# ABSTRACT

Frequentist and Bayesian Modeling in the Presence of Unmeasured Confounding

Allison E. Hainline

Director: Jeanne S. Hill, Ph.D.

Biostatistical studies of medical data are extremely important in distinguishing relationships between drugs or treatments and the patient's medical response. These studies generally use data from large health care databases, which provide immense amounts of information while allowing the researcher to analyze long-term effects that may not be shown in a typical randomized controlled trial. However, when using large databases, one must be particularly aware of the effect of unmeasured confounding on statistical models. Confounding arises when factors unrelated to the particular study have a hidden effect on observed health outcomes. Bayesian statistics provides a mechanism for model fitting which synthesizes the data with prior information about bias, allowing the researcher to control confounding through the inclusion of additional variables from independent data sets. In this thesis I will provide a background of the proposed method as well its application to two independent analyses: the prediction of low birth weight babies and the prediction of parental separation anxiety.

# FREQUENTIST AND BAYESIAN MODELING IN THE PRESENCE OF UNMEASURED CONFOUNDING

A Thesis Submitted to the Faculty of

Baylor University

In Partial Fulfillment of the Requirements for the

Honors Program

By

Allison E. Hainline

Waco, Texas

May 2013

## TABLE OF CONTENTS

# CHAPTER ONE

## Introduction

### *Background Terms and Concepts*

#### *Logistic Regression*

Logistic regression is a type of regression analysis in which the dependent variable is categorical and is predicted by one or more dependent variables which may be either categorical or continuous. The probabilities are described using a logistic function, where the outcome is a function of the explanatory variables. One model used extensively in healthcare data is the logistic model in which the dependent variable is binary. For example, one may wish to determine if the use of particular drug increases or decreases a particular disease risk. In this case, both the predictor and the outcome variable are binary. However, if it was a study of the effect of drug dosage on disease outcome, the predictor would be continuous while the outcome variable would remain binary. Additionally, one could add a confounder to the model, such as smoking status, which is a binary value.

Logistic regression is used to predict the odds of being a case (observing the disease or outcome) based on the values of the predictor variables. The odds ratio of a logistic regression analysis is the odds of having the disease divided by the odds of not having the disease.

A linear function of the independent variable,

$$logit(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ..., \tag{1}$$

is on a scale of negative infinity to positive infinity, while we want the probability of success to fall between 0 and 1. Thus, we must transform the linear function into a probability by setting the logarithm of the odds ratio equal to the linear function:

$$log_e(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \tag{2}$$

Solving for $\pi$, the probability of success, we get the logistic equation,

$$\pi = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}. \tag{3}$$

All observations must be independent of one another for the logistic model. [1, p. 181]

Multiple logistic regression is a logistic regression analysis one may use in order to consider several covariates simultaneously. With multiple logistic regression, the predictors may be a mixture of binary and continuous variables. Odds ratios may be calculated for each predictor resulting in the odds of a case for a one unit increase (for continuous variables) or with the presence of the factor (for binary variables), on average, holding all other variables constant. To demonstrate multiple regression, data from an existing study will be analyzed.

In the analysis, a logistic model that includes only the variables that contribute significantly to the outcome is defined. To do this, a logistic model that includes each of the variables in the original data set is created. The final model is then specified by using a Backward Stepwise Elimination procedure. In this procedure, the model is run with all variables included, and p-values are found for each independent variable, where the null hypothesis is that all $\beta_n = 0$ and the alternative hypothesis states that the coefficient is not equal to zero and thus contributes to the model. The independent variable with the highest p-value (above the significance level, $\alpha$) is then removed from the model. The model is then run with the remaining variables, and the independent variable with the highest p-value is removed. This process is repeated until all remaining variables have p-values below the significance level. The maximum likelihood estimates are given in the `summary()` function in R. Once we have the maximum likelihood estimates, we can define the fitted value, or the probability of a case at the predetermined variable levels.

### *Bayesian v. Frequentist Statistics*

Modern statistical inference can be divided into two main schools of thought: frequentist and Bayesian. The frequentist approach is the classical method for statistical analysis in which prior information, gained by previous studies or trials, is only used during the design stage. Bayesian statistics, by contrast, builds prior information into the formal analysis as it becomes available. Previous trials, studies in foreign countries, or expert opinion are all considered valid sources for prior information. The use of prior information in the analysis can be helpful in trial design and it can be argued that a more accurate result comes out of a Bayesian analysis. The prior information allows the researcher to decrease the scope of a trial, resulting in a more specific result.

In a frequentist analysis, the interpretation of the data depends on the intentions of the researcher. For frequentists, before the experiment is conducted, critical values must be determined, and p-values are the basis for the decision-making. Bayesian statisticians assert that hypotheses should be compared by how well they explain the data. P-values represent the probability of observing a test statistic the same or more extreme than the one observed, assuming that the null hypothesis is true. In other words, how likely one is to observe a value or one more extreme, assuming that there is no association between the dependent and independent variables. A small p-value in a frequentist analysis will result in the rejection of the null hypothesis (often assuming no association) in favor of the alternative hypothesis (often claiming association is present).

However, when the null hypothesis is rejected in favor of the alternative hypothesis, the ability of the alternative hypothesis to explain the data is never measured. A small p-value demonstrates the rarity of observing the test statistic assuming no association, but it neglects to inform exactly how much of the result was determined by the specific independent variable. Bayesian statistics allows the researcher to obtain odds ratios and predictive probabilities that can demonstrate the magnitude of the effect that the variable has on the outcome variable.

With Bayesian analysis every analysis is made up of two types of information: the data being analyzed and the prior information. The researcher is free to choose whatever prior information he/she believes helps explain the data being analyzed. [10] In Bayesian analysis, the intentions of the experimenter do not have an impact on the results because every aspect of the experiment that could affect the results is accounted for within the likelihood function. Thus, no outside forces are able to inflate the significance of the test, and the results are more accurate. [9, p. 274] As a result, the significance is determined not from pre-determined critical values or p-values, but rather, it is determined by odds ratios and predictive probabilities that demonstrate the effects of each parameter on the final result.

### *Meta-analysis*

A meta-analysis is a statistical analysis of a group of studies. The studies generally compare and contrast the results of these studies in order to identify certain patterns, relationships, or disagreements. Meta-analyses may be used as a summary of the effect across studies or as an estimation of differences among the studies. Meta-analyses are useful in that they allow the researcher to create a study with a much wider scope, meaning more variables may be available for analysis. In the context of this analysis, meta-analysis will be utilized to minimize confounding. Additionally, researchers may compare the outcomes of studies that use different samples of participants to identify differences in variance or weaknesses of the study. The results may then be used to alter future studies in order to arrive at a more accurate conclusion. [3]

### *Unmeasured Confounders*

Confounding factors, also known as hidden or lurking variables, are variables not directly included in the statistical model that have a correlation with both the independent and dependent variables. Confounding arises when factors unrelated to the particular study have a hidden effect on observed health outcomes. [5] For example, it has been observed that there is a positive correlation between ice cream consumption and drowning deaths. It would be in error to assume causality, i.e. ice cream consumption increases the probability of drowning. However, the confounding factor in the example is the outside temperature. As the temperature increases, more people go to swimming pools, and, if the fraction of people who drown remains constant, more swimmers will drown as a result of the population increase. Though the factor is obvious in this example, confounding factors are not always this easy to identify.

This becomes a particularly important problem when using health care databases. When using databases, one faces the problem of missing, poorly measured, or poorly recorded potentially confounding factors. There are several types of confounding that can be found in healthcare studies. Physicians often make decisions on treatment based on the patient's health status and willingness to take medication. Due to this, extremely ill patients are less likely to be prescribed medication, because the likelihood of benefit is so low. [2, p. S115] In addition, a patient's lifestyle may also serve as a confounding factor. For instance, a patient who takes preventative measures against disease through medication is more likely to engage in other preventative behaviors, such as healthy eating and exercise. As a result, it may appear that the use of a certain preventative medication protects against a wider variety of diseases, where the true cause is simply a patient lifestyle factor. [2, p. S115] Because the researcher often has no control over the variables that are measured in these widely used databases, unmeasured confounding is extremely likely and must be considered. Fortunately, unmeasured confounding has the potential to be controlled through combination of several data sources, or a meta-analysis.

By combining several data sources, the researcher is able to compare and contrast the variables measured in each dataset. If the researcher finds a dataset that measures a variable that is unmeasured in the dataset he/she is analyzing, the data provided by that variable may be incorporated into the model as prior information. This way, one can control hidden confounders without obtaining more data from the same source.

### *Methods and Software*

The majority of the analyses were conducted using the `R version 2.15.1`. [13] In particular, the `R` packages `ggplot2`, `plyr`, `R2OpenBUGS`, `xlsx`, `BRugs`, and `arm` were used in performing the analyses. The Bayesian analysis was conducted using the statistical software OpenBUGS. When possible, OpenBUGS was run through `R` using the `R2OpenBUGS` package.

# CHAPTER TWO

## Prediction of Low Birth Weight Babies

### *The Data*

In this initial analysis, an existing data set was used. The purpose of the study was to determine the risk factors that were associated with having a low birth weight baby. In this context, a low birth weight baby is a child that weighed less than 2500 grams at birth. The data were collected at Baystate Medical Center in Springfield, Massachusetts, as part of a study aiming to determine risk factors for delivering a low birth weight baby. 189 women were included in the study, 59 of whom had low birth weight babies. Variables of interest included age, weight of the mother at her last menstrual period (”LWT”), race (”RACE”), and the number of physician visits during the first trimester of pregnancy (”FTV”). Information was also collected on smoking status of the mother during pregnancy (”SMOKE”), history of premature labor (”PTL”), history of hypertension (”HT”), and presence of uterine irritability (”UI”).[7, p. 25]

In this example, age and weight are continuous variables, whereas the others are categorical variables. The binary variables were reported so that the value 1 corresponds to the presence of the factor and the value 0 corresponds to its absence. For example, a subject who smoked during her pregnancy would be assigned a value of 1 for the ”SMOKE” variable. The ”RACE” variable had three categorical responses: 1 for ”White”, 2 for ”Black”, and 3 for ”Other”.

In this example, the fitted value represents the predicted probability of having a low birth weight baby at specific parameter values.[7]

## *Results of a Traditional Frequentist Analysis*

### *Model Specification*

Before incorporating any analysis of confounding within the data, an analysis will be conducted in traditional frequentist fashion as well as in a simple Bayesian analysis with no confounding factor. Thus, no prior information will be added to the regression. This preliminary study will be conducted using the data previously introduced, which was used to identify potential risk factors associated with the birth of a low birth weight baby.

To create the logistic model for this data, a Backward Elimination procedure was performed. The initial model includes every variable in the dataset: mother's age, mother's weight at last menstrual period, race, smoking status during pregnancy, history of premature labor, history of hypertension, presence of uterine irritability, and number of physician visits during the first trimester. This full model, as entered into R, is shown below.

```
# Full Logistic Model
mydata = read.csv("C:/Users/Allison/Desktop/THESIS/lowbwt.csv", header = T)
mylogit <- glm(LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV, data = mydata,
    family = "binomial")
```

In order to narrow down the model to only the variables that contribute significantly to the outcome variable, LWT, we must look at the p-values for each variable. The coefficients for each variable in the full model are shown in the R summary output below. The p-values can be found in the fourth column. Several variables are significant at this point, while others are far from significance. The first step in the Backward Selection procedure is to remove the parameter with the largest p-value, i.e. the variable that contributes the least to the model. In this first round, FTV, with a p-value of 0.7085, is the least significant and must be removed.

```
summary(mylogit)$coefficients

##              Estimate Std. Error  z value Pr(>|z|)
## (Intercept) -0.07897   1.276254 -0.06188 0.950658
## AGE         -0.03585   0.036472 -0.98283 0.325690
## LWT         -0.01239   0.006614 -1.87266 0.061115
## RACE         0.45342   0.215294  2.10607 0.035198
## SMOKE        0.93728   0.398458  2.35225 0.018660
## PTL          0.54209   0.346168  1.56597 0.117357
## HT           1.83072   0.694135  2.63741 0.008354
## UI           0.72196   0.463174  1.55873 0.119060
## FTV          0.06346   0.169765  0.37381 0.708542
```

The model is then run again, with FTV removed. This procedure is repeated until each of the parameters is significant ($\alpha = 0.05$). The final model, including LWT, RACE, SMOKE, and HT, is shown in the R output below. Also shown are odds ratios for each parameter of the final model.

```
final_logit <- glm(LOW ~ LWT + RACE + SMOKE + HT, data = mydata,
                   family = "binomial")
summary(final_logit)$coefficients

##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.35754   1.010584 -0.3538 0.723495
## LWT         -0.01535   0.006523 -2.3539 0.018578
## RACE         0.48955   0.207324  2.3613 0.018211
## SMOKE        1.08002   0.383735  2.8145 0.004885
## HT           1.74427   0.687563  2.5369 0.011184
```

```
exp(cbind(OR = coef(final_logit), confint(final_logit)))
```

```
## Waiting for profiling to be done...
```

```
##                 OR  2.5 %  97.5 %
## (Intercept) 0.6994 0.1001  5.4009
## LWT         0.9848 0.9714  0.9967
## RACE        1.6316 1.0949  2.4792
## SMOKE       2.9447 1.4091  6.4002
## HT          5.7217 1.5421 24.1448
```

The final logit for the Low Birth Weight data is given by:

$$logit(\mu) = -0.3575 - 0.0154(LWT) + 0.4896(RACE) + 1.0800(SMOKE) + 1.7443(HT). \tag{4}$$

The final logistic model is given by:

$$\hat{\pi} = \frac{e^{-0.3575-0.0154(LWT)+0.4896(RACE)+1.0800(SMOKE)+1.7443(HT)}}{1+e^{-0.3575-0.0154(LWT)+0.4896(RACE)+1.0800(SMOKE)+1.7443(HT)}} \tag{5}$$

*Odds Ratios and Predictive Probabilities*

Table 1 contains the odds ratio and its 95% Confidence Interval for each predictor in the final logistic model. From the table, for each one unit increase in the mother's weight at last menstrual period, the odds of having a low birth weight baby decrease by a factor of

0.9848, on average, holding all other variables constant. Similarly, with all other variables constant, smoking increases the odds of having a low birth weight baby, on average, by a factor of 2.9447.

Table 1: Odds Ratios and Confidence Intervals for Regression Parameters

| Characteristic | OR | 95% CI |
|---|---|---|
| LWT | 0.9848 | (0.9714 , 0.9967) |
| RACE | 1.6316 | (1.0949 , 2.4792) |
| SMOKE | 2.9447 | (1.4091 , 6.4002) |
| HT | 5.7217 | (1.5421 , 24.1448) |

This final model can be used to calculate the predictive probabilities for several values of the explanatory variables. The predictive probability represents the probability of the outcome at specific explanatory variable values. For example, the predictive probability for a white female with no smoking history, no history of hypertension, and weight of 135 pounds at her last menstrual period is 0.1256, as demonstrated in equation 6.

$$\frac{e^{-0.35754-0.01535(135)+0.48955(1)+1.08002(0)+1.74427(0)}}{1+e^{-0.35754-0.01535(135)+0.48955(1)+1.08002(0)+1.74427(0)}} = 0.1256 \tag{6}$$

In other words, for a person with these parameter values, the probability of having a low birth weight child is 12.56%. Predictive probabilities may be calculated for a variety of parameter values, isolating particular variables to generalize the impact the variable has on the outcome. For example, the predictive probability for a white female who weighed 135 pounds at her last menstrual period with no history of hypertension, who smoked during her pregnancy, is 0.2972. Thus, smoking during pregnancy caused the probability of a low birth weight child to more than double. We can conclude, from this preliminary analysis, that smoking is a serious risk factor for low birth weight babies.

Table 2 provides a list of parameter values used for each model throughout this analysis. From this point on, the models will be referred to by the model number, rather than the list of parameter values.

Table 2: Model Parameter Values

| | LWT | RACE | SMOKE | HT |
|---|---|---|---|---|
| Model 1 | 135 | White | No | No |
| Model 2 | 135 | White | Yes | No |
| Model 3 | 135 | White | No | Yes |
| Model 4 | 95 | White | No | No |
| Model 5 | 135 | Black | No | No |
| Model 6 | 135 | White | Yes | Yes |
| Model 7 | 135 | Other | No | No |
| Model 8 | 95 | Black | No | No |

## *Results of the Bayesian Analysis*

### *OpenBUGS*

To compare the frequentist with the Bayesian approach, the data was also analyzed using standard Bayesian techniques. OpenBUGS, the open-sourced version of the statistical software, WinBUGS, was used for Bayesian analysis of the complex models. OpenBUGS uses Markov Chain Monte Carlo (MCMC) methods, which provide samples from the posterior distributions of the supplied parameters. With this approach, the estimated distribution for the unmeasured confounder is generated by the repeated sampling from the prior distribution after it is adjusted, unlike traditional Bayesian analysis where the posterior distribution is known. [14] When using OpenBUGS, the researcher must choose the number of ”chains” he/she wishes to sample. In this analysis, three chains are used in order to ensure that the model converges.

*Model Specification and Parameter Values*

To compare the two sets of results, I will analyze the data using the same explanatory variable values as were performed in the frequentist analysis.

$$logit(\mu) = \beta_0 + \beta_1 * (LWT) + \beta_2 * (RACE) + \beta_3 * (SMOKE) + \beta_4 * (HT), \qquad (7)$$

will be used in these analyses. In this method, loose priors are put on the parameters present in the final model of the frequentist analysis: LWT, RACE, SMOKE, and HT. Loose priors are sufficient for modeling each parameter's true distribution. Each of these parameters was given a prior that follows the same normal distribution,

$$\beta_i \sim \mathcal{N}(0,4),$$

demonstrated in Figure 1.

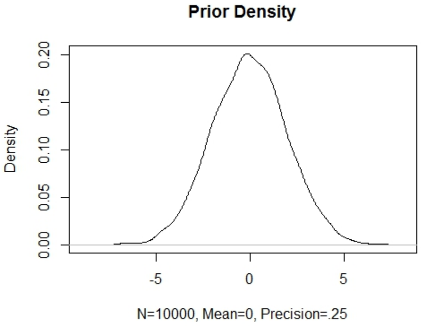


Figure 1: Prior distribution for the coefficients of LWT, RACE, SMOKE and HT: $\beta_1, \beta_2, \beta_3, and \beta_4$

Using OpenBUGS, I created a model to represent the data along with each parameter's prior distribution, detailed in Appendix B. This model was used to calculate the odds ratios for each parameter as well as the predictive probability for the model. The same values that were detailed in Table 2 for the frequentist analysis were also used for the Bayesian analysis. By using the same explanatory variable values, we can compare the predictive probabilities to demonstrate the changes that occur between the two methods.

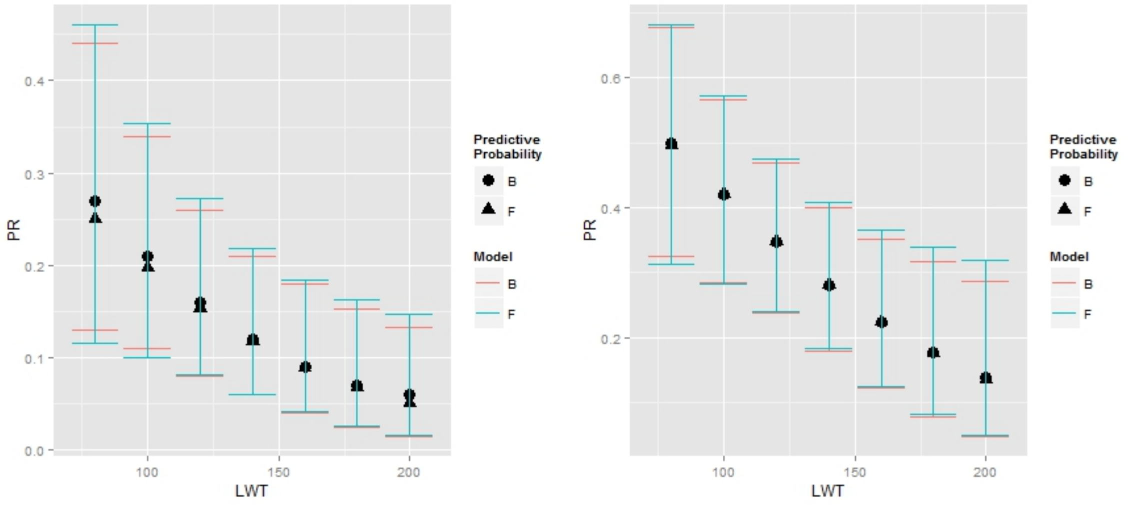
*Comparison of Predictive Probabilities*

Table 3 lists the predictive probabilities of the eight distinct sets of explanatory variables. The predictive probabilities vary between the frequentist and Bayesian approaches. In many cases, the Bayesian predictive probabilities are considerably higher than those resulting from a traditional frequentist analysis. Trials 3 and 6 (where the subject had hypertension) imply that the prior distributions had a large impact on the final predictive probability.

Table 3: Model Predictive Probabilities for Frequentist and Bayesian Analysis

| Model | Frequentist | Bayesian |
|---|---|---|
| 1 | 0.1256 | 0.1328 |
| 2 | 0.2972 | 0.2963 |
| 3 | 0.4510 | 0.4340 |
| 4 | 0.2097 | 0.2254 |
| 5 | 0.1898 | 0.1930 |
| 6 | 0.7076 | 0.6646 |
| 7 | 0.2766 | 0.2767 |
| 8 | 0.3022 | 0.3130 |

Figure 2 visually demonstrates the difference in predictive probabilities between the frequentist and Bayesian analyses. In this figure, the results are stratified by smoking



(a) SMOKE = 0

(b) SMOKE = 1

Figure 2: Frequentist predictive probabilities with 95% Confidence Intervals and Bayesian predictive probabilities with 95% Credible Intervals as LWT increases by 20 pound intervals, stratified by smoking status (RACE=1, UI=0).

status. Each plot shows both the 95% frequentist confidence interval and the 95% Bayesian credible interval around the predictive probability as the weight at last menstrual period increases by 20 pound increments. This not only demonstrates the effects of smoking and weight on the probability of having a low birth weight baby, but also shows that in all cases, the credible intervals are shorter than the confidence intervals.

The posterior distributions obtained from the output data from the Bayesian analysis are given in Figure 3. These distributions reflect the parameter values of the final logistic model. Recall that the parameters are viewed as random variables, thus a distribution is returned, rather than a single parameter value. Summary statistics for the distributions are given in Table 4.

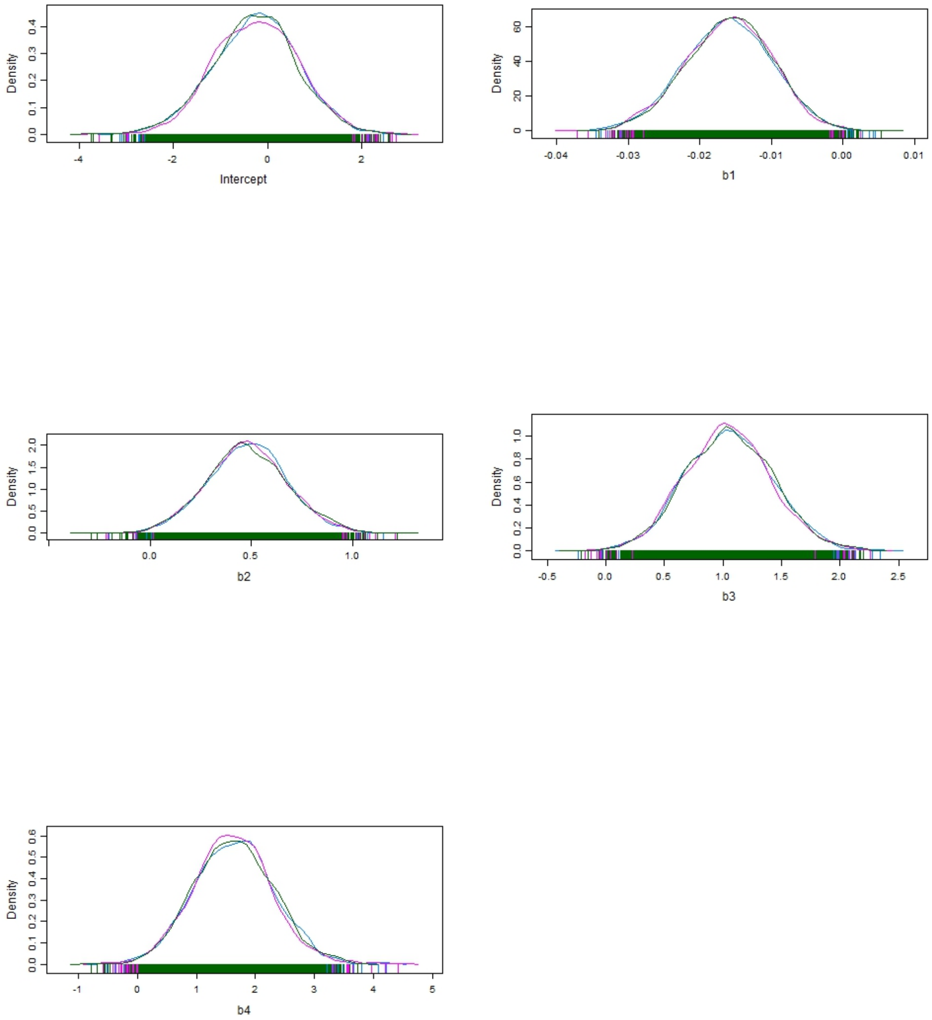


Figure 3: Posterior distributions for Low Birth Weight analysis parameters.

## Determining the Appropriate Confounder

### *Uterine Irritability as an Unmeasured Confounder*

In order to extend this analysis, it was necessary to identify a variable within the

Table 4: Summary Statistics for Low Birth Weight Analysis Parameters

| Parameter | Mean | Standard Deviation | 95% Credible Interval |
|---|---|---|---|
| b0 | -0.2537 | 0.8953 | (-2.0210 , 1.4660) |
| b1 | -0.0160 | 0.0060 | (-0.0282 , -0.0045) |
| b2 | 0.4803 | 0.1984 | (0.3478 , 0.8840) |
| b3 | 1.0493 | 0.3725 | (0.3308 , 1.7900) |
| b4 | 1.6309 | 0.6678 | (0.3620 , 2.9620) |

dataset that could be taken out of the model and re-inserted into the analysis as a simulated ”unmeasured” confounder. This methodology is meant to simulate a meta-analysis in which data on the confounder would be incorporated from a separate dataset. In the Backward Elimination procedure used to find the logistic model for the data, UI (uterine irritability) was the last insignificant variable to be taken out of the model. Because UI was taken out last, it was the most significant of the variables that were not included in the formal logistic model, thus it was a perfect candidate for the unmeasured confounder in the new analysis. The confounder was meant to be a variable that impacted the data, even though it wasn't included in the formal model. In this example, UI can be viewed as a variable that impacts the probability of a having low birth weight child, but was not measured in the dataset.

Unmeasured confounding introduces an unknown amount of uncertainty into any analysis of healthcare data in which not all variables that impact the outcome may be a part of the study. In Bayesian Sensitivity Analysis, the researcher assumes that there is one binary confounder that remains unmeasured in the data. Using Bayesian statistics, the researcher provides a loose prior distribution for the data based on prior information gained from external sources on the parameter. [12, 247]

In the method proposed by McCandless, the prior information on unmeasured confounding may be obtained from a measured confounder when the variables are known to be similar. [12, 248] In this example, the prior information, while originally from the same dataset, was used as if it were obtained from an outside source. If, perhaps, UI was not measured in the low birth weight datset, but it was thought to be a confounding factor for the outcome, it would be extremely beneficial to include it in the analysis somehow. Using McCandless' concept of the unmeasured confounder, the use of the variable measured in a separate dataset is justified.

*Bayesian Sensitivity Analysis including UI as an Unmeasured Confounder*

The analysis was conducted using a WinBUGS implementation of the Bayesian Sensitivity Analysis proposed by McCandless. [4] In this example, the same logistic model is used with the addition of an additional parameter, U, an unmeasured confounder. U is given the following prior distribution:

$$\beta_5 \sim \mathscr{N}(0, \frac{1}{\frac{log(6)}{1.96}^2}).$$

The use of $\frac{log(6)}{1.96}^2$ as the standard deviation of the prior gives a 95% chance that the odds ratio will be between 0.4 and 2.5, which is fairly uninformative. The details of this model can be found in Appendix C.

The Bayesian Sensitivity Analysis used the same 8 models of parameter values as the original Bayesian analysis in order to allow for the comparison of the two methods. If the predictive probabilities are altered by the addition of the unmeasured confounder, we may conclude that UI has some effect on the outcome variable, low birth weight.

Table 5: Bayesian Predictive Probabilities with and without Uterine Irritability as an Unmeasured Confounder

| Model | Without UI | Including UI |
|---|---|---|
| 1 | 0.1328 | 0.2217 |
| 2 | 0.2963 | 0.4277 |
| 3 | 0.4340 | 0.5842 |
| 4 | 0.2254 | 0.3371 |
| 5 | 0.1930 | 0.3063 |
| 6 | 0.6646 | 0.7785 |
| 7 | 0.2767 | 0.4089 |
| 8 | 0.3130 | 0.4414 |

Table 5 demonstrates the impact that the inclusion of UI as a confounder had on the predictive probabilities of the eight parameter settings. Every single predictive probability was increased by at least .1 due to the addition of the UI prior to the model. Mothers who exhibited uterine irritability are thus more likely to have a low birth weight child than those who did not experience UI, something that could not be predicted simply using the backward elimination technique to determine the model.

# CHAPTER THREE

# Prediction of Maternal Separation Anxiety

## The Data

For the second analysis I have chosen to use data that has not yet undergone a formal statistical analysis. These data were collected from 11 childcare centers in the Waco, Texas area during the 2011-2012 school year. Caregivers of children who attend each center (n=137) were asked to complete a survey ranking several factors of their home life as well as measures of parental separation anxiety. The home life indicators included such topics as: childcare satisfaction (”CCSTAT”), work and family factors (”WFF”), parental separation factors (”PARSS”), coping with work and family (”WFC”), etc. as well as demographics such as marital status and income. The survey contained several questions for each indicator which were summed to create totals for each category.

The originally continuous maternal separation anxiety (”PARSS”) indicator totals were dichotomized with those having PARSS totals greater than 125 coded as cases (value of 1) and those with PARSS totals less than or equal to 125 coded as zeros. This new binary variable will serve as the outcome variable for the logistic regression. Each of the predictor variables are continuous, as they are simply the summed totals of the responses for each category.

In this example, the fitted value represents the predicted probability of exhibiting maternal separation anxiety at specific parameter values.

## *Results of a Traditional Frequentist Analysis*

### *Model Specification*

As with the Low Birth Weight analysis, this data will be analyzed with both frequentist and Bayesian methods. In order to perform the frequentist analysis, we must first identify the logistic model for the data using the same Backward Elimination procedure as in Chapter 2.

The initial model includes every variable in the dataset. The full details of the Backward Elimination procedure can be found in Appendix D. The full model was narrowed down one by one until each remaining parameter was significant at the $\alpha = 0.10$ significance level.

The final logit for the data includes CCSTAT, WFF, and WFC, and is given by:

$$logit(\mu) = -6.55210 + 0.04846(CCSTAT) + 0.25161(WFF) + .11579(WFC). \tag{8}$$

The final logistic model is given by:

$$\hat{\pi} = \frac{e^{-6.55210+0.04846(CCSTAT)+0.25161(WFF)+.11579(WFC)}}{1+e^{-6.55210+0.04846(CCSTAT)+0.25161(WFF)+.11579(WFC)}} \tag{9}$$

### *Odds Ratios and Predictive Probabilities*

The odds ratios and 90% Confidence Intervals for each predictor in the final model are given in Table 6. For example, a one unit increase in the score for WFF results in the odds of parental separation anxiety increasing by a factor of 1.2861, on average, with all other parameters held constant.

Table 6: Odds Ratios and Confidence Intervals for Regression Parameters

| Characteristic | OR | 95% CI |
|---|---|---|
| CCSTAT | 1.0497 | (1.0097 , 1.0960) |
| WFF | 1.2861 | (1.1396 , 1.4739) |
| WFC | 0.8907 | (0.7968 , 0.9915) |

The final model can also be used to calculate the predictive probabilities for several sets of parameter values. In this example, the predictive probability represents the probability of parental separation anxiety ($PARSS > 125$) for a parent with specific CCSTAT, WFF, and WFC totals. For example, the predictive probability of parental separation anxiety for a parent who answered each question with the mean totals of every survey is 0.2025, as shown in equation 10. Thus, the probability of a parent with these scores having parental separation anxiety is 20.25%.

$$\frac{e^{-6.55210+.04846(68)+.25161(19)-.11579(25)}}{1+e^{-6.55210+.04846(68)+.25161(19)-.11579(25)}} = 0.2025 \qquad (10)$$

Table 7 provides a list of parameters for each model that will be used throughout the analysis of the maternal separation data. Models will be referred to by their number rather than their set of parameter values from this point on.

Table 7: Model Parameter Values

| | CCSTAT | WFF | WFC |
|---|---|---|---|
| Model 1 | 68 | 19 | 25 |
| Model 2 | 61 | 17 | 23 |
| Model 3 | 77 | 21 | 28 |
| Model 4 | 68 | 17 | 23 |
| Model 5 | 61 | 21 | 23 |
| Model 6 | 61 | 17 | 28 |
| Model 7 | 77 | 19 | 25 |
| Model 8 | 68 | 21 | 25 |
| Model 9 | 68 | 19 | 28 |

## Results of the Bayesian Analysis

### *Model Specification and Parameter Values*

As with the Low Birth Weight data, the Bayesian analysis will be conducted using OpenBUGS models with three chains. The same final model that was used in the frequentist analysis will be used for the Bayesian analysis. The variables included in the model are CCSTAT, WFF, and WFC. This analysis uses the same loose priors that were put on the low birth weight parameters. Each parameter in the analysis was given a prior that follows the same normal distribution,

$$\beta_i \sim \mathscr{N}(0,4)$$

demonstrated in Figure 4.

An OpenBUGS model was created to represent the data plus each parameter's prior distribution. The details of this model are given in Appendix E. Odds ratios for each parameter and predictive probabilities for each model were calculated using the OpenBUGS model.

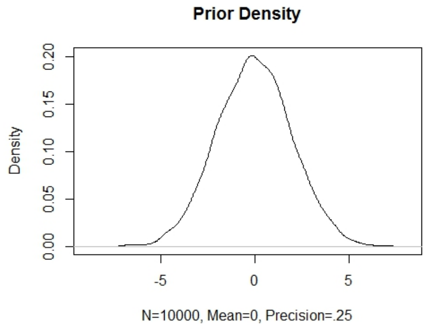


Figure 4: Prior distribution for CCSTAT, WFF, and WFC

*Comparison of Predictive Probabilities*

Table 8 gives the predictive probabilities for each of the nine parameter models for both the frequentist and Bayesian analyses.

Table 8: Model Predictive Probabilities for Frequentist and Bayesian Analysis

| Model | Frequentist | Bayesian |
|---|---|---|
| 1 | 0.2025 | 0.2183 |
| 2 | 0.1211 | 0.1699 |
| 3 | 0.3145 | 0.2600 |
| 4 | 0.1621 | 0.1913 |
| 5 | 0.2738 | 0.3311 |
| 6 | 0.0717 | 0.0979 |
| 7 | 0.2819 | 0.2536 |
| 8 | 0.2957 | 0.3030 |
| 9 | 0.1521 | 0.1586 |

Posterior probabilities obtained from the Bayesian analysis are given in Figure 5 below. Summary statistics for the parameters are shown in Table 9.

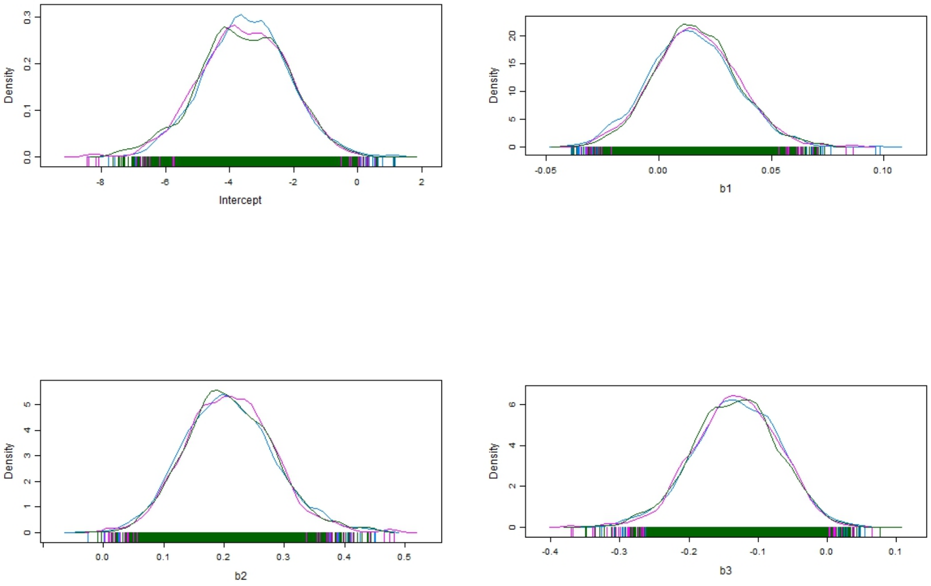


Figure 5: Posterior distributions for maternal separation anxiety analysis parameters.

Table 9: Summary Statistics for Maternal Separation Anxiety Analysis Parameters

| Parameter | Mean | Standard Deviation | 90% Credible Interval |
|---|---|---|---|
| b0 | -3.4275 | 1.4000 | (-5.5230 , -0.9611) |
| b1 | 0.0204 | 0.0185 | (-0.0064 , 0.0527) |
| b2 | 0.2264 | 0.0726 | (0.1127 , 0.3341) |
| b3 | -0.1421 | 0.0618 | (-0.2303 , -0.0349) |

## Determining the Appropriate Confounder

### *Marital Status as an Unmeasured Confounder*

Marital status was chosen as the simulated unmeasured confounder due to its predicted effect on the outcome. Logically speaking, one would expect that marital status of the mother would have an effect on her attachment to her child. Marital status was not previously included in the model, as it is a characteristic, rather than a sum of survey answers.

Each mother chose her marital status from the following options: married, partnered, single, divorced, widowed, and separated. In order to dichotomize the confounder, those who responded as married or partnered were given a ”0”, and those who answered single, divorced, widowed, or separated were given a ”1” for the confounder, marital status. Thus, the data was stratified into two groups: 1-parent households and 2-parent households.

*Bayesian Sensitivity Analysis including Marital Status as an Unmeasured Confounder*

The analysis was conducted using the same general model used for the Low Birth Weight data. The additional parameter, U, has an informative prior distribution given by

$$\beta_4 \sim \mathscr{N}(0, \frac{1}{\frac{log(6)}{1.96}^2}).$$

The details of this model can be found in Appendix F.

The Bayesian analysis with the confounder was also run with the same nine parameter values as the frequentist analysis. Table 10 gives the predictive probabilities for the Bayesian model for 2-parent households and 1-parent households. As with the Low Birth Weight data, we observe that the addition of Marital Status as a confounder significantly increases the predictive probabilities for maternal separation anxiety. In particular, parents whose scores fall under Model 5 (CCSTAT=61, WFF=21, WFC=23) and are partnered are 33.11% likely to develop separation anxiety, while their single counterparts are 45.77% likely, a 12% difference.

Figure 6 demonstrates the effect of a single variable, WFF, on the likelihood of experiencing parental separation anxiety. In every case, single parents are significantly

Table 10: Bayesian Predictive Probabilities with and without Marital Status as an Unmeasured Confounder

| Model | Partnered | Single |
|---|---|---|
| 1 | 0.2183 | 0.3193 |
| 2 | 0.1699 | 0.2664 |
| 3 | 0.2600 | 0.3549 |
| 4 | 0.1913 | 0.2876 |
| 5 | 0.3311 | 0.4577 |
| 6 | 0.0979 | 0.1611 |
| 7 | 0.2536 | 0.3510 |
| 8 | 0.3030 | 0.4174 |
| 9 | 0.1586 | 0.2402 |

more likely to experience separation anxiety. For WFF scores of 19 or higher, single parents are about twice as likely to develop separation anxiety than their partnered counterparts.

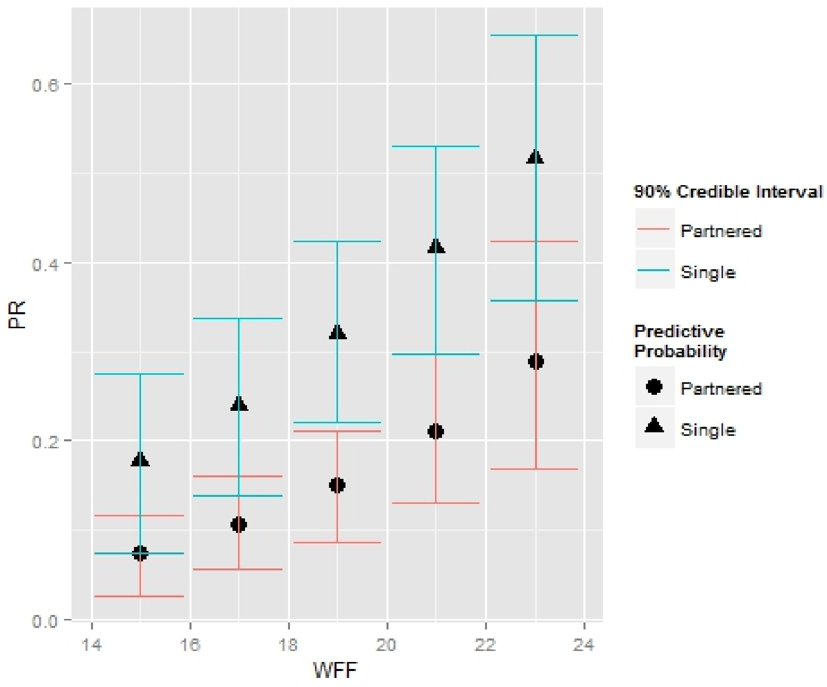


Figure 6: Predictive probabilities with 90% Credible Intervals as WFF increases by multiples of 2 points, stratified by Marital Status

Figure 7 demonstrates the difference between Frequentist and Bayesian prediction across all 9 parameter models. In this case, while the Bayesian estimates are higher than Frequentist in most cases, it turns out that at very high CCSTAT scores, the Frequentist estimates are higher. More research into the true prior distribution of CCSTAT scores may shed more light into the reasons behind the switch. In most cases, the Bayesian credible intervals are slightly shorter than the Frequentist confidence intervals.

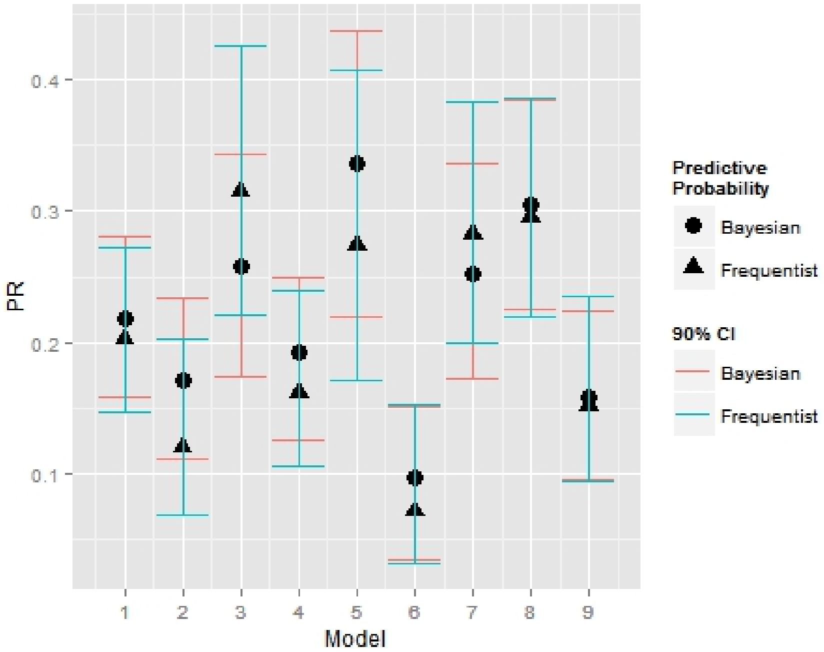


Figure 7: Comparison of Predictive probabilities with 90% Frequentist confidence intervals and 90% Bayesian credible intervals

# CHAPTER FOUR

## Discussion and Further Research

While this paper does not venture an argument in favor of one method over the other, it has revealed significant differences between Frequentist and Bayesian analyses. In many cases, the Frequentist and Bayesian estimates of the predictive probabilities turn out very different. However, simulation studies would be necessary in order to test which method performed most accurately.

The most obvious argument for Bayesian statistics in the context of this paper is its ability to incorporate outside sources of information in the form of a meta-analysis. Particularly, we explore Bayesian statistics' ability to incorporate prior information about a suspected unmeasured confounder into the prediction of the outcome variable.

This project may be expanded with different datasets where a true meta-analysis could be performed. In a situation where pertinent data is not included in the primary dataset, data from a separate analysis may be included as an unmeasured confounder as shown in this paper. The maternal separation anxiety data, in particular, may benefit from the inclusion of separate parameters that may not have been included in the original survey. As these methods are expanded and tested, the inclusion of more than one confounder will also prove helpful in prediction.

# APPENDICES

# Appendix A

## Low Birth Weight Frequentist Analysis

### *Model Specification and Odds Ratios*

```r
mydata = read.csv("C:/Users/Allison/Desktop/THESIS/lowbwt.csv",header=T)
#full model including all original variables
mylogit <- glm(LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI + FTV,
               data = mydata, family = "binomial")
summary(mylogit)
#the least significant variable, FTV, was removed
mylogit2 <- glm(LOW ~ AGE + LWT + RACE + SMOKE + PTL + HT + UI,
                data = mydata, family = "binomial")
summary(mylogit2)
#the least significant variable, AGE, was removed
mylogit3 <- glm(LOW ~ LWT + RACE + SMOKE + PTL + HT + UI, data = mydata,
                family = "binomial")
summary(mylogit3)
#the least significant variable, PTL, was removed
mylogit4 <- glm(LOW ~ LWT + RACE + SMOKE + HT + UI, data = mydata,
                family = "binomial")
summary(mylogit4)
#the least significant variable, UI, was removed, leaving all
#remaining variables significant at the .05 significance level
final_logit <- glm(LOW ~ LWT + RACE + SMOKE + HT, data = mydata,
                   family = "binomial")
summary(final_logit)
#95% confidence intervals for each parameter in the final model
```

```r
exp(cbind(OR = coef(final_logit), confint(final_logit)))
```

### *Predictive Probabilities and 95% CI for Low Birth Weight Analysis*

```r
#function to evaluate predictive probability and 95% confidence interval
pointwise.logit <- function(glm.obj, newdata, coverage=0.95)
{
  lp <- predict(glm.obj, newdata, se=T)
  a <- (1+coverage)/2
  margin.error <- lp$se.fit / lp$residual.scale * qnorm(a)
  lp.upper <- lp$fit + margin.error
  lp.lower <- lp$fit - margin.error
  upper <- exp(lp.upper)
  upper <- upper/(1+upper)
  lower <- exp(lp.lower)
  lower <- lower/(1+lower)
  fit <- exp(lp$fit)
  fit <- fit/(1+fit)
  list(upper=upper, fit=fit,lower=lower)
}
#parameter values plug in for LWT, RACE, SMOKE, HT for specific
#predictive probabilities
pointwise.logit(final_logit, data.frame(LWT=200, RACE=1, SMOKE=1, HT=0),
                coverage=.95)
```

# Appendix B

## Low Birth Weight Bayesian Analysis without Confounding

### *OpenBUGS Model without Confounding*

```
model{
  for (i in 1:n) {
    #Linear regression on logit
    logit(p[i]) <- bo + b1 * x[i] + b2 * r[i] + b3 * s[i] + b4 * h[i]
    #likelihood function for each data point
    y[i] ~dbern(p[i])
  }
  or1 <- exp(b1)
  or2 <- exp(b2)
  or3 <- exp(b3)
  or4 <- exp(b4)
  #probability of remission at any given level
  pr <- exp(bo + b1 * 135 + b2 * 1 + b3 * 0 + b4 * 0)/
           (1 + exp(bo + b1 * 135 + b2 * 1 + b3 * 0 + b4 * 0))
  bo ~ dnorm ( 0.0, 0.25) #prior for intercept
  b1 ~ dnorm (0.0, 0.25) #prior for LWT
  b2 ~ dnorm (0.0, 0.25) #prior for RACE
  b3 ~ dnorm ( 0.0, 0.25) #prior for SMOKE
  b4 ~ dnorm (0.0, 0.25) #prior for HT
}
```

### *R2OpenBUGS Code without Confounding*

```
library(arm)
library(BRugs)
library(R2OpenBUGS)
library(xlsx)
library(ggplot2)
library(plyr)
n <- nrow(df2)
#mu.gamma = c(0,0)
y <- df2$parssval
x <- df2$ccstat_total
r <- df2$wff_total
s <- df2$wfc_total
data <- #data
  list(
    #response
    y=c(),
    #weight at last menstrual period
    x=c(),
    #race
    r=c(),
    #smoking status
    s=c(),
    #hypertension
    h=c(),
    #sample size
    n=189)
```

```
parameters <- c("or1", "or2", "or3", "or4", "pr", "bo", "b1", "b2", "b3", "b4")
#run simulation, save coda
lbwt.sim <- bugs(data, inits=NULL, parameters, model.file="lbwtmodel.txt", n.chains=3,
                      n.iter=10000, n.burnin=6000, codaPkg = TRUE)
#coda from simulation
lbwt.coda <- read.bugs(lbwt.sim)
#posterior for b1
b1 <- lbwt.coda[,1]
densityplot(b1, xlab="b1")
#posterior for b2
b2 <- lbwt.coda[,2]
densityplot(b2, xlab="b2")
#posterior for b3
b3 <- lbwt.coda[,3]
densityplot(b3, xlab="b3")
#posterior for b4
b3 <- lbwt.coda[,4]
densityplot(b4, xlab="b4")
#posterior for intercept, b0
bo <- lbwt.coda[,4]
densityplot(bo, xlab="Intercept")
#summary statistics for the simulation
summary(lbwt.coda)
```

# Appendix C

## Low Birth Weight Bayesian Analysis with Confounding

### *OpenBUGS Model with Confounding*

```
model
{
  for(i in 1:n){

#Bernoulli response for outcome
y[i] ~ dbern(p[i])

#Bernoulli response for unobserved confounder
U[i] ~ dbern(q[i])

#logit model for outcomes
logit(p[i]) <- bo + b1 * x[i] + b2 * r[i] + b3 * s[i] + b4 * h[i]
                   + lambda * U[i]

#logit model for unobserved confounder
logit(q[i]) <- gamma[1] + gamma[2] * s[i] + psi * r[i]

}

#diffuse normals for intercept, observed covariates for outcome model

bo ~ dnorm ( 0.0, 0.25) #prior for intercept
b1 ~ dnorm (0.0, 0.25) #prior for LWT
```

```
b2 ~ dnorm (0.0, 0.25) #prior for RACE

b3 ~ dnorm ( 0.0, 0.25) #prior for SMOKE

b4 ~ dnorm (0.0, 0.25) #prior for HT


#informative normals for bias parameters


lambda ~ dnorm(0, c1pr)

gamma[1:2] ~ dmnorm(mu.gamma[], R[ , ])

psi ~ dnorm(0, c4pr)


c1 <- log(6)/1.96

c2 <- log(6)/1.96

c3 <- log(6)/1.96

c4 <- log(6)/1.96


c1pr <- pow(c1, -2)

c4pr <- pow(c4, -2)


c23 <- -c3 * c3/2


R[1:2,1:2] <- inverse(S[1:2,1:2])


S[1,1] <- c2 * c2

S[2,2] <- c3 * c3

S[1,2] <- c23

S[2,1] <- c23

mu.gamma[1]<-0
mu.gamma[2]<-0


or1 <- exp(b1)
or2 <- exp(b2)
or3 <- exp(b3)
or4 <- exp(b4)
#m <- bo/b1
#probability of remission at any given level
pr <- exp(bo + b1 * 135 + b2 * 1 + b3 * 0 + b4 * 0 + lambda * 1)/
      (1 + exp(bo + b1 * 135 + b2 * 1 + b3 * 0 + b4 * 0 + lambda * 1))
}
```

### *R2OpenBUGS Code with Confounding*

```
library(arm)
library(BRugs)
library(R2OpenBUGS)
library(xlsx)
library(ggplot2)
library(plyr)
data <- #data
  list(
    #response
    y=c(),
    #weight at last menstrual period
    x=c(),
    #race
    r=c(),
    #smoking status
    s=c(),
    #hypertension
    h=c(),
    #uterine irritability-the unmeasured confounder
    u=c(),
    #sample size
    n=189)
mu.gamma = c(0,0)
parameters <- c("or1", "or2", "or3", "or4", "pr", "bo", "b1", "b2",
                "b3", "b4")
lbwt.sim <- bugs(data, inits=NULL, parameters, model.file="lbwtmodel.txt",
                 n.chains=3, n.iter=10000, n.burnin=6000)
lbwt.sim$summary
#record coda
lbwt.coda <- read.bugs(lbwt.sim)
#posterior for b1
b1 <- lbwt.coda[,1]
densityplot(b1, xlab="b1")
#posterior for b2
b2 <- lbwt.coda[,2]
densityplot(b2, xlab="b2")
#posterior for b3
b3 <- lbwt.coda[,3]
densityplot(b3, xlab="b3")
```

```r
#posterior for b4

b4 <- lbwt.coda[,4]

densityplot(b4, xlab="b4")

#posterior for intercept, b0

bo <- lbwt.coda[,5]

densityplot(bo, xlab="Intercept")
```

# Appendix D

## Maternal Separation Anxiety Frequentist Analysis

### *Separation Anxiety Data Manipulation, Model Specification, and Odds Ratios with 90% CI*

```r
library(ggplot2)

library(plyr)

library(xlsx)

data=read.xlsx("childcare.xlsx",sheetIndex=1,header=T,as.data.frame=T)

df2=mutate(data, parssval = 0)

#dichotomize the response variable at value parsstotal=125

for(i in 1:length(df2)){

  if(df2$parsstotal[i] > 125){

    df2$parssval[i] <- 1

  }else if (df2$parsstotal[i] <= 125){

    df2$parssval[i] <- 0

  }

}

fit <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +

           df2$wfc_total + df2$wfmp_total + df2$ff_total +

           df2$sstotal + df2$cdpk_total + df2$na_total,

           family="binomial")

summary(fit)

#model with ff removed

fit2 <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +

            df2$wfc_total + df2$wfmp_total + df2$sstotal +

            df2$cdpk_total + df2$na_total, family="binomial")
```

```
summary(fit2)
#model with wfmp removed
fit3 <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +
            df2$wfc_total + df2$sstotal + df2$cdpk_total + df2$na_total,
            family="binomial")
summary(fit3)
#model with cdpk removed
fit4 <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +
            df2$wfc_total + df2$sstotal + df2$na_total, family="binomial")
summary(fit4)
#model with ss removed
fit5 <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +
            df2$wfc_total + df2$na_total, family="binomial")
summary(fit5)
#model with na removed--final model--all variables significant at .1 alpha level
fit6 <- glm(df2$parssval ~ df2$ccstat_total + df2$wff_total +
            df2$wfc_total, family="binomial")
summary(fit6)
#odds ratios for coefficients with 90% confidence interval
exp(cbind(OR = coef(fit6), confint(fit6, level=0.90)))
```

# Appendix E

## Maternal Separation Anxiety Bayesian Analysis without Confounding

### *OpenBUGS Model without Confounding*

```
model{
  for (i in 1:n) {
    #Linear regression on logit
    logit(p[i]) <- bo + b1 * x[i] + b2 * r[i] + b3 * s[i]
    #likelihood function for each data point
    y[i] ~ dbern(p[i])
  }

  or1 <- exp(b1)
  or2 <- exp(b2)
  or3 <- exp(b3)

  #probability of remission at any given level
  pr <- exp(bo + b1 * 68 + b2 * 19 + b3 * 28)/
           (1 + exp(bo + b1 * 68  + b2 * 19 + b3 * 28))

  bo ~ dnorm ( 0.0, 0.25) #prior for intercept
  b1 ~ dnorm (0.0, 0.25) #prior for ccstat
  b2 ~ dnorm (0.0, 0.25) #prior for wff
  b3 ~ dnorm ( 0.0, 0.25) #prior for wfc


}
```

*R2OpenBUGS Code without Confounding*

```
library(arm)
library(BRugs)
library(R2OpenBUGS)
library(xlsx)
library(ggplot2)
library(plyr)
library(coda)
childcare <- read.xlsx("childcare.xlsx", sheetIndex=1, header=TRUE,
                       as.data.frame=T)
df2=mutate(childcare, parssval = 0)
for(i in 1:length(df2)){
  if(df2$parsstotal[i] > 125){
    df2$parssval[i] <- 1
  }else if (df2$parsstotal[i] <= 125){
    df2$parssval[i] <- 0
  }
}
n <- nrow(df2)
mu.gamma = c(0,0)
y <- df2$parssval
x <- df2$ccstat_total
r <- df2$wff_total
s <- df2$wfc_total
data <- list("n", "y", "x", "r", "s")
parameters <- c("bo","b1", "b2", "b3", "or1", "or2", "or3", "pr")
childcare.sim <- bugs(data, inits=NULL, parameters, model.file="childcare.txt",
                      n.chains=3, n.iter=10000, n.burnin=6000, codaPkg=TRUE)

childcare.coda <- read.bugs(childcare.sim)

#calculates summary statistics for each variable

summary(childcare.coda)

#calculates 90% Bayesian Credible Intervals for each chain

HPDinterval(childcare.coda,prob=0.9)
```

# Appendix F

## Maternal Separation Anxiety Bayesian Analysis with Confounding

### *OpenBUGS Model with Confounding*

```
model
{
  for(i in 1:n){

   #Bernoulli response for outcome
y[i] ~ dbern(p[i])

#Bernoulli response for unobserved confounder
U[i] ~ dbern(q[i])

#logit model for outcomes
logit(p[i]) <- bo + b1 * x[i] + b2 * r[i] + b3 * s[i] + lambda * U[i]

#logit model for unobserved confounder
logit(q[i]) <- gamma[1] + gamma[2] * s[i] + psi * r[i]

}

#diffuse normals for intercept, observed covariates for outcome model

bo ~ dnorm ( 0.0, 0.25) #prior for intercept
b1 ~ dnorm (0.0, 0.25) #prior for ccstat
b2 ~ dnorm (0.0, 0.25) #prior for wff
```

```
b3 ~ dnorm ( 0.0, 0.25) #prior for wfc

#informative normals for bias parameters

lambda ~ dnorm(0, c1pr)
gamma[1:2] ~ dmnorm(mu.gamma[], R[ , ])
psi ~ dnorm(0, c4pr)

c1 <- log(6)/1.96
c2 <- log(6)/1.96
c3 <- log(6)/1.96
c4 <- log(6)/1.96

c1pr <- pow(c1, -2)
c4pr <- pow(c4, -2)

c23 <- -c3 * c3/2

R[1:2,1:2] <- inverse(S[1:2,1:2])

S[1,1] <- c2 * c2
S[2,2] <- c3 * c3
S[1,2] <- c23
S[2,1] <- c23
mu.gamma[1]<-0
mu.gamma[2]<-0
or1 <- exp(b1)
```

```
or2 <- exp(b2)
or3 <- exp(b3)
#probability of remission at any given level
pr <- exp(bo + b1 * 68 + b2 * 23 + b3 * 25  + lambda * 1)/
         (1 + exp(bo + b1 * 68 + b2 * 23 + b3 * 25  + lambda * 1))
}
```

### *R2OpenBUGS Code with Confounding*

```
library(arm)
library(BRugs)
library(R2OpenBUGS)
library(xlsx)
library(ggplot2)
library(plyr)
library(coda)
childcare <- read.xlsx("childcare.xlsx", sheetIndex=1, header=TRUE,
                       as.data.frame=T)
df2=mutate(childcare, parssval = 0)
for(i in 1:length(df2)){
  if(df2$parsstotal[i] > 125){
    df2$parssval[i] <- 1
  }else if (df2$parsstotal[i] <= 125){
    df2$parssval[i] <- 0
  }
}
df2=mutate(df2, house = 0)
```

```r
for(i in 1:length(df2)){
  if(df2$ms[i] == 1){
    df2$house[i] <- 0 }
  if(df2$ms[i] == 2){
    df2$house[i] <- 0 }
  if(df2$ms[i] == 3){
    df2$house[i] <- 1 }
  if(df2$ms[i] == 4){
    df2$house[i] <- 1 }
  if(df2$ms[i] == 5){
    df2$house[i] <- 1
  }else if (df2$ms[i] == 6){
    df2$house[i] <- 1
  }
}
n <- nrow(df2)
mu.gamma = c(0,0)
y <- df2$parssval
x <- df2$ccstat_total
r <- df2$wff_total
s <- df2$wfc_total
U <- df2$house
data <- list("n", "y", "x", "r", "s", "U")
inits <- c(list(bo = 0, b1 = 1, b2 = 1, b3 = 1),
           list(bo = 1, b1 = 2, b2 = 2, b3 = 2),
           list(bo = .5, b1 = .5, b2 = .5, b3 = .5))
parameters <- c("or1", "or2", "or3", "pr")
```

```
childcare.sim <- bugs(data, inits=NULL, parameters,
                      model.file="childcareconfound.txt", n.chains=3,
                      n.iter=10000, n.burnin=6000, codaPkg=TRUE)
childcare.coda <- read.bugs(childcare.sim)
#calculates summary statistics for each variable
summary(childcare.coda)
#calculates 90% Bayesian Credible Intervals for each chain
HPDinterval(childcare.coda,prob=0.9)
```

# REFERENCES

[1] Bolstad, W. M., 2010. *Understanding Computational Bayesian Statistics*. New York: John Wiley & Sons, Inc.

[2] Brookhart, Sturmer, et al. June 2010. Confounding Control in Healthcare Database Research: Challenges and Potential Approaches. *Medical Care* 48(6.1).

[3] Crombie I. and Davies, H. 2009. What is meta-analysis? *Hayward Medical Communications*.

[4] Faries D, Peng X, Pawaskar M, Price K, Stamey JD, Seaman JW Jr. 2013. Evaluating the Impact of Unmeasured Confounding with Internal Validation Data: An Example Cost Evaluation in Type 2 Diabetes. *Value in Health* 16(2): 259-66.

[5] Fewell, Z., Smith, G. D., and Sterne, J. A. C. 2007. The Impact of Residual and Unmeasured Confounding in Epidemiologic Studies: A Simulation Study. *American Journal of Epidemiology* 166(6): 646-655.

[6] Hock, E., McBride, S., Gnezda, M.T., (1989). Maternal separation anxiety: Mother infant separation from the maternal perspective. *Child Development* 60: 793-802.

[7] Hosmer, D. W. and S. Lemeshow. 2000. *Applied Logistic Regression: Second Edition*. New York: John Wiley \& Sons, Inc.

[8] Hosmer, D. W. and S. Lemeshow. 2000. Low Birth Weight Data. Available: http://www.umass.edu/statdata/statdata/data/lowbwt.txt.

[9] Kruschke, J. K. 2011. *Doing Bayesian Data Analysis*. Burlington, MA: Academic Press.

[10] Lavine, M. 2000. *What is Bayesian statistics and why everything else is wrong* [Online]. Available: http://www.math.umass.edu/~lavine/whatisbayes.pdf [2012, October 10].

[11] Malakoff, D. 1999. Statistics: Bayes Offers a 'New' Way to Make Sense of Numbers. *Science*. 286(5444): 1460-1464.

[12] McCandless, L., Gustafson, P., Levy, A. R., & Richardson, S. 2010. Hierarchical Priors for Bias Parameters in Bayesian Sensitivity Analysis for Unmeasured Confounding. *International Journal of Environmental Research and Public Health*. 7: 1520-1539.

[13] R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[14] Steenland, K. and Greenland, S. 2004. Monte Carlo sensitivity analysis and Bayesian analysis of smoking as an unmeasured confounder in a study of silica and lung cancer. *American Journal of Epidemiology* 160(4):384-92.