{1/2}(\hat{\beta}_0)(\hat{\beta}_0 - \beta_0). \tag{3.21}$$

Using (3.19), we can now define

$$\rho(\beta_0, \hat{\beta}_0) = \left\{ \frac{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\hat{\beta}_0, \hat{\boldsymbol{\lambda}})|}{|j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\beta_0, \hat{\boldsymbol{\lambda}}_{\beta_0})|} \right\}^{1/2}, \tag{3.22}$$

where

$$j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\hat{\beta}_0, \hat{\boldsymbol{\lambda}}) = \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{\hat{\sigma}^2} & \frac{1}{\hat{\sigma}^4} \sum_{i=1}^n x_i h_i \\ \frac{1}{\hat{\sigma}^4} \sum_{i=1}^n x_i h_i & \frac{1}{\hat{\sigma}^6} \sum_{i=1}^n h_i^2 - \frac{n}{2\hat{\sigma}^4} \end{pmatrix},$$

and

$$j_{\boldsymbol{\lambda}\boldsymbol{\lambda}}(\beta_0, \hat{\boldsymbol{\lambda}}_{\beta_0}) = \begin{pmatrix} \frac{\sum_{i=1}^n x_i^2}{\hat{\sigma}^2_{\beta_0}} & \frac{1}{\hat{\sigma}^4_{\beta_0}} \sum_{i=1}^n x_i h_i \\ \frac{1}{\hat{\sigma}^4_{\beta_0}} \sum_{i=1}^n x_i h_i & \frac{1}{\hat{\sigma}^6_{\beta_0}} \sum_{i=1}^n h_i^2 - \frac{n}{2\hat{\sigma}^4_{\beta_0}} \end{pmatrix}.$$

Thus, from (3.21) and (3.22), we have

$$q(\beta_0) = t(\beta_0)\rho(\beta_0, \hat{\beta}_0), \tag{3.23}$$

and using (3.18) and (3.23) we get

$$r^*(\beta_0) = r(\beta_0) + \frac{1}{r(\beta_0)} \log \left\{ \frac{q(\beta_0)}{r(\beta_0)} \right\},$$

which has an asymptotic standard normal distribution. Inference on β_1 and σ^2 follows a similar development to β_0.

We will be analyzing a simulated sample for the Wald, likelihood root, and r^* methods. Let $\beta_0 = 5$, $\beta_1 = 3$, and $\sigma^2 = 2$. We will simulate a sample of independent y_i's of size $n = 30$ from a $N(\beta_0 + x\beta_1, \sigma^2)$ distribution where we choose $x \sim N(0, 4)$. For this example we have chosen the endpoints for β_0 of $(2, 8)$ enclosing a grid of 500 points. This is an arbitrary choice designed to be outside the 95% interval for β_0. In practice one would not know the true value of the parameter and could choose the endpoints using alternative methods. One such method is to calculate the MLE and information for a parameter and set the endpoints to the MLE plus and minus 3 standard deviations. The grid size should be chosen to reflect the amount of precision needed.

The MLE for β_0 using the numerical algorithm described in Section 3.3 is approximately 5.11, seen in Table 3.1. Notice that the approximate MLE appears equal to the true MLE, also in Table 3.1. The two MLE's are equal to at least six decimal places. We also see that the confidence interval for β_0 using the r^* method is $(4.57, 5.64)$. Comparatively, the intervals using both the Wald and likelihood root methods are $(4.59, 5.63)$. In this example, the inference for the three methods is similar.

Table 3.1 also includes the results for β_1 and σ^2. As with β_0 the true and numerical MLE's agree out to six decimal places. The intervals for β_1 are very similar, while the intervals for σ^2 are close but not as much as the other two parameters. Note that the length of the r^* interval for σ^2 is shorter than that of the Wald or likelihood ratio.

Table 3.1: Likelihood results for the three parameters of the simple linear regression example.

Method	β_0		β_1		σ^2	
True MLE	5.11		2.96		2.06	
Numerical MLE	5.11		2.96		2.06	
Statistic	2.5%	97.5%	2.5%	97.5%	2.5%	97.5%
Wald	4.59	5.63	2.83	3.10	1.01	3.11
Likelihood Root	4.59	5.63	2.83	3.10	1.29	3.58
r^*	4.57	5.64	2.81	3.10	1.18	3.12

3.5 Example: Gamma Data

The r^* method is most useful when statistics such as the Wald pivot have a poor normal approximation to the parameter of interest. We have adapted an example from Brazzale et al. (2007) illustrating this using gamma data. We have included more detail than in their presentation.

Consider a sample $y_1, ..., y_n$ from the gamma distribution

$$f(y; \psi, \lambda) = \frac{\lambda^\psi y^{\psi-1}}{\Gamma(\psi)} \exp(-\lambda y), \quad y > 0, \lambda, \psi > 0,$$

where our parameter of interest is ψ. The log likelihood is

$$\ell(\psi, \lambda) = \psi \sum_{i=1}^{n} \log y_i - n\lambda\bar{y} + n(\psi \log \lambda - \log \Gamma(\psi)) - \sum_{i=1}^{n} \log y_i. \tag{3.24}$$

Taking the derivative with respect to λ we have

$$\frac{\partial \ell(\psi, \lambda)}{\partial \lambda} = -n\bar{y} + \frac{n\psi}{\lambda}.$$

Setting to zero and solving gives the MLE

$$\begin{aligned} 0 = -n\bar{y} + \frac{n\psi}{\hat{\lambda}} \Rightarrow n\bar{y} &= \frac{n\psi}{\hat{\lambda}} \\ \Rightarrow \hat{\lambda} = \hat{\lambda}_\psi &= \frac{\psi}{\bar{y}}. \end{aligned} \tag{3.25}$$

The negation of the second derivative with respect to λ yields

$$\begin{aligned} j_{\lambda\lambda}(\psi,\lambda) &= -\frac{\partial^2 \ell(\psi,\lambda)}{\partial \lambda^2} \\ &= -\left(-\frac{n\psi}{\lambda^2}\right) \\ &= \frac{n\psi}{\lambda^2}. \end{aligned} \tag{3.26}$$

Using (3.16) we see

$$\begin{aligned} \rho(\psi,\hat{\psi}) &= \left\{ \frac{|j_{\lambda\lambda}(\hat{\psi},\hat{\lambda})|}{|j_{\lambda\lambda}(\psi,\hat{\lambda}_\psi)|} \right\}^{1/2} \\ &= \left\{ \frac{\frac{n\hat{\psi}}{\hat{\lambda}^2}}{\frac{n\psi}{\hat{\lambda}_\psi^2}} \right\}^{1/2} \\ &= \left\{ \frac{\frac{n\hat{\psi}}{\left(\frac{\hat{\psi}}{\bar{y}}\right)^2}}{\frac{n\psi}{\left(\frac{\psi}{\bar{y}}\right)^2}} \right\}^{1/2} \\ &= \left\{ \frac{\frac{\bar{y}^2}{\hat{\psi}}}{\frac{\bar{y}^2}{\psi}} \right\}^{1/2} \\ &= \left(\frac{\psi}{\hat{\psi}}\right)^{1/2}, \end{aligned} \tag{3.27}$$

and $\hat{\psi}$ can be calculated numerically. From (3.24) and (3.25) we can get the profile log likelihood

$$\begin{aligned} \ell_p(\psi) &= \psi \sum_{i=1}^{n} \log y_i - n\bar{y}\left(\frac{\psi}{\bar{y}}\right) + n\left(\psi \log \frac{\psi}{\bar{y}} - \log\Gamma(\psi)\right) - \sum_{i=1}^{n} \log y_i \\ &= \psi \sum_{i=1}^{n} \log y_i - n\psi + n\psi \log \frac{\psi}{\bar{y}} - n \log\Gamma(\psi) - \sum_{i=1}^{n} \log y_i \\ &= \psi\left[\left(\sum_{i=1}^{n} \log y_i\right) - n\right] + n\left[\psi \log \frac{\psi}{\bar{y}} - \log\Gamma(\psi)\right] - \sum_{i=1}^{n} \log y_i. \end{aligned} \tag{3.28}$$

Now taking the derivative of (3.28) with respect to ψ gives

$$\begin{aligned} \frac{\partial \ell_p(\psi)}{\partial \psi} &= \left[\sum_{i=1}^{n} \log y_i\right] - n + n\log\psi + n - n\log\bar{y} - n\Psi(\psi) \\ &= \sum_{i=1}^{n} \log y_i + n\log\psi - n\log\bar{y} - n\Psi(\psi), \end{aligned}$$

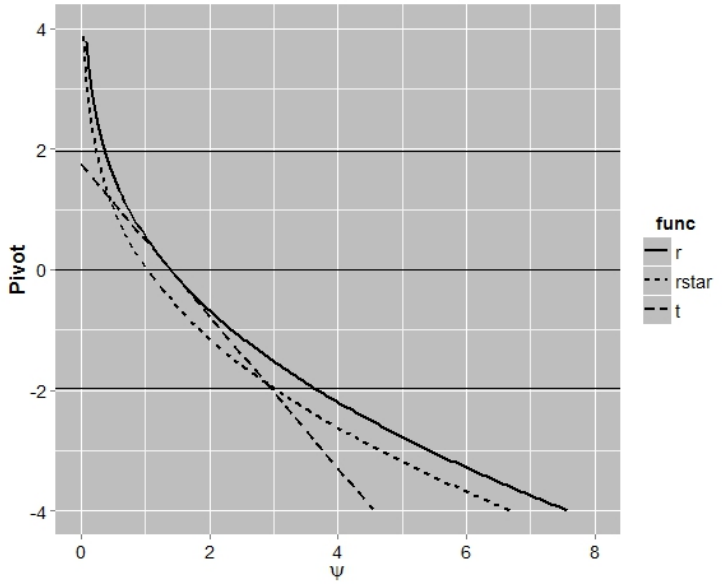

Figure 3.1: Inference for shape parameter ψ of gamma sample of size $n = 5$. The likelihood root, Wald pivot, and modified likelihood root. The horizontal lines are at 0, ± 1.96.

where $\Psi(\psi) = \Gamma'(\psi)/\Gamma(\psi)$ is the digamma function. The second derivative with respect to ψ produces

$$\begin{aligned} j_p(\psi) &= -\frac{\partial^2 \ell_p(\psi)}{\partial \psi^2} \\ &= -\left[\frac{n}{\psi} - n\Psi'(\psi)\right] \\ &= n\left[\Psi'(\psi) - \frac{1}{\psi}\right]. \end{aligned} \tag{3.29}$$

Using (3.27), (3.28), and (3.29) we can calculate $r(\psi)$, $t(\psi)$, and $r^*(\psi)$ in order to find confidence intervals for ψ.

We generate five observations from a $Gamma(1, 1)$ density, which are 1.22, 0.32, 0.42, 0.25, and 2.55. We can see in Figure 3.1 (similar to the right hand panel of Figure 2.3 in Brazzale et al. (2007)) the likelihood root, $r(\psi)$, Wald, $t(\psi)$, and modified likelihood root, $r^*(\psi)$, statistics plotted against ψ. We see that the Wald pivot is giving poor results because of the poor normal approximation. The modified likelihood root, $r^*(\psi)$, gives results that are more appropriate than the

Wald, and the 95% confidence interval is an improvement over $r(\psi)$. Thus, we see that the modified likelihood root gives appropriate inference even in situations where common methods, such as the Wald, are not suitable.

3.6 The r^* Method for the Mixed Response Model

Recall the bivariate discrete and continuous model from (3.1). This model is far more complex than either the simple linear regression or gamma models. We will now apply the r^* method to this model.

We know that the r^* approximation is excellent when the parameter is a member of an exponential family. We have attempted to show that the bivariate binomial and normal model is a member of the exponential family without success. As an alternative method of showing that the mixed response model has regularity conditions similar to those in the exponential family, Figure 3.2 shows a contour plot of the bivariate log likelihood of γ_1 and β_1, as well as the marginal log likelihood plots. We see a slight ridge in the contour plot, but for the most part, both γ_1 and β_1 are symmetric and 'unimodal'. The contour plots and log likelihoods are similar for $\gamma_0, \beta_0, \gamma_2$, and β_2, but the ridges are slightly more pronounced. The log likelihood for λ is also symmetric and unimodal, however the log likelihood for σ is positively skewed, although it is unimodal.

The likelihood for the bivariate model is listed in (2.5) and so the log likelihood is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\gamma}, \lambda, \sigma) = \sum_{i=1}^{n} [y_{1i} \log \pi_i + (1 - y_{1i}) \log(1 - \pi_i)] - \frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_{2i} - \mu_i)^2,$$

where μ_i and π_i are defined in (2.4) and (2.6). There is no closed form for the MLE's so they will be calculated using the numerical method described in Section 3.3. As with the simple linear regression case, we will employ the r^* procedure individually for each of the eight parameters in our model.

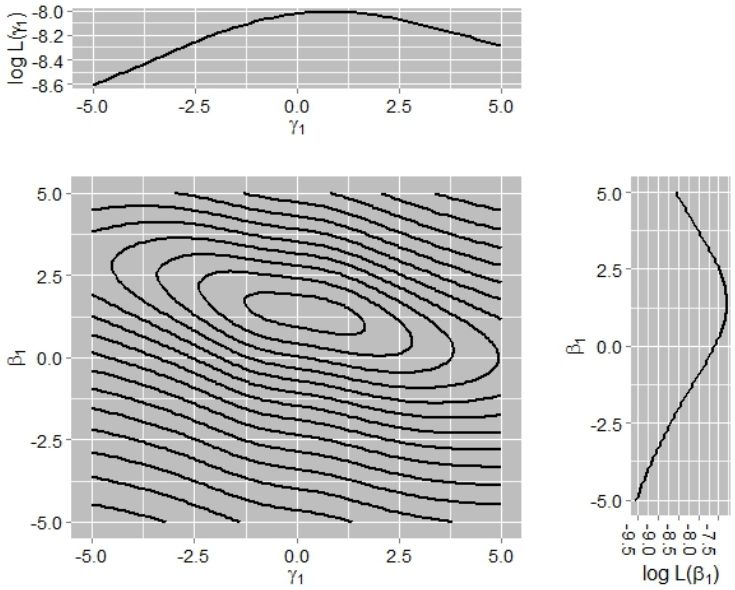

Figure 3.2: Contour plot and log likelihoods for γ_1 and β_1.

Let γ_1 be the parameter of interest on the first iteration through the method and let $\boldsymbol{\theta} = (\gamma_0, \gamma_2, \boldsymbol{\beta}, \lambda, \sigma)$ be the vector of nuisance parameters. The profile log likelihood is

$$\ell_p(\gamma_1) = \max_{\boldsymbol{\theta}} \ell(\gamma_1, \boldsymbol{\theta}) = \ell(\gamma_1, \hat{\theta}_{\gamma_1}).$$

The likelihood root is then

$$r(\gamma_1) = \text{sign}(\hat{\gamma}_1 - \gamma_1) \left[2 \left\{\ell_p(\hat{\gamma}_1) - \ell_p(\gamma_1)\right\}\right]^{1/2}.$$

The observed information, $j(\boldsymbol{\gamma}, \boldsymbol{\beta}, \lambda, \sigma)$, was calculated using *Mathematica 8.0.* The code used to calculate the observed information is included in Appendix B. The profile information for γ_1 is

$$j_p(\gamma_1) = \left\{ j^{\gamma_1 \gamma_1}(\gamma_1, \hat{\theta}_{\gamma_1}) \right\}^{-1},$$

and thus

$$t(\gamma_1) = j_p^{1/2}(\hat{\gamma}_1)(\hat{\gamma}_1 - \gamma_1).$$

Next,

$$\rho(\gamma_1, \hat{\gamma}_1) = \left\{ \frac{|j_{\boldsymbol{\theta\theta}}(\hat{\gamma}_1, \hat{\boldsymbol{\theta}})|}{|j_{\boldsymbol{\theta\theta}}(\gamma_1, \hat{\boldsymbol{\theta}}_{\gamma_1})|} \right\}^{1/2},$$

and so

$$q(\gamma_1) = t(\gamma_1)\rho(\gamma_1, \hat{\gamma}_1),$$

such that

$$r^*(\gamma_1) = r(\gamma_1) + \frac{1}{r(\gamma_1)} \log \left\{ \frac{q(\gamma_1)}{r(\gamma_1)} \right\}.$$

As before, $r^*(\gamma_1)$ has an asymptotic standard normal distribution. We calculate $r^*(\gamma_1)$ for a grid of γ_1 values. For some specified significance level, α, we then find the γ_1's such that $\Phi(r^*(\gamma_1)) \approx \alpha$ and $\Phi(r^*(\gamma_1)) \approx 1 - \alpha$.

We follow a similar development with the other seven parameters in our model. The result is that we have $100(1 - \alpha)\%$ confidence intervals for each parameter. These intervals can then be compared to the credible sets from Chapter Two. We now look at the weight-loss/blood pressure and weight-loss/suicide examples from Sections 2.3.

3.7 Examples

3.7.1 Weight Loss/Blood Pressure

Consider the situation of a weight loss drug from Section 2.3.2 where we are interested in the percent body weight lost and the occurrence of an increase of at least 10 mm/hg in systolic blood pressure. As before, $y_{1i} = 1$ is our safety indicator for a blood pressure increase and is 0 otherwise, and y_{2i} is the percent of body weight lost while on the drug. From Section 2.3.2 we have two covariates. The first covariate, x_{1i}, is age, which we generate as $N(40, 16)$ and then standardize to a $N(0, 1)$. The second covariate, x_{2i}, is gender, and has a $Bernoulli(0.5)$ distribution. We have chosen to target $\pi_i = 0.05$ and $\mu_i = 5$ which results in parameter values of $\beta_0 = 3$, $\beta_1 = 1.33$, $\beta_2 = 4$, $\gamma_0 = -2.44$, $\gamma_1 = 0.41$, $\gamma_2 = -1.45$, $\lambda = -5$ and $\sigma = 1$.

We generate a sample of size $n = 600$, and we use the same seed as in the Bayesian example.

Table 3.2a shows the results of the blood pressure example. Notice that the true value for each parameter is listed along with the MLE's and lower and upper bounds to the 95% confidence interval. The MLE's are fairly close to the true value of the parameters although the estimates for γ_0, γ_2, and λ are not as close as the others. The 95% confidence intervals include the true value of all 8 parameters, even β_0 by a small margin, which is not evident in Table 3.2a. We have included the results for the blood pressure example using the Bayesian approach for comparison. These are reproduced for convenience and listed in Table 3.2b.

Table 3.2: The results of the 95% confidence intervals and credible sets for the weight loss/systolic blood pressure example using the r^* and Bayesian methods respectively.

(a) The r^* method.

Par	True	2.5%	MLE	97.5%
β_0	3.00	3.00	3.08	3.18
β_1	1.33	1.23	1.32	1.40
β_2	4.00	3.83	3.95	4.08
γ_0	-2.44	-2.89	-2.62	-2.34
γ_1	0.41	0.30	0.52	0.73
γ_2	-1.45	-2.60	-1.71	-1.01
λ	-5.00	-5.20	-4.80	-4.42
σ	1.00	0.92	0.97	1.03

(b) The Bayesian method.

Par	True	2.5%	mean	97.5%
β_0	3.00	2.81	3.01	3.18
β_1	1.33	1.21	1.32	1.43
β_2	4.00	3.74	3.97	4.20
γ_0	-2.44	-2.81	-2.38	-1.99
γ_1	0.41	0.12	0.42	0.72
γ_2	-1.45	-3.30	-2.10	-1.12
λ	-5.00	-5.07	-4.69	-4.32
σ	1.00	0.97	1.03	1.09

3.7.2 Weight Loss/Suicidal Ideations

Recall the weight loss scenario where the adverse event of interest, y_{1i}, was a suicide attempt or not. We continue to define the efficacy response, y_{2i}, as the percent body weight lost. Age and gender are our covariates and we target $\pi_i = 0.02$ and $\mu_i = 10$. These target values result in values of the parameters of $\beta_0 = 13$, $\beta_1 = 1.67$,

$\beta_2 = -6$, $\gamma_0 = -4.6$, $\gamma_1 = 0.24$, $\gamma_2 = 1.1$, $\lambda = 10$, and $\sigma = 5$. We use the same seed as in the Bayesian suicide example in Section 2.3.1 and simulate a sample of size $n = 600$.

Recall from the simulation for the Bayesian model in Chapter Two that, as π_i approached zero, estimation became increasingly problematic. As we shall see in Section 3.8, the r^* method suffers in the same way. Here we apply the r^* method to the same data used to illustrate the small π_i behavior for the Bayesian model. Unfortunately, applying the r^* method to the same data as in Section 2.3.1 we encountered an error. This error occurred because of the method used to calculate the r^* interval. Recall that the Wald pivot is used in the r^* method and the square root of the profile information is used to calculate the Wald pivot. The observed information function of the profile log likelihood is $j_p(\psi) = \left\{ j^{\psi\psi}(\psi, \hat{\lambda}_\psi) \right\}$, where $j^{\psi\psi}(\psi, \hat{\lambda}_\psi)$ is the (ψ, ψ) block of the inverse of the observed information matrix $j(\psi, \lambda)$. Brazzale et al. (2007) suggests an alternative for the profile observed information when ψ is a scalar, as in our scenario. They suggest

$$j_p(\psi) = \frac{|j(\psi, \hat{\lambda}_\psi)|}{|j_{\lambda\lambda}(\psi, \hat{\lambda}_\psi)|}.$$

Using either method, occasionally the profile information will be negative for one or more parameters, which results in an error in the algorithm. This happens more often than not for our model when our binary response variable has small probabilities. We will discuss this further in Section 3.8.

3.8 Simulation

We would like to compare the r^* method to the Bayesian method detailed in Chapter Two. Recall from Section 2.4 that we targeted three possible values each of π_i and μ_i in the Bayesian simulation for a total of nine design points. The table of design points is reproduced in Table 3.3. We then had two values each of λ and σ that, combined with the nine design points produced 36 total simulations. As in

Table 3.3: The design points over which the r^* simulation will be run.

Efficacy	$\pi = 0.02$	$\pi = 0.05$	$\pi = 0.20$
$\mu = 5$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (3, 1.33, 4)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (3, 1.33, 4)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (3, 1.33, 4)$
$\mu = 10$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (13, 1.67, -6)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (13, 1.67, -6)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (13, 1.67, -6)$
$\mu = 20$	$\boldsymbol{\gamma} = (-4.6, 0.24, 1.1)$ $\boldsymbol{\beta} = (16, 1, 8)$	$\boldsymbol{\gamma} = (-2.44, 0.41, -1.45)$ $\boldsymbol{\beta} = (16, 1, 8)$	$\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$ $\boldsymbol{\beta} = (16, 1, 8)$

the Bayesian simulation, let λ be either 10 or -5 and σ be either 1 or 5.

Errors occurred in the r^* algorithm for some of the simulations. The errors that occurred were mostly due to issues with the profile log likelihood that we have previously discussed. There were also problems with the optimization routine as well as with singularity in the information matrices. There were errors in every simulation where $\pi_i = 0.02$, consistent with the results from Section 3.7.2. A few of the simulations where $\pi_i = 0.05$ were successful, however none of those three design points ran all four combinations of λ and σ without error. As a result, we focus our discussion of the simulation and its results on design points three and nine, where $\pi_i = 0.20$ and $\mu_i = 5$ and 20, respectively. Each of these design points completed all four of the λ and σ combinations.

As before, let x_{1i} be a standardized age covariate and let x_{2i} be gender, with equal probability of men and women. We will generate a sample of size $n = 600$ for $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. We will compute π_i for each subject using the current γ values and generate the 600 values of y_1. Finally, we use x_{1i}, x_{2i}, π_i, and y_{1i}, along with the current parameter values to calculate each of the μ_i which can be used to generate each of the y_{2i}.

Once the data is generated, we use x_{1i}, x_{2i}, y_{1i}, and y_{2i} to implement the r^* method. This process, of generating the data then estimating parameters, is

replicated 50 times for each of the 36 simulations. Note that we have used the same seeds as in the simulation for the Bayesian model in Section 2.4, allowing us to compare the results.

Table 3.4 shows the results from the simulation for the r^* method for DP3 with $\lambda = 10$ and $\sigma = 1$. We see that the average MLE is very close to the true value for each of the eight parameters. The widths of all eight intervals is fairly narrow as well. Note that, for this design point, we targeted the largest of our three values of $\pi_i = 0.20$, as well as $\mu_i = 5$. Across all simulation results when $\pi_i = 0.20$ we have only three parameters with an average MLE around 0.15 away from its true value and the vast majority of the parameters are within 0.05 of their true values. The simulations that target $\pi_i = 0.05$ have parameters with average MLE values that are up to 0.26 away from their true values, with a larger number of parameters with average MLE values further from their true values. This supports the conclusion that the estimation of the model has more difficulty as π_i gets smaller. Also of interest in Table 3.4 are the coverage percentages for each parameter. Note that the coverage of both λ and σ are very good here, at 96%, and they are consistently in the 80's and 90's across all simulations. The other six parameters are a different story. The regression parameters for both safety and efficacy have exceedingly bad coverage values with the best coverage coming for β_1 at 46% and the worst for γ_2 at 26%.

Table 3.5 shows the simulation results for DP3 with $\lambda = 10$ and $\sigma = 5$. The only distinction between the simulations for Tables 3.4 and 3.5 is the change from $\sigma = 1$ to $\sigma = 5$. There are two notable differences in the results. The most obvious change occurs in the coverage of the parameters. The coverage has increased for the larger value of σ for the safety and efficacy regression parameters by an average of 40 to 50 percentage points. This is due, in part, to the other difference between the simulation results, which is the widths. The widths are 3 to 4 times wider when

Table 3.4: Simulation results for the r^* method for DP3 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.91	3.02	3.12	0.21	0.42
β_1	1.33	1.22	1.32	1.42	0.20	0.46
β_2	4.00	3.82	3.97	4.12	0.30	0.32
γ_0	-1.73	-1.81	-1.75	-1.68	0.13	0.36
γ_1	0.11	0.04	0.10	0.16	0.12	0.44
γ_2	0.64	0.56	0.64	0.72	0.16	0.26
λ	10.00	9.82	10.03	10.23	0.41	0.96
σ	1.00	0.95	1.00	1.06	0.11	0.96

Table 3.5: The simulation results for the r^* method for DP3 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.57	3.02	3.46	0.89	0.82
β_1	1.33	0.89	1.31	1.73	0.83	0.94
β_2	4.00	3.36	4.00	4.62	1.26	0.88
γ_0	-1.73	-1.90	-1.72	-1.54	0.35	0.82
γ_1	0.11	-0.06	0.09	0.25	0.31	0.90
γ_2	0.64	0.41	0.64	0.86	0.44	0.78
λ	10.00	8.95	9.96	10.96	2.01	0.90
σ	5.00	4.69	4.97	5.25	0.56	0.92

$\sigma = 5$ as compared to $\sigma = 1$. This says that the simulation variability in the MLE routine is such that when the width's are smaller the r^* method cannot capture the true value of the parameters. This is a trend we see across all the simulation results. We also notice a similar trend, although not as drastic, when we change the magnitude of λ, but in the opposite direction. When λ has a large magnitude ($\lambda = 10$) the coverage is, on average, 10 to 20 percentage points lower than when λ has a smaller magnitude ($\lambda = -5$). This is also due to a difference in widths, but in this case the safety regression parameters are the only widths affected by the change in magnitude.

Figure 3.3 summarizes the simulations for the four combinations of λ and σ for DP3. The horizontal line is at 3, the true value of β_0. The four box plots represent the four combinations of λ and σ. The four points are at the average MLE value

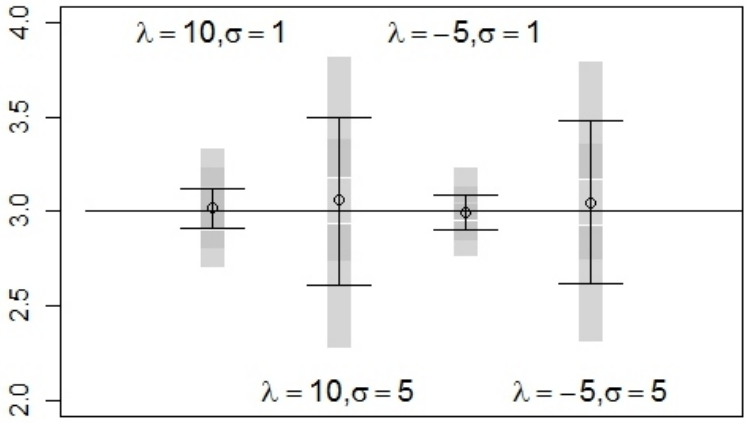

Figure 3.3: Spread of 95% confidence intervals for β_0 in DP3.

of the 50 replications for each simulation. The upper and lower boundaries of the plot are the average of the 97.5% and 2.5% confidence bounds, respectively, and the three grey boxes extend one simulation standard deviation above and below each of the average MLE, 2.5%, and 97.5% estimates. We see the difference in widths between when $\sigma = 1$ and $\sigma = 5$. The difference in widths between the two values of λ is difficult to see in this figure, however Figure 3.4 shows this more clearly for γ_0 in DP3. Figure 3.5 shows the box plots for all of the r^* simulations where $\lambda = 10$. We can see that the width of the 95% confidence intervals and the simulation variability of the estimates increases as σ increases as well as when π_i decreases. Finally, the estimation of σ does not appear to be affected by the values of the other parameters. We will compare the Bayesian simulation from Section 2.4 with the r^* simulation in Section 3.9.

3.9 Comparison of the Bayesian Simulation to the r^* Simulation

Chapter Two presents the Bayesian development for model (3.1). We are interested in knowing how the Bayesian model compares to the frequentist r^* model.

The simulation studies from Chapters Two and Three used the same data, so the comparisons between the results can be made directly. Recall that the r^*

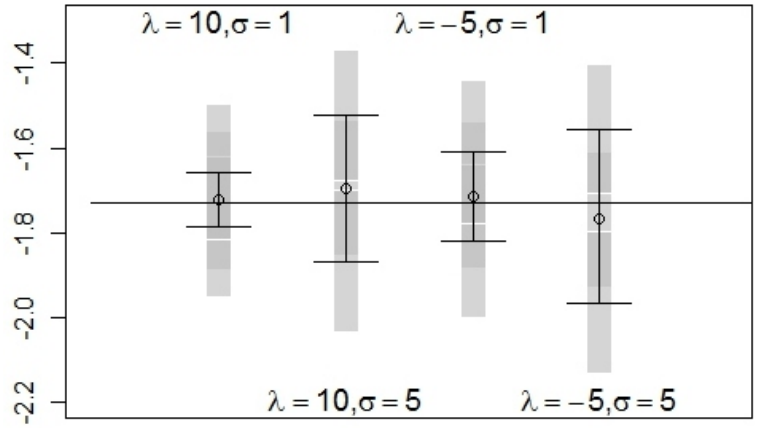

Figure 3.4: Spread of 95% confidence intervals for γ_0 in DP3.

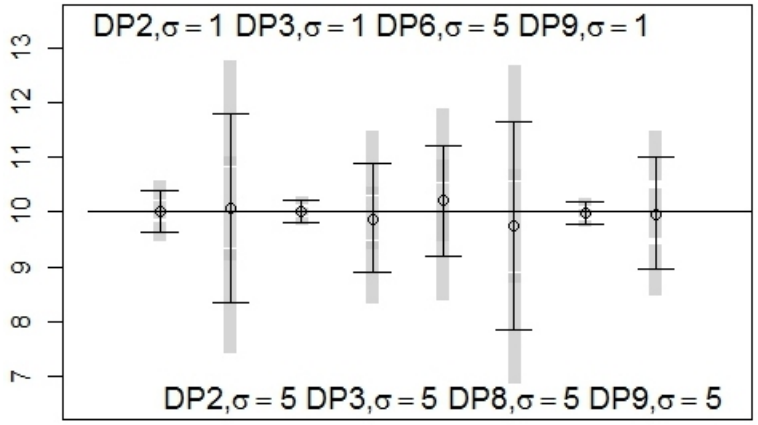

Figure 3.5: Spread of 95% confidence intervals for $\lambda = 10$.

simulations resulted in errors in certain areas, specifically where $\pi_i = 0.02$ and $\mu_i = 10$. The results for the Bayesian simulation are more complete, but we will only be able to compare the methods where we have both sets of results.

Figure 3.6 illustrates the simulation results for both the Bayesian and r^* methods. The figure is similar to the plots from Sections 2.4 and 3.8. The horizontal line is still the true value of the parameter and each box plot depicts either the average posterior mean and 95% credible set or the average MLE and 95% confidence interval along with the simulation variability for each respective simulation. The main difference between Figure 3.6 and the previous figures is that each label of the λ and σ combination refers to two box plots with the plot from the Bayesian simulation on the left and the r^* simulation on the right. The most obvious thing to notice on Figure 3.6 is that the r^* simulations have consistently narrower intervals. This is a trend in the results for all of the efficacy and safety regression parameters. We do notice, however, that even with narrower intervals the simulation variability for the MLE and 95% confidence intervals is, at best, similar to that of the average posterior mean and 95% credible sets. This is easy to see in Figure 3.6 when $\lambda = 10$ and $\sigma = 1$ because of the overlap in the grey bars. The simulation variability for the r^* method also occasionally shows up slightly larger than that of the Bayesian method.

Figure 3.7 compares the results across DP3 and DP9 when $\lambda = -5$. Note that DP3 and DP9 both target $\pi_i = 0.20$, resulting in $\boldsymbol{\gamma} = (-1.73, 0.11, 0.64)$. Figure 3.7 is interesting because the simulation results appear almost identical for both methods. The average posterior means, average 95% credible sets, and even the simulation variability for each simulation look very similar to the average MLE's and average 95% credible sets. The results for λ in other design points are similar to what we see here.

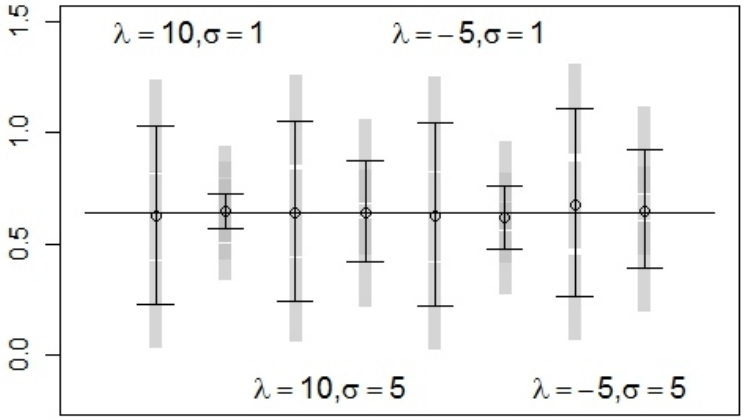

Figure 3.6: Comparison of simulation results for the Bayes and r^* methods where the plots are alternating between the two methods starting with the Bayesian results on the far left. Results are for γ_2 in DP3.

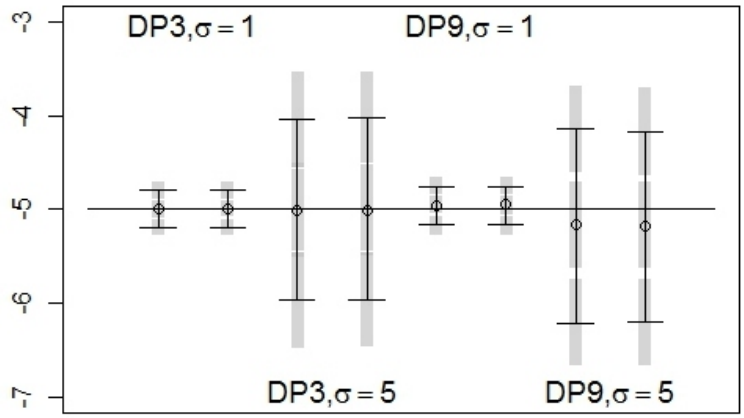

Figure 3.7: Comparison of simulation results for the Bayes and r^* methods where the plots are alternating between the two methods starting with the Bayesian results on the far left. Results are for $\lambda = -5$ across DP3 and DP9.

Table 3.6: Comparison of coverage for the Bayesian and r^* approaches on DP3 with two combinations of λ and σ.

(a) $\lambda = 10$ and $\sigma = 1$

Parameter	Bayesian Coverage	r^* Coverage
β_0	0.90	0.42
β_1	1.00	0.46
β_2	0.94	0.32
γ_0	0.94	0.36
γ_1	0.96	0.44
γ_2	0.96	0.26
λ	0.96	0.96
σ	0.98	0.96

(b) $\lambda = 10$ and $\sigma = 5$

Parameter	Bayesian Coverage	r^* Coverage
β_0	0.96	0.82
β_1	0.94	0.94
β_2	0.98	0.88
γ_0	0.94	0.82
γ_1	0.98	0.90
γ_2	0.94	0.78
λ	0.90	0.90
σ	0.98	0.92

The results for σ have similar interval widths for both methods. The average posterior means and the MLE's do not line up as closely as they do for λ.

We can also compare the average coverage values for the two methods. Table 3.6 gives side-by-side comparison of the coverage for two combinations of λ and σ in DP3. Table 3.6a depicts the comparison when $\lambda = 10$ and $\sigma = 1$. For the regression parameters in both the safety and efficacy portions of the model, the Bayesian method is vastly superior to the r^* method. We also notice that Table 3.6b, where $\lambda = 10$ and $\sigma = 5$, shows a big improvement in the coverage for the r^* method, however the Bayesian method still has a clear advantage. The only difference between the two tables is the change from $\sigma = 1$ to $\sigma = 5$. This dependence on the size of σ is a trend we see throughout the simulation results, which was mentioned in Section 3.8. The coverage for λ and σ is comparable in both methods, although across all simulations the Bayesian method appears to have a slight advantage.

We have compared the simulations for the Bayesian and r^* approaches to our bivariate model. The r^* approach has consistently smaller widths for the 95% intervals but does not have improved simulation variability. The coverage for the r^* method is extremely dependent on the magnitude of σ. Even when the coverage

for the r^* method is better, the Bayesian approach has visibly better values. In practice, this would lead to a much higher type I error for the r^* approach. Thus, we recommend the Bayesian method.

3.10 Discussion

Chapter Three looked at frequentist methods to the bivariate discrete and continuous model. We introduced the p^* method and discussed its usage as well as why we have not implemented it here. We then introduced the r^* method. We looked at some specific examples of the r^* method and compared some of those examples to those used in the Bayesian approach. We then implemented a full scale simulation using the r^* method and compared it to the Bayesian simulation from Section 2.4. The results indicated that the Bayesian method should be preferred.

CHAPTER FOUR

Bayesian Power Study for Multiple Testing in Mixed Responses

4.1 Introduction

Assessing the operating characteristics of a statistical model is an important step in determining the usefulness of the model. In particular, type I error rates and power are two of the more important characteristics. There is a vast literature on all aspects of operating characteristics for a wide variety of statistical models and inferential methods.

Operating characteristics such as power and type one error rate are, of course, repeated sampling constructs, averaging, as it were, over the sample space. Bayesian methods condition on the data, averaging over the parameter space. Nevertheless, operating characteristics are important to Bayesian practice in some contexts. As it happens, Bayesian methods often have very good frequentist properties (Carlin and Louis, 2009).

We are particularly interested in the regulatory context in which the use of Bayesian methods in biopharmaceutical research must be accompanied by a thorough examination of operating characteristics. See, for example, the FDA's guidance on the use of Bayesian statistics in medical device trials (FDA, 2010), which was issued in draft form in 2006, and finalized in 2010. Pennello and Thompson (2007) also give insight into regulatory requirements for submissions using Bayesian analysis, specifically in the area of medical devices. In addition, Rubin (1984) talks about the use of 'frequency calculations' when performing a Bayesian analysis. See also Lee and Chu (2012) and Winkler (2001).

An operating characteristic of particular concern in multiple testing problems is the family-wise error rate (FWER). Suppose, for example, we are interested in

testing five endpoints in a frequentist analysis at the 5% type I error level for each endpoint, then the probability of obtaining at least one erroneous significant result can be almost 23%. This 23% is called the family-wise error rate and we must be concerned with controlling this error. Three commonly used frequentist strategies for controlling FWER include multiplicity adjusted significance levels, multiplicity adjusted p-values, and simultaneous confidence intervals (Dmitrienko et al., 2010, p. 46).

An alternative method of controlling the FWER in a Bayesian analysis is through exchangeability as discussed, for example, in Gelman et al. (2004). The FDA guidance also discusses the role of exchangeability (see p. 30) as a possible Bayesian approach to testing multiple endpoints.

As noted, there are several frequentist approaches to FWER problems, including the multiplicity adjusted significance levels described in Dmitrienko et al. (2010). In view of the fact that Bayesian methods often exhibit good frequentist properties, we wish to investigate the application of these multiplicity adjustment methods to Bayesian modeling—specifically to the joint statistical model studied in Chapter Two. We discuss two of these methods, the fixed-sequence procedure and the fallback procedure, in Sections 4.3.1 and 4.3.2 respectively.

In this chapter, we begin in Section 4.2 by redefining the Bayesian regression model and constructing a Bayesian power simulation structure, analogous to the sample size simulation used by Stamey et al. (2013). Section 4.3 discusses multiple testing and some of the methods used to adjust for multiplicity, including the fixed-sequence and fallback procedures. Section 4.4 presents the simulation and discusses the results. We finish in Section 4.5 with a discussion.

4.2 Bayesian Power

We will continue to use the mixed response bivariate regression model introduced in Chapter Two and continued in Chapter Three. Define

$$Y_{1i} \sim \text{Bernoulli}(\pi_i), \tag{4.1}$$

and

$$Y_{2i}|Y_{1i} \sim N(\mu_i, \sigma^2),$$

with link functions

$$\text{logit}(\pi_i) = \gamma_0 + \gamma_1 x_{1i} + \gamma_2 x_{2i},$$

and

$$\mu_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \lambda(y_{1i} - \pi_i).$$

Wang and Gelfand (2002) develop a Bayesian sample size determination method. They are interested in finding the sample size, n, to optimize endpoints such as the average coverage, average power, or average length of the posterior interval for specific parameters. The method involves specifying two sets of priors rather than one, as in a traditional Bayesian analysis. The 'design' or 'sampling' priors are elicited to represent the fixed planning estimates that are chosen for a traditional sample size determination. These probability distributions are chosen so that they allow for some uncertainty in the parameter values used to generate the data. The data are generated using parameters from the design priors. The process is repeated for each iteration in the simulation. The 'analysis' or 'fitting' priors are those used in a conventional Bayesian analysis. After the data are generated using the parameters taken from the design priors, the analysis priors are used in the Bayesian model. In practice, the analysis priors can be relatively non-informative but the design priors must include some information. A Bayesian sample size determination utilizes the

steps previously mentioned for various sample sizes until a sample size can be found that achieves the endpoint of interest.

Stamey et al. (2013) implement a Bayesian sample size determination on (4.1), both for the case where Y_2 is dependent on Y_1, and also the case where Y_1 and Y_2 are independent. The independence is achieved by removing $\lambda(y_{1i} - \pi_i)$ from the right side of the $Y_{2i}|Y_{1i}$ link function in (4.1). Stamey et al. (2013) focus on comparing the performance of the independent and dependent models for a treatment indicator in the portion of the model connected to the continuous response. They look at the average length criterion (ALC), average coverage criterion (ACC), average power criterion (APC) and type I error rates. They also briefly look at the APC and type I error for a simultaneous hypothesis test on the treatment indicator for both the binary and continuous responses. We will extend their work by comparing the power of their simultaneous hypothesis test while using a Bonferroni adjustment for multiplicity to the power achieved when using two methods to control the FWER.

Following Stamey et al. (2013) we use the following approach to calculate the power:

(1) Specify the analysis priors for all parameters and the design priors for all parameters except those of interest, the binary and continuous treatment covariates.

(2) Specify the effect size of interest for the parameters of interest. The effect size for the continuous treatment regression parameter will be a real number where larger numbers indicate larger improvement. The effect size for the binary treatment regression parameter will be specified as a real number that can be translated through the logit transformation into a change in the probability of a binary response for the baseline levels of the other binary regression parameters.

(3) For $l = 1, 2, ..., M$ Monte Carlo iterations, at each sample size:

   (a) Generate values of regression and variance parameters from their design priors.

   (b) Simulate continuous covariates from normal distributions (or suitable alternatives) over expected ranges. Similarly, simulate binary covariates from Bernoulli distributions.

   (c) Calculate π_i and μ_i from model (4.1) and generate y_{1i} and y_{2i} for $i = 1, 2, ..., n$ using the parameters generated from the design priors and the simulated covariates.

   (d) To each simulated data set generated in Step 3c, fit the Bayesian model using the analysis priors. A Gibbs sampler can be employed at this step.

   (e) Calculate the posterior probability that the continuous treatment parameter exceeds 0 (or other threshold of interest). Similarly, calculate the posterior probability that the binary treatment parameter exceeds 0. Calculate an indicator of acceptance or rejection based on a simultaneous hypothesis test or some multiplicity adjusted variant for the FWER of interest, α.

(4) Fit a curve or surface through the Bayesian power values and find an adequate sample size combination for the desired power.

Suppose for model (4.1) that we consider a clinical trial where Y_1 is a safety indicator, that is, let $y_1 = 1$ if the subject experiences an adverse event and 0 otherwise. Also, let Y_2 be a continuous efficacy variable. This is similar to the situation discussed in Stamey et al. (2013). We are interested in simultaneous testing of the treatment parameters, γ_1 and β_1. Let H_1 correspond to hypothesis

$$\begin{aligned} H_0 &: \gamma_1 = 0 \\ H_A &: \gamma_1 > 0, \end{aligned} \tag{4.2}$$

and H_2 correspond to hypothesis

$$\begin{aligned} H_0 &: \beta_1 = 0 \\ H_A &: \beta_1 > 0. \end{aligned} \tag{4.3}$$

Note that H_1 tests if $\gamma_1 > 0$ which corresponds to a subject with the treatment having a smaller probability of safety than a subject without the treatment. We reject the simultaneous hypothesis test if $P(\gamma_1 > 0) \geq \alpha_1$ and $P(\beta_1 > 0) \geq \alpha_2$, for some significance levels α_1 and α_2. In the safety and efficacy scenario this corresponds to a significant increase in the probability of an adverse event due to treatment and a significant increase in efficacy due to treatment. In the Bayesian context we compare the marginal posterior probabilities that $\gamma_1 > 0$ and $\beta_1 > 0$ to the significance level(s) of interest. We discuss methods of controlling the FWER under multiplicity in the following section.

4.3 Multiple Testing

Testing of multiple endpoints is common in the literature. Perlman and Wu (2004) discuss testing of multivariate endpoints with one-sided alternative hypotheses, Gönen et al. (2003) introduce an Bayesian algorithm for simultaneous testing of two-sample multivariate endpoints, and Pocock et al. (1987) give an overview of the use of multiple testing in clinical trials. In our case, we are interested in simultaneous testing of our two treatment parameters, but are concerned with controlling the FWER as well.

One method for controlling the FWER is through a Bonferroni adjustment, originated by Dunn (1961). The Bonferroni correction is one of the most widely used multiplicity adjustment methods, but is also well known to be very conservative, especially for large numbers of endpoints. While the use of a Bonferroni correction is not necessary in the Bayesian analysis of a single data set, power, coverage, and other operating characteristics may benefit from its application across repeated samples.

There are several methods that attempt to avoid the overly conservative performance associated with Bonferroni adjustments. For an overview, see Dmitrienko et al. (2010), Chapter Two. The main focus of their discussion is the use of 'stepwise' adjustments such as the Holm procedure and variations therof, like the Schaffer procedure. Similarly, they discuss a variety of extensions to the Bonferroni procedure. Most of the methods that Dmitrienko et al. (2010) discuss, determine the order of the hypotheses to be tested a-posteriori, for example, using the magnitude of the p-value. Procedures such as the fixed-sequence, fallback, and reverse fixed-sequence procedures are the exception. These procedures all specify an order of the hypotheses to be tested beforehand. The order of testing may correspond to clinical importance, such as by concentration in a dose finding study. This seems to be intuitive and we choose the fixed-sequence and fallback procedures to test against the Bonferroni approach.

The use of the fixed-sequence and fallback procedures on a Bayesian analysis, described in Sections 4.3.1 and 4.3.2, respectively, can be justified in a similar manner to the Bonferroni approach. In our simulation we will compare power using a Bonferroni adjustment, to that of the fixed sequence procedure and the fallback procedure.

4.3.1 The Fixed-Sequence Procedure

Maurer et al. (1995) introduced the fixed-sequence testing approach. Westfall and Krishen (2001) use the fixed-sequence approach, among others, in their paper. We follow the development of the fixed-sequence procedure in Dmitrienko et al. (2010, p. 56). Specify the order of our hypotheses to be tested as $H_1, ..., H_m$, a-priori. The fixed-sequence procedure starts by testing H_1 and carries on sequentially without a multiplicity adjustment as long as H_j, $j = 1, ..., m$ is rejected at the jth step. We reject H_k if $p_k \leq \alpha, \quad k = 1, ..., j$, where α is our FWER and p_k is the

posterior probability against H_k. Consider hypotheses (4.2) and (4.3). If H_1 is tested first, then we reject H_1 if the posterior probability that $\gamma_1 > 0$ is greater than α. If this is the case, then we can move on to test H_2, that $\beta_1 > 0$, also using the posterior probability at the α level. Thus, we test hypotheses sequentially until we find the first non-significant hypothesis and we fail to reject every hypothesis following the non-significant one. This stoppage is the reason the FWER is controlled in the fixed-sequence procedure. The order of the hypotheses is incredibly important because if a non-significant result occurs then none of the following hypothesis are even tested.

4.3.2 The Fallback Procedure

Dmitrienko et al. (2010, p. 57) also discuss the implementation of the fallback procedure. For this procedure we will continue to have ordered hypotheses $H_1, ..., H_m$. We are also still concerned with controlling α, the FWER. We can specify weights $w_1, ..., w_m$ that correspond to each hypothesis such that $\alpha_j = \alpha w_j$ assigns a portion of the FWER to each hypothesis. When $w_j = w_k = \alpha/m$ for all $j \neq k$ then the error rate allocation is equal to that of the Bonferroni adjustment. The method follows m steps. In step one we test H_1 at the $\alpha_1 = \alpha w_1$ level. If $p_1 \leq \alpha_1$ then we reject H_1. Steps $k = 2, ..., m$ proceed in order by testing H_k at $\alpha_k = \alpha_{k-1} + \alpha w_k$ if H_{k-1} is rejected, and at $\alpha_k = \alpha w_k$ otherwise. Again, if $p_k \leq \alpha_k$, we reject H_k. This procedure will end up testing every one of the ordered hypotheses, but the significance level will change depending on the acceptance or rejection of the preceding hypotheses. The ability to test every hypothesis is an advantage over the fixed-sequence procedure. The fallback procedure simplifies to the fixed-sequence procedure when $w_1 = 1$ and $w_2 = \cdots = w_m = 0$. For the power simulation in Section 4.4 we will set $w_1 = w_2 = \alpha/2$.

4.4 Simulation

We follow a similar development to that of the example in Stamey et al. (2013). Their example is based on an actual clinical trial. Let $y_1 = 1$ be an indicator of the occurrence of an adverse event and let y_2 be a continuous efficacy response. Let $x_1 = 1$ indicate that the subject was given the treatment rather than the placebo and let x_2 be some continuous covariate. Table 4.1 gives the design and analysis priors for each of the parameters. We will be looking exclusively at testing γ_1 and β_1 simultaneously, so we set degenerate design priors at $\gamma_1 = 0.85$ and $\beta_1 = 75$. Notice that the baseline value for γ_0 will generally stay between -0.8 and 0 which corresponds to adverse event probabilities between 0.31 and 0.5. There are three different analysis priors for both γ_1 and β_1. These correspond to optimistic, reference, and skeptical priors for each parameter. The reference and skeptical priors for γ_1 were not included in Stamey et al. (2013), but have been added because we are interested in the simultaneous test. The optimistic priors have means centered at the effect of interest with standard deviations that place some information around that point, as might be available from a previous study. The reference priors are centered at 0 with large 'non-informative' standard deviations. The skeptical priors are centered at 0 and have standard deviations that are informative. The skeptical priors are designed to support the null hypothesis of no effect.

We have generated 1000 replications of the parameter values and data sets for various sample sizes from 50 up to 400. The Bayesian model was fit using WinBUGS through *R*. We kept 4500 iterations after a burn-in of 3000 iterations. Thinning was set at 2 to avoid autocorrelation.

We have recorded the power of the Bayesian approach to individually detect when γ_1 and β_1 are greater than 0. We have also recorded when γ_1 and β_1 are greater than 0 simultaneously using the Bonferroni adjustment, the fixed-sequence procedure, and the fallback procedure, both when γ_1 and β_1 are tested first. Our

Table 4.1: Design and analysis priors and covariate distributions.

Parameter	Design Prior	Analysis Prior
γ_0	$N(-0.4, 0.2)$	$N(-0.4, 1.00)$
		$N(0.85, 1.00)$
γ_1	0.85 (fixed value)	$N(0, 1.00)$
		$N(0, 30)$
γ_2	$N(0.1, 0.05)$	$N(0.1, 1.00)$
β_0	$N(1200, 100)$	$N(1200, 316)$
		$N(75, 69)$
β_1	75 (fixed value)	$N(0, 69)$
		$N(0, 3162)$
β_2	$N(15, 5)$	$N(15, 25)$
λ	$N(400, 30)$	$N(400, 408)$
σ	Uniform(140, 160)	Uniform(0, 800)
X_1	Bernoulli(0.5)	NA
X_2	Uniform(3, 23)	NA

interest is in the simultaneous test as might be the case in a clinical trial. As a result, we have not recorded the results of the individual hypotheses for each of our methods. Thus, the fixed sequence procedure will have the same result, regardless of the order the hypotheses are tested, because all hypotheses are tested at the FWER, α, until one hypothesis fails to be rejected. We then fail to reject the remainder of the tests as well. On the other hand, the fallback procedure adjusts the individual error rates throughout the testing process, and results in different powers when changing the order of the hypotheses. The FWER will be controlled at $\alpha = 0.05$. Note that we have specified the weights for the fallback procedure as $w_1 = w_2 = \alpha/2$ for each ordering of the hypotheses.

Figure 4.1 plots the powers for various sample sizes using the optimistic priors for the two treatment parameters individually along with the simultaneous power using the Bonferroni adjustment. Notice that the individual powers for both the efficacy and safety parameters individually are close enough to be indistinguishable. Also, notice that there is a significant drop off from the individual powers to the simultaneous power. This occurs because the individual powers are tested at the 95% significance level, and the Bonferroni adjustment requires that we test each hypothesis at the $1 - 0.05/2 = 97.5\%$ significance level. Note that this drop off between the Bonferroni power and the individual powers also follows across the simulation results for the reference and skeptical priors also.

Figure 4.2 shows the powers across various sample sizes for our four simultaneous methods using the optimistic priors. Note that the line labeled 'Fallback Efficacy' refers to the fallback method when testing the efficacy treatment parameter first. Similarly the fallback method when testing the safety treatment parameter first is labeled 'Fallback Safety.' Notice that the Bonferroni adjustment has the worst power of the four methods. This is a result of the conservative nature of the method. The fixed-sequence method has better power than both of the fall-

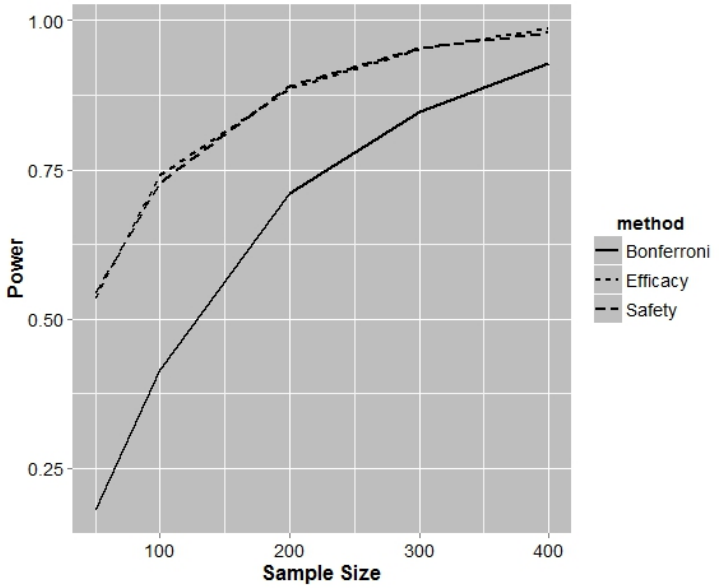

Figure 4.1: The individual powers across sample sizes for the efficacy and safety treatment parameters plotted with the simultaneous power using the Bonferroni adjustment. The optimistic priors were used for this plot.

back methods. This is intuitive because the fixed-sequence procedure by definition does not make an adjustment to the FWER. Thus, because of the adjustment in the fallback procedure, the fixed-sequence will perform as well or better than the fallback method every time. This is important to note because if interest is in the simultaneous hypothesis test exclusively and not in the individual hypotheses, then the fixed-sequence procedure will always produce better results than the fallback method. The two fallback methods perform similarly in this case. Power is slightly higher when safety is tested first. As the sample size increases though, the two fallback methods become essentially equivalent. As with the previous figure, the results using the reference and skeptical priors follow the same trends as are seen here.

Figure 4.3 shows the results for the fixed-sequence procedure across the optimistic, reference, and skeptical priors. Notice that there is a definite improvement from the skeptical prior to the reference prior and then again from the reference prior to the optimistic prior. The improvement starts at approximately 10% between each

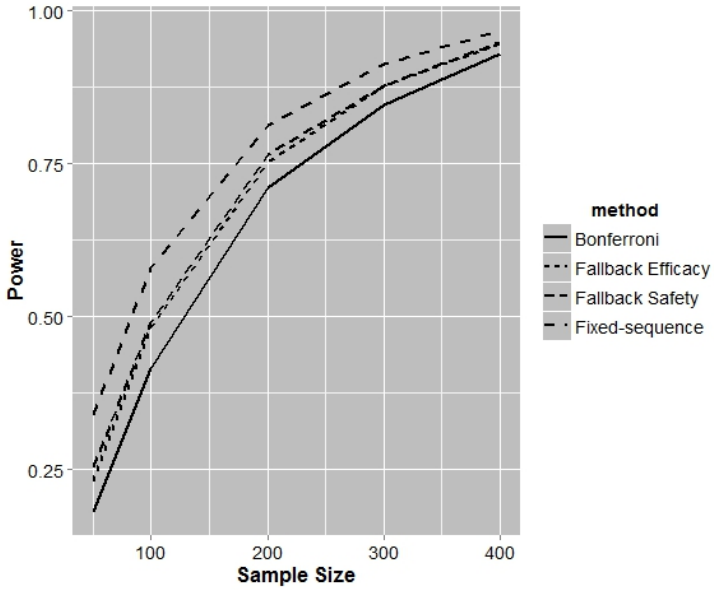

Figure 4.2: Comparison of the four simultaneous methods for the optimistic priors.

of the skeptical and reference priors as well as the reference and optimistic priors. The improvement steadily decreases as the sample size increases. These results are intuitive because we would expect that using prior information should provide an improvement over a so-called 'non-informative' prior such as the reference prior. Similarly, we would hope that placing an informative prior over the null hypothesis would weight the results toward the null, even when the null is false. Thus, we should incorporate prior information into our prior distribution, if possible, to improve the chances of rejecting a false null hypothesis. The differences between the prior types show similar results across each of the different simultaneous methods.

In Chapter Two, as the probability of an adverse event in the safety regression model decreased, the model had a more difficult time estimating the safety regression parameters. This is seen in the increase in simulation variability. For the power study, we hypothesized that, if the baseline probability of an adverse event in the safety regression model were to be more rare, then the power of the test would decrease. So far the results have all used the design prior for γ_0 that is listed in

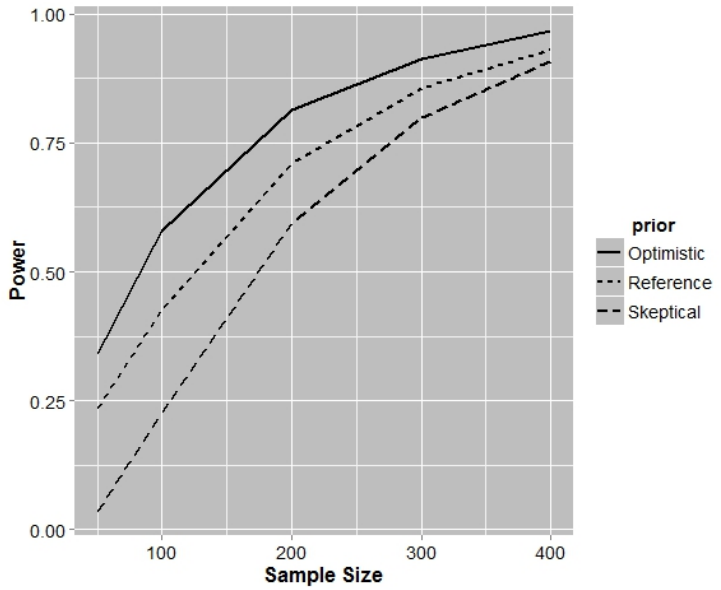

Figure 4.3: Comparison of the optimistic, reference, and skeptical priors using the fixed-sequence method.

Table 4.1 at $N(-0.4, 0.2)$. This puts the baseline probability of an adverse event approximately between 0.31 and 0.5. We ran simulations similar to those mentioned previously but we changed the design prior for γ_0 to $N(-1.2, 0.2)$ and $N(-2.0, 0.2)$. These new design priors correspond to baseline values between 0.16 and 0.31 and then 0.08 and 0.16, respectively. The results using the fixed-sequence method for the three different design priors are shown in Figure 4.4. Performance is similar across sample sizes for all three design priors, with none yielding a clear advantage. This makes sense because, as we saw in Chapter Two, there was little difference in coverage when simulation variability increased due to decreased adverse event probabilities.

Finally, Figure 4.5 is a plot of the individual powers and simultaneous Bonferroni adjustment when the design prior for γ_0 is $N(-1.2, 0.2)$, similar to Figure 4.1. This is interesting because the individual efficacy and safety powers show an obvious difference, when there was no difference previously. In this case the power for the individual safety parameter has increased approximately 4% from the original design

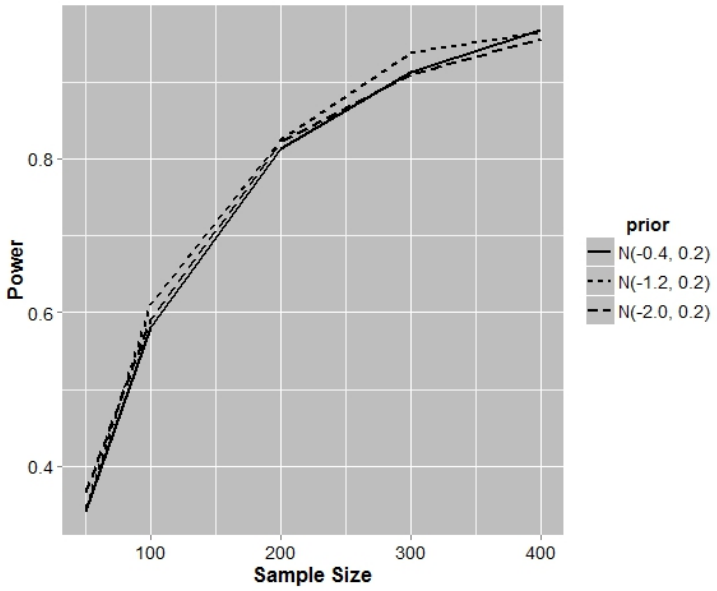

Figure 4.4: Comparison of the three design priors for γ_0.

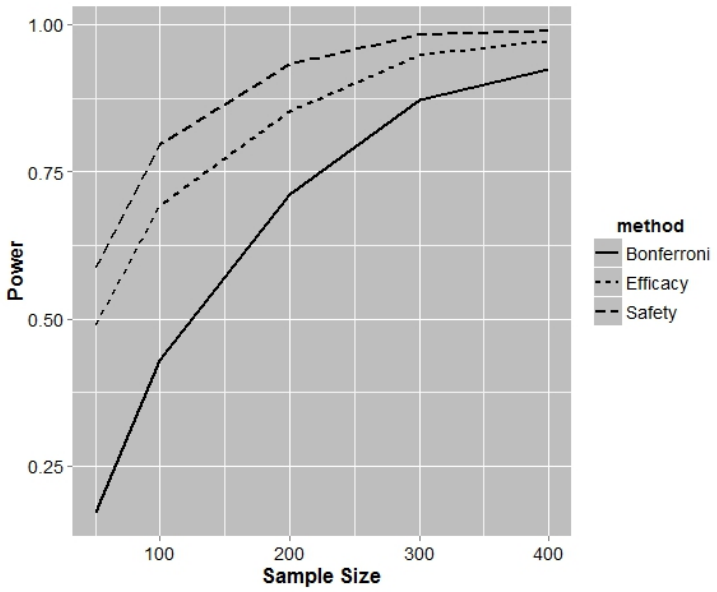

Figure 4.5: The individual powers for the efficacy and safety regression parameters as well as the power for the simultaneous Bonferroni adjustment when the design prior for γ_0 is $N(-1.2, 0.2)$.

prior at $N(-0.4, 0.2)$ and the power for the individual efficacy parameter has decreased approximately 4%. As we saw in Figure 4.4 the resulting simultaneous tests are approximately equal using the fixed-sequence approach. These results would become more important if the fallback procedure was being used. When the power for safety increases and efficacy decreases, as it does here, the fallback procedure when testing safety first increases its advantage over the fallback procedure when testing efficacy first. Thus, in practice, a knowledge of what the individual powers are doing in relation to each other should suggest which procedure to use in the simultaneous test. (We have tabulated these results by method and sample size and they are included in Appendix D.)

4.5 Discussion

In this chapter we have considered multiplicity adjustment methods in a Bayesian sample size determination simulation in order to explore the operating characteristics of controlling the FWER. We introduced the Bayesian sample size determination scheme, following that of Stamey et al. (2013). We discussed various multiplicity adjustment methods and the validity of using them in a Bayesian analysis. We then introduced the fixed-sequence and fallback procedures and performed the simulation in order to determine their effectiveness relative to the Bonferroni adjustment. The results suggest that the fixed-sequence procedure will save the most power while controlling the FWER if we are interested solely in the simultaneous test rather than the individual hypotheses. We also saw that the incorporation of any prior information available will save power over 'non-informative' or skeptical priors. We finished by noting that adjusting the baseline adverse event probability does not seem to have an effect on the simultaneous power using the fixed-sequence method, but that the individual efficacy and safety powers are affected which corresponds to an effect on the fallback procedures.

CHAPTER FIVE

Conclusions

In the dissertation we consider the bivariate model with one binary and one continuous response using the marginal and conditional model specification. We condition the continuous response on the binary response in order to incorporate a link between the two models that allows for some correlation between the two responses. We implement a Bayesian approach to model the bivariate regression model and compare it to a frequentist, higher-order asymptotic approach. Finally, we extend a previous Bayesian sample size determination study by improving the operating characteristics of the model using multiplicity adjustments.

In Chapter Two we introduce the bivariate discrete and continuous regression model. We examine the performance of a Bayesian approach to estimating the parameters in the model using a simulated example in a biopharmaceutical context where the binary response is some safety indicator and the continuous response is some efficacy variable. The simulation indicates that the Bayesian model estimates the parameters well. The simulation also shows increased simulation variability in the estimates when the probability of a safety event approaches zero. We discuss the possibility that an induced prior on the adverse event probability is affecting the estimates.

In Chapter Three we look at a frequentist method of estimation using higher-order approximations, which is based on the likelihood ratio, called the r^* method. This approximation is built for estimation of models that do not meet the normality assumptions that are required for methods such as the Wald and score statistics. We chose the r^* method over another higher-order approximation, the p^* approach, because the r^* method affords convenient elimination of nuisance parameters. The

r^* approximation also appears to estimate the parameters fairly well. We conclude Chapter Three with a comparison of the simulation results using the Bayesian approach to those using the r^* approach. Both models appear to be sensitive to small safety probabilities, sometimes prohibitively so in the case of the r^* method. This is due to an inconsistency in the profile information matrix that results in an error in the estimation algorithm. When we do obtain results, the r^* method results in 95% confidence intervals that are slightly narrower than the 95% Bayesian credible sets. However, the MLE's also exhibit the increased variability that we saw for the Bayesian approach in Chapter Two and combined with the narrower intervals, the coverage of the 95% intervals in the r^* approach is much worse than the Bayesian approach. We conclude by recommending the use of the Bayesian method for this model.

In Chapter Four we extend a Bayesian sample size determination study using the bivariate binary and continuous model. We hypothesize that interest is in both safety and efficacy and is incorporated through a joint treatment endpoint. We are interested in the operating characteristics of the model, specifically the power, as might be useful in a regulatory setting. Although Bayesian inference does not generally use methods based on repeated measurements, in this case, we claim that this approach is appropriate through simulation in order to satisfy the requirements of a regulatory submission. We use multiplicity adjustment techniques such as the fixed-sequence and fallback approaches to improve the power of the bivariate model. These two approaches are appropriate when the hypotheses of interest are ordered *a-priori*. The results suggest that using methods for multiplicity yield an increase in power over the overly conservative Bonferroni adjustment with the fixed-sequence method showing the most improvement. We also see that incorporating any prior information will save power over either 'non-informative' or skeptical priors. Finally, we show that an adjustment to the baseline adverse event probability affects the

individual safety and efficacy parameter powers. This results in an adjustment using the fallback method, dependent upon which parameter is tested first, but does not seem to affect the fixed-sequence procedure.

APPENDICES

APPENDIX A

Selected Simulation Results for the Bayesian Model

These results represent one of the four simulations for each of the nine design points (DP1 — DP9). The included tables are all for the case where $\lambda = 10$ and $\sigma = 1$. The parameter estimation is similar to what is shown here, for all other simulations. The differences that can be seen in the results, all relate to the widths of the 95% credible sets. The widths of the credible sets for the efficacy regression parameters are tied to both λ and σ. When λ decreases to -5 from 10, the width of the credible set also decreases slightly. When σ increases to 5 from 1, the credible set width also increases fairly drastically. The width of the credible sets for λ and σ also increase with the magnitude of σ and do not depend on the magnitude of λ. Finally, the safety regression parameter credible set widths do not appear to be affected by the magnitude of either λ or σ.

Table A.1: Simulation results for DP1 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	3.00	2.86	3.01	3.19	0.32	0.96
β_1	1.33	1.19	1.31	1.43	0.23	0.94
β_2	4.00	3.73	3.99	4.27	0.54	1.00
γ_0	-4.60	-7.00	-5.14	-3.85	3.15	0.92
γ_1	0.24	-0.25	0.22	0.67	0.92	0.98
γ_2	1.10	-0.04	1.46	3.44	3.47	0.96
λ	10.00	9.48	10.08	10.69	1.20	0.94
σ	1.00	0.95	1.01	1.07	0.11	0.96

Table A.2: Simulation results for DP2 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	3.00	2.74	3.04	3.38	0.63	0.96
β_1	1.33	1.18	1.31	1.43	0.26	0.96
β_2	4.00	3.57	3.94	4.28	0.71	0.96
γ_0	-2.44	-2.85	-2.43	-2.04	0.81	0.96
γ_1	0.41	0.18	0.39	0.59	0.41	1.00
γ_2	-1.45	-2.77	-1.76	-0.94	1.84	0.96
λ	10.00	9.63	10.00	10.36	0.74	0.94
σ	1.00	0.95	1.01	1.06	0.11	0.96

Table A.3: Simulation results for DP3 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	3.00	2.63	3.03	3.46	0.83	0.90
β_1	1.33	1.06	1.32	1.57	0.51	1.00
β_2	4.00	3.30	3.93	4.57	1.28	0.94
γ_0	−1.73	−2.07	−1.74	−1.44	0.63	0.94
γ_1	0.11	−0.06	0.10	0.25	0.31	0.96
γ_2	0.64	0.23	0.62	1.03	0.80	0.96
λ	10.00	9.82	10.03	10.23	0.41	0.96
σ	1.00	0.95	1.01	1.07	0.12	0.98

Table A.4: Simulation results for DP4 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	13.00	12.87	13.02	13.20	0.33	0.92
β_1	1.67	1.53	1.65	1.77	0.24	0.94
β_2	−6.00	−6.31	−6.05	−5.77	0.54	0.96
γ_0	−4.60	−6.71	−4.97	−3.75	2.96	0.94
γ_1	0.24	−0.32	0.15	0.60	0.92	0.88
γ_2	1.10	−0.23	1.22	3.11	3.34	0.96
λ	10.00	9.50	10.08	10.67	1.18	0.96
σ	1.00	0.95	1.00	1.06	0.11	0.92

Table A.5: Simulation results for DP5 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	13.00	12.69	12.98	13.30	0.61	0.92
β_1	1.67	1.52	1.65	1.79	0.26	0.94
β_2	−6.00	−6.34	−5.98	−5.65	0.69	0.96
γ_0	−2.44	−2.95	−2.52	−2.12	0.84	0.94
γ_1	0.41	0.20	0.41	0.62	0.43	0.98
γ_2	−1.45	−2.48	−1.54	−0.77	1.71	0.96
λ	10.00	9.65	10.02	10.39	0.74	0.98
σ	1.00	0.95	1.00	1.06	0.11	1.00

Table A.6: Simulation results for DP6 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	13.00	12.62	13.01	13.44	0.82	0.94
β_1	1.67	1.39	1.66	1.91	0.52	0.94
β_2	−6.00	−6.65	−6.02	−5.38	1.28	0.96
γ_0	−1.73	−2.08	−1.75	−1.45	0.63	0.92
γ_1	0.11	−0.06	0.10	0.25	0.31	0.94
γ_2	0.64	0.25	0.64	1.05	0.81	0.94
λ	10.00	9.80	10.00	10.20	0.40	1.00
σ	1.00	0.94	0.99	1.05	0.11	1.00

Table A.7: Simulation results for DP7 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	16.00	15.86	16.01	16.18	0.32	0.98
β_1	1.00	0.86	0.98	1.09	0.23	0.94
β_2	8.00	7.73	7.99	8.26	0.53	0.96
γ_0	−4.60	−6.50	−4.93	−3.80	2.70	1.00
γ_1	0.24	−0.28	0.18	0.62	0.90	0.98
γ_2	1.10	−0.12	1.24	2.93	3.05	0.96
λ	10.00	9.38	9.97	10.56	1.18	0.94
σ	1.00	0.94	1.00	1.06	0.11	0.94

Table A.8: Simulation results for DP8 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	16.00	15.72	16.01	16.34	0.63	0.96
β_1	1.00	0.85	0.98	1.10	0.26	0.92
β_2	8.00	7.61	7.97	8.31	0.70	0.98
γ_0	−2.44	−2.88	−2.45	−2.06	0.82	0.98
γ_1	0.41	0.17	0.38	0.58	0.41	0.94
γ_2	−1.45	−2.68	−1.68	−0.88	1.80	0.94
λ	10.00	9.62	9.98	10.35	0.73	0.92
σ	1.00	0.95	1.00	1.06	0.11	0.98

Table A.9: Simulation results for DP9 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	Mean	97.5%	Width	Coverage
β_0	16.00	15.61	16.00	16.43	0.82	0.92
β_1	1.00	0.72	0.98	1.23	0.51	0.86
β_2	8.00	7.34	7.96	8.59	1.26	0.96
γ_0	-1.73	-2.08	-1.76	-1.45	0.63	0.92
γ_1	0.11	-0.06	0.10	0.25	0.31	0.84
γ_2	0.64	0.24	0.64	1.04	0.80	0.92
λ	10.00	9.80	10.00	10.20	0.41	1.00
σ	1.00	0.95	1.00	1.06	0.11	0.96

APPENDIX B

Mathematica Code to Calculate the Observed Information

(* The log likelihood *)

$$f\left[\gamma_{-0}, \gamma_{-1}, \gamma_{-2}, \beta_{-0}, \beta_{-1}, \beta_{-2}, \lambda_{-}, \sigma_{-}\right] :=$$

$$-0.5 * \text{Log}[2 * \pi * \sigma\text{^}2] - (1 + \text{Exp}\left[\gamma_0 + \gamma_1 * x_1 + \gamma_2 * x_2\right]) +$$

$$(y_1 * (\gamma_0 + \gamma_1 * x_1 + \gamma_2 * x_2)) - (1/(2 * \sigma\text{^}2)) * (y_2 - \beta_0 - \beta_1 * x_1 - \beta_2 * x_2 - \lambda *$$

$$(y_1 - (\text{Exp}\left[\gamma_0 + \gamma_1 * x_1 + \gamma_2 * x_2\right] / (1 + \text{Exp}\left[\gamma_0 + \gamma_1 * x_1 + \gamma_2 * x_2\right])))) \text{\^}2$$

(* Returns the elements of the observed information matrix *)

$$\text{Simplify}\left[-D\left[f\left[\gamma_0, \gamma_1, \gamma_2, \beta_0, \beta_1, \beta_2, \lambda, \sigma\right], \left\{\left\{\gamma_0, \gamma_1, \gamma_2, \beta_0, \beta_1, \beta_2, \lambda, \sigma\right\}, 2\right\}\right]\right]$$

APPENDIX C

Simulation Results using the r^* Method

Table C.1: Simulation results for the r^* method for DP2 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.94	3.05	3.14	0.20	0.40
β_1	1.33	1.23	1.32	1.40	0.18	0.86
β_2	4.00	3.78	3.93	4.05	0.26	0.42
γ_0	-2.44	-2.54	-2.39	-2.24	0.30	0.58
γ_1	0.41	0.28	0.40	0.52	0.25	0.80
γ_2	-1.45	-2.35	-1.67	-1.18	1.16	0.86
λ	10.00	9.64	10.01	10.37	0.73	0.94
σ	1.00	0.94	1.00	1.06	0.11	1.00

Table C.2: Simulation results for the r^* method for DP2 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.49	2.93	3.36	0.87	0.90
β_1	1.33	0.98	1.38	1.78	0.80	0.90
β_2	4.00	3.45	4.08	4.68	1.22	0.82
γ_0	-2.44	-2.76	-2.45	-2.11	0.65	0.92
γ_1	0.41	0.14	0.44	0.74	0.60	0.86
γ_2	-1.45	-2.53	-1.62	-0.79	1.75	0.94
λ	10.00	8.23	10.01	11.81	3.58	0.96
σ	5.00	4.65	4.93	5.20	0.55	0.90

Table C.3: Simulation results for the r^* method for DP2 with $\lambda = -5$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.92	3.01	3.11	0.19	0.64
β_1	1.33	1.24	1.33	1.42	0.17	0.88
β_2	4.00	3.87	3.99	4.13	0.26	0.66
γ_0	-2.44	-2.70	-2.45	-2.21	0.48	0.64
γ_1	0.41	0.19	0.40	0.60	0.42	0.88
γ_2	-1.45	-2.40	-1.55	-0.91	1.49	0.80
λ	-5.00	-5.40	-5.03	-4.66	0.73	0.86
σ	1.00	0.94	1.00	1.05	0.11	0.94

Table C.4: Simulation results for the r^* method for DP3 with $\lambda = -5$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.89	2.98	3.08	0.19	0.50
β_1	1.33	1.24	1.33	1.42	0.18	0.76
β_2	4.00	3.89	4.02	4.16	0.28	0.38
γ_0	−1.73	−1.83	−1.72	−1.61	0.22	0.50
γ_1	0.11	0.02	0.12	0.22	0.20	0.66
γ_2	0.64	0.48	0.62	0.76	0.27	0.46
λ	−5.00	−5.18	−4.98	−4.78	0.40	0.96
σ	1.00	0.94	1.00	1.05	0.11	0.88

Table C.5: Simulation results for the r^* method for DP3 with $\lambda = -5$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	3.00	2.61	3.05	3.48	0.87	0.84
β_1	1.33	0.91	1.32	1.72	0.81	0.98
β_2	4.00	3.32	3.93	4.55	1.23	0.84
γ_0	−1.73	−1.97	−1.77	−1.56	0.41	0.82
γ_1	0.11	−0.07	0.11	0.29	0.36	0.94
γ_2	0.64	0.42	0.68	0.94	0.52	0.78
λ	−5.00	−5.98	−4.97	−3.97	2.02	0.92
σ	5.00	4.67	4.95	5.23	0.56	0.96

Table C.6: Simulation results for the r^* method for DP6 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	13.00	12.60	13.06	13.50	0.89	0.84
β_1	1.67	1.31	1.73	2.15	0.84	0.86
β_2	−6.00	−6.78	−6.14	−5.50	1.27	0.86
γ_0	−1.73	−1.89	−1.72	−1.54	0.35	0.80
γ_1	0.11	−0.01	0.15	0.30	0.31	0.74
γ_2	0.64	0.40	0.62	0.84	0.44	0.80
λ	10.00	9.17	10.15	11.15	1.98	0.86
σ	5.00	4.73	5.01	5.29	0.56	1.00

Table C.7: Simulation results for the r^* method for DP8 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.54	16.00	16.44	0.89	0.84
β_1	1.00	0.64	1.04	1.44	0.80	0.94
β_2	8.00	7.44	8.07	8.68	1.24	0.88
γ_0	−2.44	−2.77	−2.45	−2.11	0.66	0.84
γ_1	0.41	0.07	0.37	0.66	0.59	0.88
γ_2	−1.45	−2.31	−1.47	−0.70	1.61	0.92
λ	10.00	8.00	9.81	11.62	3.62	0.88
σ	5.00	4.71	5.00	5.26	0.55	0.90

Table C.8: Simulation results for the r^* method for DP8 with $\lambda = -5$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.47	15.90	16.34	0.87	0.80
β_1	1.00	0.61	1.00	1.40	0.80	0.94
β_2	8.00	7.49	8.11	8.71	1.21	0.78
γ_0	−2.44	−2.77	−2.43	−2.06	0.71	0.90
γ_1	0.41	0.09	0.42	0.75	0.65	0.92
γ_2	−1.45	−2.52	−1.63	−0.78	1.74	0.94
λ	−5.00	−7.04	−5.26	−3.41	3.63	0.98
σ	5.00	4.71	4.99	5.26	0.56	0.96

Table C.9: Simulation results for the r^* method for DP9 with $\lambda = 10$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.87	15.98	16.08	0.21	0.46
β_1	1.00	0.90	1.00	1.09	0.19	0.42
β_2	8.00	7.87	8.02	8.17	0.30	0.46
γ_0	−1.73	−1.83	−1.77	−1.70	0.13	0.40
γ_1	0.11	0.05	0.11	0.17	0.12	0.48
γ_2	0.64	0.59	0.67	0.75	0.16	0.40
λ	10.00	9.81	10.01	10.21	0.40	1.00
σ	1.00	0.94	1.00	1.05	0.11	0.94

Table C.10: Simulation results for the r^* method for DP9 with $\lambda = 10$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.53	15.98	16.42	0.90	0.84
β_1	1.00	0.62	1.03	1.45	0.84	0.88
β_2	8.00	7.36	8.00	8.63	1.27	0.84
γ_0	-1.73	-1.94	-1.76	-1.58	0.36	0.68
γ_1	0.11	-0.05	0.10	0.26	0.32	0.80
γ_2	0.64	0.43	0.66	0.88	0.45	0.70
λ	10.00	8.97	9.99	11.01	2.04	0.94
σ	5.00	4.71	5.00	5.27	0.56	1.00

Table C.11: Simulation results for the r^* method for DP9 with $\lambda = -5$ and $\sigma = 1$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.89	15.99	16.08	0.19	0.66
β_1	1.00	0.92	1.01	1.10	0.18	0.70
β_2	8.00	7.87	8.01	8.14	0.27	0.60
γ_0	-1.73	-1.83	-1.73	-1.61	0.22	0.66
γ_1	0.11	-0.01	0.09	0.19	0.20	0.76
γ_2	0.64	0.50	0.63	0.77	0.27	0.64
λ	-5.00	-5.17	-4.97	-4.77	0.40	0.94
σ	1.00	0.93	0.99	1.04	0.11	0.88

Table C.12: Simulation results for the r^* method for DP9 with $\lambda = -5$ and $\sigma = 5$.

Parameter	True	2.5%	MLE	97.5%	Width	Coverage
β_0	16.00	15.57	16.00	16.44	0.88	0.82
β_1	1.00	0.59	1.00	1.41	0.81	0.92
β_2	8.00	7.37	7.99	8.61	1.24	0.82
γ_0	-1.73	-1.97	-1.77	-1.56	0.41	0.76
γ_1	0.11	-0.06	0.12	0.30	0.36	1.00
γ_2	0.64	0.43	0.68	0.94	0.52	0.80
λ	-5.00	-6.18	-5.17	-4.16	2.03	0.96
σ	5.00	4.70	4.99	5.27	0.56	0.92

APPENDIX D

Simulation Results for the Power Study

Table D.1: Simulation results for the power study using the optimistic priors.

	Individual Powers				Fallback	
Sample Size	Efficacy	Safety	Bonferroni	Fixed-Sequence	Efficacy First	Safety First
50	0.535	0.544	0.181	0.341	0.231	0.255
100	0.741	0.729	0.414	0.580	0.482	0.490
200	0.885	0.891	0.711	0.813	0.752	0.766
300	0.952	0.954	0.847	0.912	0.878	0.878
400	0.987	0.979	0.929	0.968	0.947	0.948

Table D.2: Simulation results for the power study using the reference priors.

	Individual Powers				Fallback	
Sample Size	Efficacy	Safety	Bonferroni	Fixed-Sequence	Efficacy First	Safety First
50	0.328	0.391	0.160	0.235	0.183	0.200
100	0.552	0.578	0.314	0.426	0.343	0.371
200	0.790	0.823	0.593	0.710	0.638	0.652
300	0.905	0.921	0.764	0.857	0.791	0.820
400	0.962	0.960	0.876	0.930	0.898	0.908

Table D.3: Simulation results for the power study using the skeptical priors.

	Individual Powers				Fallback	
Sample Size	Efficacy	Safety	Bonferroni	Fixed-Sequence	Efficacy First	Safety First
50	0.112	0.184	0.015	0.035	0.019	0.024
100	0.360	0.444	0.121	0.225	0.154	0.170
200	0.711	0.754	0.436	0.595	0.500	0.513
300	0.870	0.887	0.669	0.800	0.709	0.745
400	0.949	0.949	0.827	0.907	0.861	0.873

Table D.4: Simulation results for the power study with the optimistic priors when the design prior for γ_0 is changed to $N(-1.2, 0.2)$.

	Individual Powers				Fallback	
Sample Size	Efficacy	Safety	Bonferroni	Fixed-Sequence	Efficacy First	Safety First
50	0.490	0.587	0.171	0.343	0.217	0.260
100	0.693	0.796	0.430	0.611	0.485	0.527
200	0.852	0.932	0.712	0.824	0.733	0.791
300	0.948	0.982	0.873	0.937	0.883	0.918
400	0.971	0.990	0.925	0.964	0.931	0.957

Table D.5: Simulation results for the power study with the optimistic priors when the design prior for γ_0 is changed to $N(-2.0, 0.2)$.

	Individual Powers				Fallback	
Sample Size	Efficacy	Safety	Bonferroni	Fixed-Sequence	Efficacy First	Safety First
50	0.511	0.595	0.195	0.367	0.242	0.291
100	0.678	0.786	0.430	0.591	0.471	0.528
200	0.852	0.927	0.685	0.821	0.721	0.770
300	0.920	0.975	0.847	0.910	0.861	0.892
400	0.960	0.991	0.916	0.955	0.922	0.946

BIBLIOGRAPHY

Barndorff-Nielsen, O. E. (1983), "On a formula for a distribution of the maximum likelihood estimator," *Biometrika*, 70, 343–365.

— (1986), "Inference on full or partial parameters based on the standardized signed log likelihood ratio," *Biometrika*, 73, 307–322.

— (1990), "Approximate interval probabilities," *Journal of the Royal Statistical Society, Series B*, 52, 485–496.

Barndorff-Nielsen, O. E. and Cox, D. R. (1994), *Inference and Asymptotics*, London: Chapman and Hall.

Barreto, L. S., Cysneiros, A. H., and Cribari-Neto, F. (2013), "Improved Birnbaum-Saunders inference under type II censoring," *Computational Statistics and Data Analysis*, 57, 68–81.

Brazzale, A. R., Davison, A. C., and Reid, N. (2007), *Applied Asymptotics: Case Studies in Small-Sample Statistics*, Cambridge University Press.

Butler, R. W. (2007), *Saddlepoint Approximations with Applications*, Cambridge University Press.

Carlin, B. P. and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, Taylor and Francis Group, LLC, 3rd ed.

Christensen, R., Johnson, W., Branscum, A., and Hanson, T. E. (2011), *Bayesian Ideas and Data Analysis*, Taylor and Francis Group, LLC.

Conaway, M. R. and Petroni, G. R. (1996), "Designs for Phase II Trials Allowing for a Trade-off Between Response and Toxicity," *Biometrics*, 52, 1375–1386.

Cortese, G. and Ventura, L. (2012), "Accurate higher-order inference on $P(Y < X)$," *Computational Statistics*, online.

Cox, D. R. (1972), "The Analysis of Multivariate Binary Data," *Applied Statistics*, 21, 113–120.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2010), *Multiple Testing Problems in Pharmaceutical Statistics*, Taylor and Francis Group, LLC.

Dunn, O. J. (1961), "Multiple Comparisons Among Means," *Journal of the American Statistical Association*, 56, 52–64.

Efron, B. (1998), "R.A. Fisher in the 21st century," *Statistical Science*, 13, 95–122.

FDA (2010), "Guidance for the Use of Bayesian Statistics in Medical Device Clinical Trials," Http://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071121.pdf.

Fitzmaurice, G. M. and Laird, N. M. (1995), "Regression Models for a Bivariate Discrete and Continuous Outcome with Clustering," *Journal of the American Statistical Association*, 90, 845–852.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, Chapman and Hall/CRC, 2nd ed.

Gönen, M., Westfall, P. H., and Johnson, W. O. (2003), "Bayesian Multiple Testing for Two-Sample Multivariate Endpoints," *Biometrics*, 59, 76–82.

Gueorguieva, R. and Agresti, A. (2001), "A correlated probit model for multivariate repeated measures of mixtures of binary and continuous responses," *Journal of the American Statistical Association*, 96, 1102–1112.

Jiang, L. and Wong, A. (2012), "On standardizing the signed root log likelihood ratio statistic," *Statistics and Probability Letters*, 82, 833–839.

Lauritzen, S. L. and Wermuth, N. (1989), "Graphical Models for Associations between Variables, some of which are Qualitative and some Quantitative," *The Annals of Statistics*, 17, 31–57, correction in 17, 1916.

Lee, J. J. and Chu, C. T. (2012), "Bayesian Clinical Trials in Action," *Statistics in Medicine*, 31, 2955–2972.

Maurer, W., Hothorn, L. A., and Lehmacher, W. (1995), "Multiple comparisons in drug clinical trials and preclinical assays: a priori ordered hypotheses," *Biometrie in der Chemisch-in-Pharmazeutischen Intustrie*, 6, 3–18, J. Vollman (editor). Fischer-Verlag, Stuttgart.

McCulloch, C. (2008), "Joint modelling of mixed outcome types using latent variables," *Statistical Methods in Medical Research*, 17, 53–73.

Nelder, J. A. and Mead, R. (1965), "A simplex method for function minimization," *Computer Journal*, 7, 308–313.

Olkin, I. and Tate, R. F. (1961), "Multivariate Correlation Models with Mixed Discrete and Continuous Variables," *Annals of Mathematical Statistics*, 32, 448–465, correction in 36, 343-344.

Pace, L., Salvan, A., and Ventura, L. (2011), "Adjustments of profile likelihood through predictive densities," *Annals of the Institute of Statistical Mathematics*, 63, 923–937.

Pawitan, Y. (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press Inc.

Pennello, G. and Thompson, L. (2007), "Experience with Reviewing Bayesian Medical Device Trials," *Journal of Biopharmaceutical Statistics*, 18, 81–115.

Perlman, M. D. and Wu, L. (2004), "A Note on One-Sided Tests with Multiple Endpoints," *Biometrics*, 60, 276–279.

Pocock, S. J., Gellar, N. L., and Tsiatis, A. A. (1987), "The Analysis of Multiple Endpoints in Clinical Trials," *Biometrics*, 43, 487–498.

Robert, C. P. (2007), *The Bayesian Choice*, Springer Science+Business Media, LLC.

Robertson, H. T. and Allison, D. B. (2009), "Drugs Associated with More Suicidal Ideations Are also Associated with More Suicide Attempts," *PLoS ONE*, 4.

Rubin, D. B. (1984), "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151–1172.

Seaman III, J. W., Seaman Jr., J. W., and Stamey, J. D. (2012), "Hidden Dangers of Specifying Noninformative Priors," *The American Statistician*, 66, 77–84.

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

Teixeira-Pinto, A. and Normand, S.-L. T. (2009), "Correlated bivariate continuous and binary outcomes: Issues and applications," *Statistics in Medicine*, 28, 1753–1773.

Thall, P. F. and Cheng, S.-C. (1999), "Treatment Comparisons Based on Two-Dimensional Safety and Efficacy Alternatives in Oncology Trials," *Biometrics*, 55, 746–753.

Thall, P. F. and Cook, J. D. (2004), "Dose-Finding Based on Efficacy-Toxicity Trade-Offs," *Biometrics*, 60, 684–693.

Thomas, A., O'Hara, B., Ligges, U., and Sturtz, S. (2006), "Making BUGS Open," *R News*, 6, 12–17.

Wang, F. and Gelfand, A. E. (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.

Wei, H. (2012), "Bayesian Modelling of Mixed Outcome Types Using Random Effect," Ph.D. thesis, Baylor University.

Westfall, P. H. and Krishen, A. (2001), "Optimally weighted, fixed sequence and gatekeeper multiple testing procedures," *Journal of Statistical Planning and Inference*, 99, 25–40.

Winkler, R. L. (2001), "Why Bayesian Analysis hasn't caught on in Healthcare Decision Making," *International Journal of Technology Assessment in Health Care*, 17, 56–66.