Estimation of the mitigated fraction from ordinal data in the evaluation of vaccine efficacy

by

Steephanson Anthonymuthu

B.S., Rajarata University of Sri Lanka, 2010
M.S., Sam Houston State University, 2015

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Vaccine efficacy can be established through the estimation of several numerical measures. A common measure of efficacy for vaccines, especially those designed to prevent a given disease, is the prevented fraction. Unfortunately, the prevented fraction can be used only when the outcome is dichotomous. It is worth noting that some useful vaccines reduce the severity of the targeted disease rather than entirely prevent its occurrence. The concept of the mitigated fraction was introduced in veterinary medicine to quantify the reduction in the severity of disease occurring in vaccinated animals as compared to non-vaccinated animals. The USDA's Center for Veterinary Biologics (CVB) recommends a form of the mitigated fraction proposed by Siev (2005) which can be easily calculated when the disease severity can be graded by some continuous measure or by some discrete assessment resulting in unambiguous ranks. Current CVB guidance suggests that the mitigated fraction be estimated non-parametrically via the use of the Wilcoxon rank sum statistics.

A survey of recent literature indicates a growing interest in measures of efficacy when the outcome variable is ordinal, especially when observations are clustered or measured longitudinally. Here, a parametric approach assuming a generalized linear mixed model (GLMM) with latent variable is developed for data collected in a completely randomized design (CRD) or a randomized complete block design (RCBD) and is then evaluated through simulation. Results show this parametric approach works well for both the CRD and RCBD. The GLMM approach can be extended to studies where more than two treatments are compared whereas the method of Siev (2005) can handle only two treatment groups (vaccinated and non-vaccinated). Furthermore, a Bayesian statistical approach has been briefly explored to estimate the mitigated fraction from an ordinal response observed in a completely randomized design.

Extension of this Bayesian statistical approach for vaccine trials will also be discussed as future work.

Estimation of the mitigated fraction from ordinal data in the evaluation of vaccine efficacy

by

Steephanson Anthonymuthu

B.S., Rajarata University of Sri Lanka, 2010
M.S., Sam Houston State University, 2015

A DISSERTATION

submitted in partial fulfillment of the requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Statistics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Christopher I. Vahl

Copyright

Abstract

Vaccine efficacy can be established through the estimation of several numerical measures. A common measure of efficacy for vaccines, especially those designed to prevent a given disease, is the prevented fraction. Unfortunately, the prevented fraction can be used only when the outcome is dichotomous. It is worth noting that some useful vaccines reduce the severity of the targeted disease rather than entirely prevent its occurrence. The concept of the mitigated fraction was introduced in veterinary medicine to quantify the reduction in the severity of disease occurring in vaccinated animals as compared to non-vaccinated animals. The USDA's Center for Veterinary Biologics (CVB) recommends a form of the mitigated fraction proposed by Siev (2005) which can be easily calculated when the disease severity can be graded by some continuous measure or by some discrete assessment resulting in unambiguous ranks. Current CVB guidance suggests that the mitigated fraction be estimated non-parametrically via the use of the Wilcoxon rank sum statistics.

A survey of recent literature indicates a growing interest in measures of efficacy when the outcome variable is ordinal, especially when observations are clustered or measured longitudinally. Here, a parametric approach assuming a generalized linear mixed model (GLMM) with latent variable is developed for data collected in a completely randomized design (CRD) or a randomized complete block design (RCBD) and is then evaluated through simulation. Results show this parametric approach works well for both the CRD and RCBD. The GLMM approach can be extended to studies where more than two treatments are compared whereas the method of Siev (2005) can handle only two treatment groups (vaccinated and non-vaccinated). Furthermore, a Bayesian statistical approach has been briefly explored to estimate the mitigated fraction from an ordinal response observed in a completely randomized design.

Extension of this Bayesian statistical approach for vaccine trials will also be discussed as future work.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to express my wholehearted gratitude to Dr. Christopher I. Vahl, my major professor, for all his knowledge, guidance, and suggestions. He always encourages and motivates me throughout my entire Ph.D. studies. I especially thank Dr. Gyuhyeong Goh, Dr. Michael Higgins, Dr. Qing Kang and Dr. Steve Dritz for serving on my committee, and for their valuable guidance through my study process. Also, I would like to extend my gratitude to Dr. Majid Jaberi-Douraki for his willingness to serve as the outside chairperson in the examining committee for my doctoral degree.

I would also like to thank the entire faculty, and staff in the Department and thank all the professors for their excellent courses and for their valuable knowledge. Especially, Dr. James Neill, Dr. Juan Du, Dr. Gary Gadbury, Dr. Weixing Song and Dr. Cen Wu who taught me and helped me during my studies Statistics at Kansas State University.

Special thanks go to all current and former students, of whom which there are too many to mention, who influenced my statistical education. In particular, I thank my classmate, Brianna Cunningum, for the unforgettable memories and valuable discussions we had.

Once again, I want to express my sincerest gratitude to the Department for their kindness, providing me with the opportunity to complete my Ph.D. degree.

Thanks are due foremost to my wonderful wife, Sanjeewani, and my lovely daughter, Sarah Sanujini, for supporting me throughout my entire education career. Thank you Sanjeewani for your patience, sacrifices and all your support. Since we both were statisticians, it was really an unforgettable and valuable discussions and arguments we had over statistics.

Dedication

From the bottom of my heart, I sincerely dedicate this work to my father, my mother, my sisters, my wife, and my daughter.

Chapter 1 - Introduction

A vaccine is generally a fluid containing weakened or dead microorganisms known to cause disease. A vaccine is typically applied to an individual subject by injection but may be taken orally or via a nasal mist. The main purpose of vaccination is to make the immune system in vaccinated subjects ready in advance to prevent or cure an infection. Vaccination (a way of getting a vaccine) is one of the most remarkable health benefits and became popular with the great achievement of a smallpox vaccine and now with the current pandemic of Covid-19. The primary focus has been on how well the vaccine protects the vaccinated subjects compared with non-vaccinated subjects. Clinical trials are conducted to evaluate vaccine efficacy (VE). The "efficacy" of a vaccine refers to its ability to either prevent infection or reduce the incidence and/or severity of the associated disease in the target population (Mehrota, 2006).

1.1 Relative risk and vaccine efficacy

Here some of the technical terms and definitions in vaccine fields are briefly discussed. Suppose a subject is selected at random from a population and let A and B denote two events. Specifically, let A be the event that the subject is vaccinated. Then the complement of A, denoted by $\bar{A}$, is the event that the subject is not vaccinated. Let B be the event that subject has a certain health-related characteristic, e.g., the subject develops a disease such as Covid-19. Now let $P(B|A)$ denote the **risk** of the occurrence of B given A, i.e. the risk that a vaccinated subject develops Covid-19. Then $P(B|\bar{A})$ denotes the risk of the occurrence of B given $\bar{A}$, i.e. the risk that an unvaccinated subject develops Covid-19. The two groups of subjects defined by A and $\bar{A}$, e.g. the vaccinated and unvaccinated subjects, are often compared through the relative risk (RR) which is defined as the ratio of $P(B|A)$ to $P(B|\bar{A})$, i.e.

$$RR = \frac{P(B|A)}{P(B|\bar{A})}.$$

Halloran and Longini (1997) define vaccine efficacy (VE) to be a measure based on the relative risk of a particular outcome of the form $VE = 1 - RR$.

Following Centers for Disease Control and Prevention (CDC), it can clearly be understood that VE is measured by calculating the risk of disease among vaccinated and unvaccinated subjects and determining the percentage reduction in risk of disease among vaccinated subjects relative to unvaccinated subjects. The greater the percentage reduction of disease in the vaccinated group, the greater the vaccine efficacy (CDC, 2012). A vaccine is 100% efficacious if $VE = 1$.

1.2 Prevented Fraction (PF)

Medical interventions are intended to either reduce or prevent disease. Statistical tools are used to estimate the effect of an intervention on disease reduction or prevention. As Siev (2005) stated, when a medical intervention is intended to prevent a dichotomous outcome (e.g., clinical signs of disease are observed on a subject as presence or absence), an estimator known as the prevented fraction (PF) is commonly used to measure the effect of the intervention on preventing disease. As the usual estimator for efficacy of vaccine, PF is often simply termed vaccine efficacy (VE) in vaccine studies (Siev ,2005). It is to be clearly understood that the prevented fraction (PF) is commonly used for interventions other than vaccines whereas vaccine efficacy (VE) is used in vaccine studies. As discussed in the Halloran et al. (1997), VE may be constructed from other parameters that are related in some way to the probability of disease transmissions.

According to the Centers for Disease Control and Prevention (CDC), PF measures the proportionate reduction in cases (disease occurrences) among the vaccinated group (CDC, 2012).

Let us consider some examples based on prevented fraction (PF):

Ex 1: Pfizer-BNT162b2 mRNA Covid-19 vaccine

Pollack et al. (2020) discussed a clinical trial for the safety and efficacy of the BNT162b2 mRNA Covid-19 vaccine. In a clinical trial, a total of 43,548 participants underwent randomization, of whom 43,448 received injections: 21,720 with BNT162b2 and 21,728 with placebo. There were 8 cases of Covid-19 with onset at least 7 days after the second dose among participants assigned to receive BNT162b2 and 162 cases among those assigned to placebo. Letting n_1 be the number of people in the vaccinated group and n_2 be the number of people in the control group, then $n_1 = 21{,}720$ and $n_2 = 21{,}728$.

Similarly, letting x_1 be the number of people infected with the virus in the vaccinated group and x_2 be the number of people infected with the virus in the control group, we have $x_1 = 8$ and $x_2 = 162$.

Following the convention about relative risk (RR) discussed in Section 1.1, let A be the vaccinated group and $\bar{A}$ be the unvaccinated group where B is defined to be the event the subject develops Covid-19 7 or more days after the second dose. The corresponding estimated risks are then defined to be

$$P(B|A) = x_1 \,/\, n_1 = 8/21{,}720 \quad = 0.000368$$

$$P(B|\bar{A}) = x_2 \,/\, n_2 = 162/21{,}728 = 0.007456$$

$$PF = 1 - \frac{P(B|A)}{P(B|\bar{A})} = \frac{0.000368}{0.007456} = 0.9506438.$$

PF of 95.1% indicates a 95.1% reduction in disease occurrence among the vaccinated group.

So, the vaccinated group experienced 95.1% fewer disease cases than they would have if they had not been vaccinated.

Ex 2: Outbreak of varicella (chickenpox) in Oregon in 2002

Tugwell et al. (2004) investigated a chickenpox outbreak that started in an Oregon elementary school in October 2001, after public schools began phasing in a varicella vaccination requirement for enrollment. In that chickenpox outbreak study, varicella was diagnosed in 18 of 152 vaccinated children compared with 3 of 7 unvaccinated children.

Here $n_1 = 152$, $n_2 = 7$, $x_1 = 18$, and $x_2 = 3$

Using the above formula, PF will be equal 0.72. So, the vaccinated group experienced 72% fewer varicella cases than they would have if they had not been vaccinated.

1.3 Mitigated Fraction

Prevented fraction (PF) is used in studies where the response observed on subjects is a dichotomous outcome, e.g., clinical signs of disease are observed in a subject as presence or absence. Recent studies indicate a growing interest in the vaccine's ability to reduce the severity of the targeted disease rather than preventing it entirely. The concept of the mitigated fraction was introduced in veterinary medicine to quantify the reduction in the severity of disease occurring in vaccinated animals as compared to non-vaccinated animals. The USDA's Center for Veterinary Biologics (CVB) recommends a form of the mitigated fraction proposed by Siev (2005) which can be calculated when the disease severity is graded by some continuous measure or by some discrete assessment resulting in unambiguous ranks. Siev (2005) outlined and structured the formulation of the mitigated fraction in a non-parametric approach based on Wilcoxon rank sum statistics. A motivated example was given in Siev (2005) describing a study to determine the efficacy of a vaccine in terms of finding the mitigated fraction for swine respiratory disease with two treatment groups, one group of pigs treated with the vaccine and other group of pigs treated with a placebo (i.e. the control group).

1.4 The motivated example for mitigated fraction from Siev (2005)

Siev (2005) describes a study to determine the efficacy of a vaccine for swine respiratory disease with two treatment groups, pigs treated with the vaccine and pigs treated with a placebo (i.e. the control group). In this study all the pigs were exposed to the pathogen and sacrificed. A postmortem examination of each pig's lungs was conducted to measure the extent of gross lesions. Two observers independently sketched on a grid of the dorsal and ventral surfaces of each of the seven lung lobes. The fraction of each lobe was taken as the average of the two surfaces over the two observers. The lobe fractions were then weighted and summed to arrive at the fraction of the lungs affected by gross lesions. This response variable is clearly continuous. Siev (2005) pointed out a possible analysis would be to convert the fraction of gross lung lesions to a dichotomous response, say unaffected (i.e. 0% lesions) and affected (greater than 0%). This variable could then be analyzed in terms of the prevented fraction (PF). As Siev (2005) mentioned, important information is lost when considering only presence/absence since it ignores the severity of the affected individuals. However, Siev (2005) mentioned that if disease severity can be graded by some continuous measure or perhaps a discrete assessment that results in unambiguous ranks (i.e. no ties), then the mitigated fraction could be estimated in a fully non-parametric approach.

1.5 Mathematical formulation for mitigated Fraction

The mitigated fraction (MF) is defined as the relative increase in the probability that a vaccinate's disease will be less severe than a non-vaccinate's disease (Siev, 2005). Let Y represent a response variable measured on an individual denoting the severity of its disease. Without a loss of generality, assume that larger values of Y indicate greater severity. Now let Y_1 be the observed value of this variable for an individual randomly selected from the control (placebo) group and let Y_2 be the response for a randomly selected individual from the vaccine group.

Mathematically, that mitigated fraction can be expressed as:

$$MF = P(Y_1 > Y_2) - P(Y_2 > Y_1).$$

The mitigated fraction can be reformulated in the following way. A measure that summarizes the relative effect size of Y_1 and Y_2 is

$T(Y_1, Y_2) = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ (Kruskal, 1957).

Vargha and Dalancy (1998) called this the measure of stochastic superiority of variable Y_1 over variable Y_2. If $T(Y_1, Y_2) = 0.5$ then we can say that Y_1 and Y_2 are stochastically equal or identically distributed. Further we can note from Vargha et al. (1998) that when $T(Y_1, Y_2) > 0.5$ (or < 0.05), outcomes of Y_1 tend to be larger (or smaller) than outcomes of Y_2. For simplicity, from now we will drop (Y_1, Y_2) from $T(Y_1, Y_2)$ and let $T = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ as the measure of relative effect size.

In terms of T, *MF* can be expressed as,

$$MF = 2T - 1.$$

The measure T could be estimated in terms of Wilcoxon rank sum statistics as discussed in Siev (2005).

1.6 Wilcoxon rank sum statistics and the mitigated fraction

Wolf and Hogg (1971) showed that $P(Y_2 < Y_1)$ can be estimated using Wilcoxon rank sum statistics as shown below.

$P(Y_2 < Y_1) = \int F(z)dG(z)$, where F and G are the distribution function of Y_2 and Y_1 respectively. Since F and G are unknown, Wolf et al. (1971) showed that the value of $P(Y_2 < Y_1)$ can approximated estimated by using empirical distribution functions as estimates for F and G.

$P(Y_2 < Y_1) \approx \int F_{Y_2}(z)dG_{Y_1}(z) = \frac{W_1 - N_1(N_1+1)/2}{N_2N_1}$, where F_{Y_2} and G_{Y_1} are empirical distribution functions as estimates for F and G respectively, W_1 is sum of the ranks in control group (Wilcoxon rank sum statistics), $N_1$and N_2 are the number of observations from the control and vaccine groups respectively.

Because Y_1 and Y_2 are continuous random variables, $P(Y_1 = Y_2) = 0$ and $T = P(Y_1 > Y_2)$ and $MF = 2T - 1$.

Now, estimated MF will be equal to $\left\{\frac{2W_1 - N_1(1+N_2+N_1)}{N_2N_1}\right\}$. This provides a fully non-parametric approach for MF.

1.7 Ridit analysis and MF

Ridit was introduced by Bross (1958) and the term ridit is derived from the initials of "**R**elative to an **I**dentified **D**istribution" and it is analogous to the probit and logit models. Ridit assumes an underlying empirical continuous distribution whereas the probit and logit assume underlying standard normal and logistic distributions, respectively. Ridit is a non-parametric approach for ordinal data whereas the probit and logit are parametric approaches. Ridit analysis plays a main role in the estimation of mitigated fraction for an ordinal response. In fact, the mean ridit determines the estimated value for the mitigated fraction from an ordinal response. For the analysis of ridit, it is important to have a population group to treat as a reference group. Bross (1958) mentioned this population group as the 'identified distribution'. This can be thought as a control (or placebo) group in the usual analysis of treatment versus control. A ridit score is defined as the proportion of all individuals from the reference group falling in the lower ranking categories plus half the proportion falling in the given category (Fleiss, 2003). Fleiss (2003) states that the ridit analysis assumes there are discrete categories representing intervals of an underlying but unobservable continuous distribution. The operations can be viewed as a method of assigning a score (or weight) to the graded categories.

The mean ridit for a given group with same categories of the reference group can be calculated and viewed as a probability. Fleiss (2003) states, "The mean ridit for a group is the probability that a randomly selected individual from it has a value indicating greater severity or seriousness than a randomly selected individual from the standard group." Similarly, Bross (1958) says the mean ridit is an estimate of the chance that an individual in a given class is "worse off" than an individual in the reference class.

Mathematical form of ridit analysis

An alternative approach to selecting scores for ordinal categories used in a ridit analysis is to use the data themselves to determine the scores. One such scoring method uses the average cumulative proportions for the ordinal response variable (Agresti, 2010).

For illustration purposes, let Y_1 be the observed value of this variable for an individual randomly selected from the control (Placebo) group and let Y_2 be the response for a randomly selected individual from the treatment (Vaccine) group. Let J be the number of ordinal categories in both vaccine and control groups.

Let $p_1, p_2, \ldots, p_J$ be the proportions for the control group categories and let $q_1, q_2, \ldots, q_J$ be the proportions for vaccine group categories. For the reference group, the ridit scores for category j is given by

$$a_j = \sum_{k=1}^{j-1} p_k + \frac{1}{2}\, p_j \quad for\ j = 1,2,\ldots,J.$$

Let $T(Y_2, Y_1)$ be the mean ridit of the treatment group with respect to the control group which is defined as

$$T(Y_2, Y_1) = \sum_{j=1}^{J} q_j a_j.$$

Following Agresti (2010), it can be shown that

$$T(Y_2, Y_1) = P(Y_2 > Y_1) + \frac{1}{2} P(Y_1 = Y_2).$$

Using the above results, *MF* can be expressed in terms of $T(Y_2, Y_1)$ as,

$$MF = 1 - 2 * T(Y_2, Y_1).$$

Following Agresti (2010), the mean ridit of the control group with respect itself, i.e. $T(Y_1, Y_1)$, is given by

$$T(Y_1, Y_1) = \sum_{j=1}^{J} p_j a_j = \sum_{j=1}^{J} p_j \left(\sum_{k=1}^{j-1} p_k + \frac{1}{2}\, p_j \right)$$

$$= 2 \sum \sum\nolimits_{k<j} p_j p_k + \sum_{j} p_j^2.$$

Therefore,

$$T(Y_1, Y_1) = \frac{\left(\sum p_j\right)^2}{2} = 0.5$$

The above result clearly shows the reference group compared to itself has the mean ridit of 0.50 which is intuitively what it should be. If the mean ridit of the treatment group with respect to the control group is 0.5, then we can say both groups are similar in terms of the severity.

1.8 Mitigated fraction from an ordinal response

Now let us assume there is an interest in a measure of disease severity where the outcome variable is ordinal, possibly when observations are clustered or measured longitudinally. Ordinal outcomes in terms of measuring disease severity are widely popular in the areas of veterinary medicines and epidemiological fields. For example, disease severity might be measured as 'none', 'light', 'mild', 'moderate', or 'severe'. Also, it is common in many clinical trials having outcomes which are measured on an ordered categorical scale.

Currently, the MF is estimated for continuous responses. Siev (2005) implied that MF can be calculated when the disease severity is graded by some continuous measure or by some discrete assessment resulting in unambiguous ranks. Further, the MF R-package (Siev, 2012) supports only the continuous response variable to calculate MF. Siev (2005) briefly introduced the interest in estimating MF for mitigated fraction for ordinal data and he referred the idea of using normal latent continuous random variable discussed in Poon (2004) and mean ridit (Bross, 1958) to estimate MF for ordinal data. Latent variable approaches will be discussed in the subsequent chapters to estimate MF for ordinal data. Even though Siev (2005) very briefly introduced the interest in estimating MF for mitigated fraction for ordinal data, he never discussed any methods explicitly to estimate MF for ordinal data. In this work, several methods are discussed to estimate MF for an ordinal response. A parametric approach assuming a generalized linear mixed model (GLMM) with latent variable is developed for estimating MF for ordinal response. To accomplish this, the following sections are outlined as follows: Chapter 2 discusses the theatrical framework, Chapter 3 discusses estimating MF when the vaccine and placebo groups are in a completely randomized design, Chapter 4 discusses estimating MF for ordinal clustered data, Chapter 5 briefly discusses estimating MF for ordinal data from a Bayesian perspective, and Chapter 6 discusses conclusion, extension and future work.

Chapter 2 - Theoretical framework

Multinomial distribution analysis with ordinal data for mitigated fraction is preferable for the response variables such as disease severity rating (with possible outcomes none, light, mild, moderate, or severe) or approval rating (with possible outcomes strongly disagree, disagree, neutral, agree, or strongly agree).

A multinomial response Y can be considered for ordered categorical outcomes, which could possibly be denoted by $1, 2, 3, \dots, J$ with probabilities $\tilde{\pi}_j = P(Y = j)$ for $j = 1, \dots, J$. Each observation of Y belongs to exactly one category, i.e. Y takes a value j if a particular ordinal observation falls in the j'th category and the probabilities sum to one, i.e. $\sum_{j=1}^{J} \tilde{\pi}_j = 1$.

When n_j is the number of observations from a given sample belonging to category j with $\sum_{j=1}^{J} n_j = n$, the probability mass function for the multinomial distribution is

$$f(n_1, n_2, \dots, n_J) = \frac{n!}{n_1!\ n_2,! \dots n_J!} (\tilde{\pi}_1)^{n_1}\ (\tilde{\pi}_2)^{n_2} \dots (\tilde{\pi}_J)^{n_J}.$$

Cumulative link models

Let Y denote the response variable for an ordinal observation, and let $\boldsymbol{X} = (x_1, x_2, \dots, x_p)'$ denote a p-dimensional vector of explanatory variables.

One can model the relationship between the ordinal random variable Y and its corresponding vector of explanatory variables using the cumulative probability model.

A basic cumulative link model is

$$\pi_j = G(\eta_j), \eta_j = \gamma_j - \boldsymbol{X}'\boldsymbol{\beta} \text{ for } j = 1, \dots, J - 1,$$

where $\pi_j = P(Y \leq j|\boldsymbol{X}) = \tilde{\pi}_1 + \cdots + \tilde{\pi}_j$ with $\sum_{j=1}^{J} \tilde{\pi}_j = 1$ are cumulative probabilities; η_j is the linear predictor; $\boldsymbol{\beta} = \left(\beta_1, \beta_2, \dots, \beta_p\right)'$ is a p-dimensional vector of parameters that describe the effects of the explanatory variables; and $\gamma_j{}'s, j = 1,2, \dots, J$ are thresholds (also known as cut-points or intercepts) and they are representing the baseline value of the transformed cumulative probability for category j.

The above cumulative link model written as

$$G^{-1}[P(Y \leq j|\boldsymbol{X})] = \gamma_j - \boldsymbol{X}'\boldsymbol{\beta} = \gamma_j - \beta_1 x_1 - \cdots - \beta_p x_p \tag{1.1}$$

for $j = 1, \dots, J - 1$. Here, $G^{-1}(\cdot)$ is an arbitrary link function which links the cumulative probabilities to the linear predictor. The $G^{-1}(\cdot)$ is a (known) monotone-increasing function mapping the interval $(0, 1)$ onto the real line $(-\infty, \infty)$. Here, model (1.1) assumes an identical effect $\boldsymbol{\beta}$ of the predictors for each cumulative probability.

If G^{-1} is the inverse of the standard logistic cumulative density function (cdf), then model (1.1) is a cumulative logit model, while if G^{-1} is the inverse of the standard extreme value (minimum) cdf, then model (1.1) is called the cumulative complementary log-log (clog-log) or proportional hazards model. Different types of cumulative probability models for ordinal data will briefly be discussed in the subsequent sections.

The cell probabilities, $\tilde{\pi}_1, \dots, \tilde{\pi}_j$, for the cumulative link modes can be driven as follows

$$P(Y = j|\boldsymbol{X}) = \tilde{\pi}_j = G\left(\gamma_j - \boldsymbol{X}'\boldsymbol{\beta}\right) - G\left(\gamma_{j-1} - \boldsymbol{X}'\boldsymbol{\beta}\right) \text{ for } j = 1, \dots, J$$

with $\gamma_0 = -\infty$ and $\gamma_J = \infty$.

2.1 The cumulative logit model

When the G^{-1} in model (1.1) is the inverse of the standard logistic cdf, then the model (1.1) is a cumulative logit or proportional odds model given as

$$logit[P(Y \leq j|\boldsymbol{X})] = log \frac{P(Y \leq j|\boldsymbol{X})}{1 - P(Y \leq j|\boldsymbol{X})} = \gamma_j - \boldsymbol{X'\beta}, for \ \ j = 1, \dots, J - 1$$

and $P(Y \leq j|\boldsymbol{X})$ can be obtained as,

$$P(Y \leq j|\boldsymbol{X}) = \frac{exp\,(\gamma_j - \boldsymbol{X'\beta})}{1 + exp\,(\gamma_j - \boldsymbol{X'\beta})} \ for \ \ j = 1, \dots, J - 1.$$

and the cell probabilities under logit model are

$$P(Y = j|\boldsymbol{X}) = \tilde{\pi}_j = \frac{exp\,(\gamma_j - \boldsymbol{X'\beta})}{1 + exp\,(\gamma_j - \boldsymbol{X'\beta})} - \frac{exp\,(\gamma_{j-1} - \boldsymbol{X'\beta})}{1 + exp\,(\gamma_{j-1} - \boldsymbol{X'\beta})} \text{ for } j = 1, \dots, J$$

with $\gamma_0 = -\infty$ and $\gamma_J = \infty$.

2.1.1 The proportional odds property

The odds ratio (OR) of observing event $Y \leq j$ at $\boldsymbol{X} = \boldsymbol{X_1}$ relative to the same event at $\boldsymbol{X} = \boldsymbol{X_2}$ is

$$OR = \frac{P(Y \leq j|\boldsymbol{X_1})/P(Y > j|\boldsymbol{X_1})}{P(Y \leq j|\boldsymbol{X_2})/P(Y > j|\boldsymbol{X_2})} = exp((\boldsymbol{X_2} - \boldsymbol{X_1})'\boldsymbol{\beta})$$

The cumulative odds ratio is proportional to the difference between $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$. The same proportionality constant applies to all $J - 1$ logits. Because of this property, following McCullagh (1980), the cumulative logit model is often referred to as a proportional odds model.

2.2 The cumulative probit model

When the G^{-1} in model (1.1) is the inverse of the standard normal cdf, then the model (1.1) is a cumulative probit model given as

$$\Phi^{-1}[P(Y \leq j|\boldsymbol{X})] = \gamma_j - \boldsymbol{X'\beta} = \gamma_j - \beta_1 x_1 - \cdots - \beta_p x_p \text{ for } j = 1, \dots, J - 1$$

and hence

$$[P(Y \leq j|\boldsymbol{X})] = \Phi(\gamma_j - \boldsymbol{X'\beta}) \text{ for } j = 1, \dots, J - 1,$$

where Φ denotes the cumulative density function of the standard normal distribution. The above cumulative probit model links the cumulative probabilities to a linear predictor.

2.3 The cumulative complementary log-log model and cumulative log-log

The Gumbel distribution is sometimes called the type 1 extreme value distribution and has two forms. One is used to model the smallest extreme (minimum) values of various distributed outcomes and the other is used to model the largest extreme (maximum) values of various distributed outcomes. They are called Gumbel distribution (minimum) and Gumbel distribution (maximum) respectively (Wikipedia, 2021 and Agresti, 2010).

This project report uses the term *Gumbel distribution (maximum)* for Gumbel distribution to model the distribution of the maximum values, and the name *Gumbel distribution (minimum)* for Gumbel distribution to model the distribution of the minimum value.

2.3.1 The Cumulative complementary log-log models (clog-log)

The Gumbel distribution (minimum) has the cumulative distribution function, $F(x;\mu,\sigma) = 1 - exp\ (-exp\ ((x-\mu)/\sigma))$ with location parameter μ and scale parameter σ. The standard Gumbel distribution (minimum) with location parameter $\mu = 0$ and scale parameter $\sigma = 1$ is given as $F(x) = 1 - exp\ (-exp\ (x))$.

When the G^{-1} in model (1) is the inverse of the standard Gumbel (minimum) cdf then the model (1) is a cumulative complimentary log-log (clog-log) model given as

$$log\ (-log\ (1 - P(Y \leq |\boldsymbol{X})) = \gamma_j - \boldsymbol{X}'\boldsymbol{\beta} = \gamma_j - \beta_1 x_1 - \cdots - \beta_p x_p \text{ for } j = 1, \dots, J-1$$

and hence

$$P(Y \leq j|\boldsymbol{X}) = 1 - exp\big(-exp\big((\gamma_j - \boldsymbol{X}'\boldsymbol{\beta})\big)\big) \text{ or } j = 1, \dots, J-1.$$

Here it is to be noted that the above model applies the $log\ (-log)$ as the link function for the complementary of cumulative probabilities and thus the link function is called the complimentary log-log (clog-log) link function (Agresti, 2010).

2.3.2 Cumulative log-log models

The Gumbel distribution (maximum) has the cumulative distribution function, $F(x; \mu, \sigma) = exp\ (-exp\ (-(x-\mu)/\sigma))$ with location parameter μ and scale parameter σ. The standard Gumbel distribution (maximum) with location parameter $\mu = 0$ and scale parameter $\sigma = 1$ is given as $F(x) = exp\ (-exp\ (-x))$.

When the G^{-1} in model (1) is the inverse of the standard Gumbel (maximum) cdf, then the model (1) is a cumulative log-log model given as

$$-log\ (-log\ (P(Y \leq j|\boldsymbol{X})) = \gamma_j - \boldsymbol{X'\beta} = \gamma_j - \beta_1 x_1 - \cdots - \beta_p x_p \text{ for } j = 1, \ldots, J-1$$

and hence

$$P(Y \leq j|\boldsymbol{X}) = exp\big(-exp\big(-(\gamma_j - \boldsymbol{X'\beta})\big)\big) \text{ for } j = 1, \ldots, J-1.$$

Here it is to be noted that the above model applies the $-log\ (-log)$ as the link function for the cumulative probabilities and thus the link function for the cumulative probabilities is called the log-log link function.

2.4 Latent variable motivation

Following McKelvey and Zavoina (1975), McCullagh (1980), Anderson and Philips (1981), Cox (1995), Johnson and Albert (1999), Agresti (2010), and many other authors, one can view an ordinal response variable Y is that it is generated by an unobserved continuous latent variable Y^* with $J-1$ cut points $\gamma_1 < \gamma_2 < \cdots < \gamma_{J-2} < \gamma_{J-1}$, on the continuous latent scale such that $Y = j$ if $\gamma_{j-1} < Y^* \leq \gamma_j$. Agresti (2010) notes that it may be sensible to consider an ordinal variable as a necessarily crude measurement of the continuous latent variable Y^*, which would be the response variable in an ordinary linear model. The cumulative probability model (1) can be implied by a model in which the continuous latent variable Y^* satisfies an ordinary regression model $Y^* = \boldsymbol{X'\beta} + \epsilon$ in which ϵ has a cumulative distribution function (cdf) G with constant variance (Anderson et al., 1981).

Thus, the latent variable Y^* and the observed response variable Y give the model with the probability that the response variable Y will fall in the j^{th} category or below, given $\boldsymbol{X}$,

$$P(Y \leq j|\boldsymbol{X}) = P(Y^* \leq \gamma_j|\boldsymbol{X}) = P(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j) = P(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}) = G(\gamma_j - \boldsymbol{X'\beta}).$$

Thus, the cumulative probability model (1) can be given in terms of latent variable as

$$P(Y^* \leq \gamma_j|\boldsymbol{X}) = P(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j) = P(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}) = G(\gamma_j - \boldsymbol{X'\beta}) \qquad (1.2).$$

Note that in both models (1) and (2), we assume that the predictors $\boldsymbol{X}$ do not include a column of ones because the constant is absorbed into the cut points. But in some situations, like in the Bayesian approach in Chapter 5, we assume that the predictors $\boldsymbol{X}$ include a column of ones by imposing appropriate constraints on the cut points. Detailed identification constraints will be discussed in the next section.

As mentioned above, the link function to apply for $P(Y \leq j|\boldsymbol{X})$ to obtain a linear predictor is G^{-1}, the inverse of the cumulative distribution function for the latent variable Y^*. That is,

$G^{-1}(P(Y \leq j|\boldsymbol{X})) = \gamma_j - \boldsymbol{X'\beta}$ (Agresti, 2010).

When ϵ follows an independent normal distribution with mean 0 and constant variance σ^2, $N(0, \sigma^2)$ then,

$$\begin{aligned} P(Y^* \leq \gamma_j|\boldsymbol{X}) &= P(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j) \\ &= P(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}) \end{aligned}$$

$= P\big(Z \le (\gamma_j - \boldsymbol{X'\beta})/\boldsymbol{\sigma}\big),$ hence

$$P\big(Y^* \le \gamma_j | \boldsymbol{X}\big) = \Phi\big((\gamma_j - \boldsymbol{X'\beta})/\sigma\big) \qquad (1.3)$$

where Z follows a standard normal distribution and Φ denotes the standard normal cumulative distribution function. The γ_j's, $\boldsymbol{\beta}$ and σ can be identifiable if Y^* is directly observed but since Y^* is unobservable, not all parameters are identifiable. For identifiable easiness, without loss of generality, σ can be set to 1. Jackman (2000) discussed possible identification constraints which are displayed in Table 2.1.

With $\sigma = 1$ for the expression in the equation (1.3),

$$P\big(Y^* \le \gamma_j | \boldsymbol{X}\big) = \Phi\big((\gamma_j - \boldsymbol{X'\beta}\big).$$

Now it is straight forward that $P(Y \le j|\boldsymbol{X}) = P\big(Y^* \le \gamma_j|\boldsymbol{X}\big) = \Phi\big((\gamma_j - \boldsymbol{X'\beta}\big)$ and this shows a cumulative probit model arise when ϵ follows a standard normal distribution. We will discuss different types of cumulative probability models under the latent variable structure in the subsequent sections.

Table 2.1. Identification Constraints

Constrains options	β	σ	α
1	unconstrained	fixed e.g., $\sigma = 1$	one γ_j fixed e.g., $\gamma_1 = 0$
2	drop intercept, $\beta_0 = 0$	fixed e.g., $\sigma = 1$	unconstrained
3	unconstrained	unconstrained	two γ_j's fixed e.g., $\gamma_1,\ \gamma_2 = 0$

Source: Jackman, S. (2000). Models for Ordered Outcomes**.** *Political Science* 200C

2.4.1 Probit link model motivation

When ϵ follows a standard normal distribution, $\epsilon \sim N(0,1)$, a normal continuous latent variable is given as $Y^* \sim N(\boldsymbol{X'\beta}, 1)$, and under this normal latent variable:

$$P(Y \leq j|\boldsymbol{X}) = P\big(Y^* \leq \gamma_j|\boldsymbol{X}\big) = P\big(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j\big)$$

$$= P\big(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}\big)$$

$$= P\big(Z \leq (\gamma_j - \boldsymbol{X'\beta})\ \big)$$

$$P(Y \leq j|\boldsymbol{X}) = P\big(Y^* \leq \gamma_j|\boldsymbol{X}\big)\ =\ \Phi\big((\gamma_j - \boldsymbol{X'\beta})\big).$$

where Z follows a standard normal distribution and Φ denotes the standard normal cdf. Thus, a probit model arises under the normal latent variable.

2.4.2 Logit link model motivation

When ϵ follows independent standard logistic distribution, $logistic(0, 1)$ with location parameter $\mu = 0$ and scale parameter $\sigma = 1$,

$$P\big(Y^* \leq \gamma_j|\boldsymbol{X}\big) = P\big(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j\big)$$

$$= P\big(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}\ \big)$$

$$P\big(Y^* \leq \gamma_j|\boldsymbol{X}\big) = G(\gamma_j - \boldsymbol{X'\beta}\ )$$

$$P(Y \leq j|\boldsymbol{X}) = P\big(Y^* \leq \gamma_j|\boldsymbol{X}\big) = \frac{exp\ \big((\gamma_j - \boldsymbol{X'\beta}\ \big)}{1 + exp\ ((\gamma_j - \boldsymbol{X'\beta}\ )}$$

where G denotes the standard logistic cdf. Thus, a logit model arises under the logistic latent variable.

2.4.3 Complimentary log-log model link model motivation

When ϵ follows independent standard Gumbel (minimum) distribution, $Gumbel_{min}(0,1)$ with location parameter $\mu = 0$ and scale parameter $\sigma = 1$,

$$P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = P\left(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j\right)$$

$$= P\left(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}\right)$$

$$P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = G\left(\gamma_j - \boldsymbol{X'\beta}\right)$$

$$P(Y \leq j | \boldsymbol{X}) = P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = 1 - exp\left(-\, exp\left((\gamma_j - \boldsymbol{X'\beta})\right)\right)$$

where G denotes the standard Gumbel (minimum) cdf. Thus, a complimentary log-log model arises under the Gumbel (minimum) latent variable.

2.4.4 log-log model link model motivation

When ϵ follows independent standard Gumbel (maximum) distribution, $Gumbel_{max}(0,1)$ with location parameter $\mu = 0$ and scale parameter $\sigma = 1$,

$$P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = P\left(\boldsymbol{X'\beta} + \epsilon \leq \gamma_j\right)$$

$$= P\left(\epsilon \leq \gamma_j - \boldsymbol{X'\beta}\right)$$

$$P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = G\left(\gamma_j - \boldsymbol{X'\beta}\right)$$

$$P(Y \leq j | \boldsymbol{X}) = P\left(Y^* \leq \gamma_j | \boldsymbol{X}\right) = exp\left(-\, exp\left(-(\gamma_j - \boldsymbol{X'\beta})\right)\right)$$

where G denotes the standard Gumbel (maximum) cdf. Thus, a log-log model arises under the Gumbel (maximum) latent variable.

In summary, the latent variable structure gives rise to probit, logit, log-log, and clog-log models when ϵ respectively follows distributions of a standard normal, standard logistic, standard

Gumbel (maximum), and standard Gumbel (minimum) cdf's. Table 2.2 shows the distributions of ϵ and the form of link functions $G^{-1}(\cdot)$.

Table 2.2. Distributions of $\boldsymbol{\epsilon}$ and form of $\boldsymbol{G^{-1}}(\cdot)$

Name/Model	Distribution of ϵ	Link function (G^{-1})	Inverse link (G)
probit	$N(0,\ 1)$	$\Phi^{-1}(\gamma)$	$\Phi(\eta)$
logit	$logistc(0,\ 1)$	$logit(\gamma) = log\ [\gamma/(1-\gamma)]$	$exp(\eta)\,/(1+exp(\eta))$
clog-log	$Gumbel_{min}(0,1)$	$log[-log\ (1-\gamma)]$	$1 - exp\ (-exp\ (\eta))$
log-log	$Gumbel_{max}(0,1)$	$-log[-log\ (\gamma)]$	$exp\ (-exp\ (-\eta))$

In this table, γ denotes cumulative probabilities and η denotes linear predictor. $Gumbel_{min}$ and $Gumbel_{max}$ denote, respectively, Gumbel (minimum) distribution and Gumbel (maximum) distribution.

2.5 Maximum likelihood (ML) estimation for cumulative link models parameters

Following Agresti (2010), for subject i, let $y_{ij} = 1$ if $y_i = j$ and let $y_{ij} = 0$ otherwise, $i = 1,2,\dots,n$. Then $E(Y_{ij}) = \tilde{\pi}_j(\boldsymbol{X}_i)$, the probability that observation i with explanatory variable values $\boldsymbol{X_i}$ falls in category j. With the independent observations, we obtain the likelihood function by substituting $G(\gamma_j - \boldsymbol{X_i'\beta})$ for $P(Y_i \le j|\boldsymbol{X}_i)$ in the product of multinomial probability mass functions,

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\tilde{\pi}_j(\boldsymbol{X}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}[P(Y_i \le j|\boldsymbol{X}_i) - P(Y_i \le j-1|\boldsymbol{X}_i)]^{y_{ij}}\right\},$$

$$\prod_{i=1}^{n}\left[\prod_{j=1}^{J}\tilde{\pi}_j(\boldsymbol{X}_i)^{y_{ij}}\right] = \prod_{i=1}^{n}\left\{\prod_{j=1}^{J}\left[G(\gamma_j - \boldsymbol{X_i'\beta}) - G(\gamma_{j-1} - \boldsymbol{X_i'\beta})\right]^{y_{ij}}\right\}.$$

The log-likelihood function is

$$L(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} log\left[G(\gamma_j - \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\beta}) - G(\gamma_{j-1} - \boldsymbol{X}_{\boldsymbol{i}}'\boldsymbol{\beta})\right].$$

We can see that the log-likelihood function is function of γ_j's and $\boldsymbol{\beta}$ after observing the observations y_{ij}. As discussed in Agresti (2010), each log-likelihood equation can be obtained by differentiating $L(\boldsymbol{\gamma}, \boldsymbol{\beta})$ with respect to a particular parameter and equating the derivative to zero. And then iterative methods are used to solve the log-likelihood equations and then obtain the ML estimates of the model parameters. Generally, the Newton-Rapson algorithm can be used to maximize the log-likelihood function to yield ML estimates.

2.6 Bootstrap methods and bootstrap confidence intervals

Bootstrap methods can be used to construct confidence intervals. In this section, most of terms, and notation follow closely to those found in Efron and Tibshirani (1993).

Let $\boldsymbol{X} = (x_1, x_2, \dots, x_n)$ be a random sample from an unknown probability distribution F and we wish to estimate a parameter of interest $\theta = t(F)$ based on $\boldsymbol{X}$. Let $\hat{\theta} = s(\boldsymbol{X})$ be a statistic calculated from $\boldsymbol{X}$ such that $\hat{\theta}$ estimate the parameter of interest θ. The bootstrap method can be used to estimate the standard error of $\hat{\theta}$. Bootstrap methods depend on the bootstrap sample. Let $\hat{F}$ be the empirical distribution, having probability $1/n$ on each of the observed values $x_i, i = 1,2, \dots, n$. A bootstrap sample $\boldsymbol{X}^*$ can be defined as $\boldsymbol{X}^* = (x_1^*, x_2^*, \dots, x_n^*)$ where $\boldsymbol{X}^*$ is a random sample of size n drawn from an empirical distribution $\hat{F}$ of the sample $\boldsymbol{X}$**.** Here, the $\boldsymbol{X}^*$ will be treated as a resampled version of $\boldsymbol{X}$**.** This means that the data points $(x_1^*, x_2^*, \dots, x_n^*)$ are a random sample of size n drawn with replacement from the population of n data points $(x_1, x_2, \dots, x_n)$. Corresponding to a bootstrap data set $\boldsymbol{X}^*$**,** $\hat{\theta}^*$will be a bootstrap replication of $\hat{\theta}$ such that

$$\hat{\theta}^* = s(\boldsymbol{X}^*).$$

The quantity $s(\boldsymbol{X}^*)$ can be calculated by applying the same function $s(\cdot)$ as was applied to $\boldsymbol{X}$**.** In the need of a given statistical properties, such as standard error of the estimate $\hat{\theta}$, it is necessary to have the sampling distribution of $\hat{\theta}$.

Let $se_F\hat{\theta}$ be the bootstrap estimate and it is the standard error of the statistics $\hat{\theta}$. The $se_F\hat{\theta}$ is a plug-in estimate that uses the empirical distribution $\hat{F}$ in place of the unknown distribution F. That is, the bootstrap estimate of $se_F\hat{\theta}$ can be defined by $se_F\hat{\theta}^*$(Efron et al.,1993).

As Efron et al. (1993) stated, the bootstrap algorithm works by drawing many independent bootstrap samples, evaluating the corresponding bootstrap replications, and estimating the standard error of $\hat{\theta}$ by the empirical standard deviation of the replications. The result is called the bootstrap estimate of standard error, denoted by $\widehat{se}_B$, with B is the number of bootstrap samples. The following gives the detailed description of the bootstrap procedure for estimating the standard error of $\hat{\theta} = s(\boldsymbol{X})$ from the observed data $\boldsymbol{X}$. The following algorithm has been directly taken from Efron et al. (1993).

The bootstrap algorithm procedure for estimating the standard errors:

1. First, select B independent number of bootstrap samples $\boldsymbol{X}^{*1}, \boldsymbol{X}^{*2}, \boldsymbol{X}^{*3}, \ldots \boldsymbol{X}^{*B}$, each consisting of n data values drawn with replacement from $\boldsymbol{X}$.

2. Evaluate the bootstrap replication, $\hat{\theta}^*$, corresponding to each bootstrap sample,

$$\hat{\theta}^*(b) = s(\boldsymbol{X}^{*b}) \qquad\qquad b = 1,2,\ldots,B.$$

3. Now, estimate the standard error $se_F(\hat{\theta})$ by finding the sample standard deviation of the B replications

$$\widehat{se}_B = \left\{\sum_{b=1}^{B}\left[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)\right]^2/(B-1)\right\}^{1/2}$$

where $\hat{\theta}^*(\cdot) = \sum_{b=1}^{B} \hat{\theta}^*(b)/B$. Using the generated bootstrap replications $(\hat{\theta}^{*^1}, \dots, \hat{\theta}^{*^B})$, confidence intervals for θ can be constructed.

Percentile interval using bootstrap samples

Let $\hat{\theta}^{*(\alpha)}$ indicate the $100 * \alpha$th percentile of B bootstrap replications $\hat{\theta}^*(1), \hat{\theta}^*(2), \dots, \hat{\theta}^*(B)$. The percentile interval $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ of intended coverage $1 - 2\alpha$, can be obtained directly from these percentiles, with $\hat{\theta}_{lo}$ and $\hat{\theta}_{up}$ are lower control and upper control confidence limits, respectively.

The percentile interval $(\hat{\theta}_{lo}, \hat{\theta}_{up})$ will be equal to $(\hat{\theta}^{*(\alpha)},\ \hat{\theta}^{*(1-\alpha)})$, i.e. $(\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha)},\ \hat{\theta}^{*(1-\alpha)})$.

An example from Efron et al. (1993), if $B = 2000$ and $\alpha = .05$, then the percentile interval $(\hat{\theta}^{*(.05)},\ \hat{\theta}^{*(.95)})$ will be given by the 100th (i.e. $\hat{\theta}^*(100)$)to the 1900th (i.e. $\hat{\theta}^*(1900)$) ordered values of the 2000 numbers $\hat{\theta}^*(b)$, $b = I, 2, \dots, 2000$.

Bias Corrected and Accelerated Confidence Intervals (BC_a)

Following Efron et al. (1993), the BC_a interval of intended coverage, $1 - 2\alpha$, will be given by

BC_a: $(\hat{\theta}_{lo}, \hat{\theta}_{up}) = (\hat{\theta}^{*(\alpha_1)}, \hat{\theta}^{*(\alpha_2)})$,

where

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(\alpha)})}\right)$$

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}(\hat{z}_0 + z^{(1-\alpha)})}\right).$$

Chapter 3 - Estimation of mitigated fraction for two groups on an ordinal categorical response variable

In this chapter, several statistical methods will be discussed to estimate the mitigated fraction for two independent groups. An experiment involving two treatments such as vaccine and placebo will be considered for the initial study. Later, possible methods will be discussed for studies where more than two independent groups are compared to estimate the mitigated fraction. In an experiment involving two independent groups, experimental units are randomly divided into two treatment groups, experimental units in one group are treated with the vaccine, and experimental units in the other group are treated with a placebo. The response to treatment for each subject in terms of disease severity will be assumed to be an ordinal outcome.

Here, latent variable techniques and the values of the latent variables will mainly be used to estimate the mitigated fraction from ordinal response variables. As discussed in Chapter 2, an observed ordinal response is often taken to reflect an unobserved continuous latent variable.

Cox (1995) worked with location-scale cumulative models for ordinal data to investigate several problems on medical diagnosis. Cox (1995) applied disease severity on a latent continuous random variable scale to investigate problems, namely, ultrasound ratings as an ordinal measure of disease severity of different cancer types; severity of nausea on a 6-point scale was measured in groups of patients who received chemotherapy without and with cisplatin; and severity of coronary artery disease. Whitehead, Omar, Higgins, Savaluny, Turner and Thompson (2001) discussed a latent logistic random variable for analyzing ordinal responses with applications on medical diagnosis. Recently, Saho, Ma, Chen, Pan and You (2020) discussed a latent continuous random variable scale to investigate injury severity in truck-involved rear-end collisions from ordinal outcomes. Detailed discussion about latent variable techniques can be found in Agresti (2010). Moreover, the latent variable techniques contribute significantly to the ordinal data analysis via the data augmentation technique. Especially, the Bayesian methods have become popular because the data augmentation techniques can be used effectively using latent variables. Albert and Chib (1993) used data augmentation in their implementations of the ordinal probit models on the binary and ordinal response so that the full conditional distributions of parameters in the Gibbs sampler would be closed form of densities. Notably, in some problems, the values of the latent variables may be of interest. For example, Albert (1992) used the values of the latent variable to estimate the polychoric correlation coefficient between two ordinal variables. Especially, Poon (2004) discussed a latent normal distribution model for analyzing ordinal responses with applications on medical diagnosis. Poon's (2004) method was later referred by Siev (2005) to estimate MF for ordinal response.

3.1 Estimation of mitigated fraction for two groups

Assume that there is a single factor of treatment type with two levels (vaccine and control) as an explanatory variable. Let x_1 denote the explanatory variable as a group indicator for an observation, where $x_1 = 1$ for the control group and $x_1 = 0$ for the vaccine group. Also, let Y denote an independent ordinal random variable with J categories. The cumulative probability model will be:

$$\pi_j = G(\eta_j), \eta_j = \gamma_j - \beta_1 x_1 \text{ for } j = 1, \ldots, J-1,$$

where η_j is the linear predictor, G is inverse of an arbitrary link function, and

$\pi_j = P(Y \leq j | x_1) = \tilde{\pi}_1 + \cdots + \tilde{\pi}_j$ with $\sum_{j=1}^{J} \tilde{\pi}_j = 1$ are cumulative probabilities.

The components for the model summary for this vaccine study are as follows.

- Linear predictor: $\gamma_j - \beta_1 x_1$
- Distribution: $Y \sim Multinomial(n, \tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_J)$
- Link: G^{-1}such that $G^{-1}\big(P(Y \leq j|x_1)\big) = \eta_j$

The underlying latent variable Y^* for above model will be

$$Y^* = \beta_1 x_1 + \epsilon \tag{3.1}$$

Agresti and Kateri (2017) denoted that it may be sensible to consider an ordinal categorical variable as a necessarily crude measurement of the continuous latent variable Y^*, which would be the response variable in an ordinary linear model such as a model in the expression (3.1). The distribution of Y^* will be determined by the cdf of the error term ϵ.

Following model in the expression (3.1), let Y_1^* and Y_2^* denote independent underlying latent random variables such that when $x_1 = 1$ then Y_1^* for the control group and when $x_1 = 0$ then Y_2^* for the vaccine group. Siev (2005) and Agresti et al. (2017) argued that the measure T ($T = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$) applies directly to this latent variable model in the expression (3.1). In the following sections, different link functions are used to estimate the mitigated fractions; these are probit, logit, clog-log and log-log.

3.2 Estimation of mitigated fraction under probit link

When the probit link function is used, then the cumulative probit model will be

$$[P(Y \leq j|x_1)] = \Phi(\gamma_j - \beta_1 x_1) \text{ for } j = 1, \dots, J-1.$$

The underlying latent variable, Y^*, will be $Y^* = \beta_1 x_1 + \epsilon$.

For this probit model case, as ϵ follows the standard normal distribution, a normal latent variable Y^* given as $Y^* \sim N(\beta_1 x_1, 1)$. Letting Y_1^* and Y_2^* denote independent underlying latent variables such that when $x_1 = 1$ then Y_1^* for the control group and when $x_1 = 0$ then Y_2^* for the vaccine group. Thus, $Y_1^* \sim N(\beta_1, 1)$ and $Y_2^* \sim N(0,1)$, and the mitigated fraction will be driven as follows,

$$(Y_1^* - Y_2^*) \sim N(\beta_1, 2)$$

$$P(Y_1^* > Y_2^*) = P(Y_1^* - Y_2^* > 0)$$

$$P(Y_1^* - Y_2^* > 0) = P\left(\frac{(Y_1^* - Y_2^*) - \beta_1}{\sqrt{2}} > -\frac{\beta_1}{\sqrt{2}}\right)$$

$$P(Y_1^* - Y_2^* > 0) = P\left(Z > -\frac{\beta_1}{\sqrt{2}}\right) = \Phi\left(\frac{\beta_1}{\sqrt{2}}\right)$$

where Z follows a standard normal distribution, Φ denotes the standard normal cumulative distribution function. $T = P(Y_1^* > Y_2^*) + \frac{1}{2}P(Y_1^* = Y_2^*)$ and for the underlying continuous latent variable $P(Y_1^* = Y_2^*) = 0$ and then $T = P(Y_1^* > Y_2^*)$.

$$MF = 2T - 1 = 2 * \Phi\left(\frac{\beta_1}{\sqrt{2}}\right) - 1$$

3.3 Estimation of mitigated fraction under logit link

When the logit link function is used, then the cumulative logit model will be

$$[P(Y \leq j|x_1)] = \frac{exp\ (\gamma_j - \beta_1 x_1\ )}{1 + exp\ (\gamma_j - \beta_1 x_1\ )} \text{ for } j = 1, \ldots, J - 1.$$

The underlying latent variable, Y^*, will be $Y^* = \beta_1 x_1 + \epsilon$.

Under the logit model, as ϵ follows a standard logistics distribution, a latent variable will be given as $Y^* \sim logistic(\beta_1 x_1, 1)$ and thus $Y_1^* \sim logistic(\beta_1, 1)$ and $Y_2^* \sim logistic(0,1)$. Now when it comes to the distribution of $(Y_1^* - Y_2^*)$, this one does not follow a logistic distribution or any other closed form distribution. A numerical solution can be performed using Monte Carlo simulation methods to estimate the value of $P(Y_1^* > Y_2^*)$.

Let us briefly discuss Monte Carlo integration methods.

3.2.1 Monte Carlo integration

Monte Carlo methods are numerical techniques which rely on random sampling to approximate a result of interest. Monte Carlo integration applies this process to the numerical estimation of integrals when an analytical solution to an integration problem is not feasible or hard to implement.

Expected value of a random variable

Let X be a random variable with a probability density function $f(x)$. Suppose that our goal is to find the expected value or expectation of a function of the random variable X, say $g(X)$. The expected value or expectation of $g(x)$ over a domain $\mu(x)$ is defined as

$$E[g(X)] = \int_{\mu(x)} g(x)f\ (x)d\mu(x)\,.$$

If $x_1, \ldots, x_\mathrm{n}$ is a random sample from $f(x)$, then the strong law of large numbers (SLLN) ensures that the Monte Carlo estimate converges to the true value of the integral:

$$\frac{1}{n}\sum_{i=1}^{n} g(x_i) \overset{a.s}{\rightarrow} E[g(X)] = \int_{\mu(x)} g(x)f\ (x)d\mu(x)$$

as $n \to \infty$.

3.2.2 Expected value of an indicator variable

Given a probability space $(\Omega\ , \mathcal{F}, P)$ with $A \in \mathcal{F}$, the indicator variable $I_A : \Omega \to \mathbb{R}$ is defined by $I_A(\omega) = 1$ if $\omega \in A$, otherwise $I_A(\omega) = 0$. Here Ω is the sample space; $\mathcal{F}$ is σ-field or σ-algebra; P is probability measure mapping from $\mathcal{F}$ to [0,1]; and A is a subset of Ω (Rosenthal, 2006).

The mean value of $I_A(\omega)$, that is $E(I_A(\omega))$ will be given as $E\big(I_A(\omega)\big) = P(A)$.

Using the above fact, the value of $P((Y_1^* > Y_2^*)$ will be equal to the expected value of the indicator random variable $I_{(Y_1^* > Y_2^*)}$,

$$P(Y_1^* > Y_2^*) = E(I_{(Y_1^* > Y_2^*)})$$

where $I_{(Y_1^* > Y_2^*)} = 1$ if $Y_1^* > Y_2^*$, otherwise $I_{(Y_1^* > Y_2^*)} = 0$, and now

$$\frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^* > Y_2^*)} \overset{a.s}{\rightarrow} E\left[I_{(Y_1^* > Y_2^*)}\right]$$

as $n \rightarrow \infty$ by the SLLN.

Now the estimation for $P(Y_1^* > Y_2^*)$ will be equal to $\frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^* > Y_2^*)}$, and thus the estimated MF will be equal to $2 * \frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^* > Y_2^*)} - 1$.

An alternative way to estimate the value of $P(Y_1^* > Y_2^*)$, a normal approximate solution can be deployed for the logistic distribution. The logistic distribution is a location-scale family, and it is very similar to the normal distribution. Both distributions are symmetric and bell-shaped, though the logistic distribution has heavier tails than normal distribution. Due to these similarities, it is appropriate to approximate the logistic distribution using a normal distribution. Let $F(x) = (1 + e^{-x})^{-1}$ be the logistic cumulative density function and $G(x)$ be the normal cumulative distribution function with mean 0 and standard deviation δ and thus logit(x) $\approx$ $\delta * \Phi^{-1}(x)$. Haley (1952) outlined a theoretical derivation of $\delta = 1.702$ which was based on minimx criteria, $\delta = \underset{\delta}{min}\,\underset{x}{max} \parallel F(x) - G(x) \parallel_2$. Johnson and Kotz (1970) showed graphically that a factor $\delta = \frac{\pi}{\sqrt{3}}(15/16) = 1.70044$ would approximate the logistic distribution. Further, Gregory Camilli (1994) reframed Haley's (1952) work, discussed the minimax criteria in detail, and showed the same normal approximation to the logistics distribution with $\delta = 1.70174$. Savalei (2006) proposed a value for $\delta = 1.749$ based on minimizing Kullback-Leibler (KL) information. Therefore, the standard logistic distribution will be approximated by a normal distribution with mean 0 and standard deviation δ given by one of the above discussed methods. Now, $Y_1^* \sim N(\beta_1, \delta^2)$ and $Y_2^* \sim N(0, \delta^2)$, where δ is the scale parameter and now $P(Y_1^* - Y_2^* > 0) \approx \Phi\left(\frac{\beta_1}{\sqrt{2}\delta}\right)$ and it follows immediately that estimated $MF = 2 * \Phi\left(\frac{\beta_1}{\sqrt{2}\delta}\right) - 1$.

3.3 Estimation of mitigated fraction under log-log link

When the log-log link function is used, then

$$[P(Y \leq j|x_1)] = exp\big(-exp\big(-\gamma_j + \beta_1 x_1\ \big)\big)\ for\ j = 1, \dots, J - 1.$$

The underlying latent variable, Y^*, will be $Y^* = \beta_1 x_1 + \epsilon$.

Under the log-log model, as ϵ follows a standard Gumbel distribution (maximum), a latent variable will be given as $Y^* \sim Gumbel_{max}\ (\beta_1 x_1, 1)$ and thus $Y_1^* \sim Gumbel_{max}(\beta_1, 1)$ and $Y_2^* \sim Gumbel_{max}(0,1)$. The standard Gumbel maximum distribution with location parameter equals to 0 and scale parameter equals to 1 is given as $F(x) = exp\ (-exp\ (-x))$. Following Agresti et al. (2017) and McFadden(1974), the distribution of $(Y_1^* - Y_2^*)$ follows a logistic distribution, $(Y_1^* - Y_2^*) \sim logistic(\beta_1, 1)$. Now it is trivial to show that $P(Y_1^* > Y_2^*) = \frac{exp(\beta_1)}{1+exp(\beta_1)}$ and thus $MF = 2 * \frac{exp(\beta_1)}{1+exp(\beta_1)} - 1$.

3.4 Estimation of mitigated fraction under clog-log link

When the clog-log link function is used, then the cumulative logit model will be

$$[P(Y \leq j|x_1)] = 1 - exp\big(-exp\big((\gamma_j - \beta_1 x_1)\big)\big) \text{ for } j = 1, \dots, J - 1.$$

The underlying latent variable, Y^*, will be $Y^* = \beta_1 x_1 + \epsilon$.

Under the logit model, as ϵ follows a standard Gumbel distribution (minimum), an underlying latent variable will be given as $Y^* \sim Gumbel_{min}\ (\beta_1 x_1, 1)$ and thus $Y_1^* \sim Gumbel_{min}(\beta_1, 1)$ and

$Y_2^* \sim Gumbel_{min}(0,1)$. Now when it comes to the distribution of $(Y_1^* - Y_2^*)$, this one does not follow a Gumbel distribution (minimum) or any other closed form distribution. The Monte Caro simulation methods was used to estimate $p(Y_1^* - Y_2^*)$. Thus, it immediately follows the estimation for $P(Y_1^* > Y_2^*)$ will be equal to $\frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^* > Y_2^*)}$, and thus the estimated MF will be equal to $2 * \frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^* > Y_2^*)} - 1$.

3.5 True value for mitigated fraction

Let π_{ij} be the probability for the j^{th} category for i^{th} group (either placebo or vaccine) with $j = 1,2, \dots, J$ and $i = 1,2$, then the $T = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ can be calculated by

$$T = \sum\sum_{j>k} \pi_{1j}\pi_{2k} + \frac{1}{2}\sum_{j=k} \pi_{1j}\pi_{2k}$$

for $j, k = 1,2, \dots, J$ (Agresti et al., 2017).

If the true values for $\pi_{ij}{}'s$ are known, then $MF = 2T - 1$.

If the true values for $\pi_{ij}{}'s$ are not known, then the estimated mitigated fraction equals $2\hat{T} - 1$ with

$$\hat{T} = \sum\sum_{j>k} \hat{\pi}_{1j}\hat{\pi}_{2k} + \frac{1}{2}\sum_{j=k} \hat{\pi}_{1j}\hat{\pi}_{2k}$$

where $\hat{\pi}_{1j}$ and $\hat{\pi}_{2k}$ are fitted values from the corresponding link function of cumulative probability model. It is to be noted that the mitigated fraction can be estimated without latent variable assumption using the above measure and estimated MF equal to $2\hat{T} - 1$.

3.6 Simulation study

A simulation study was conducted to assess the performance of the proposed method to calculate the mitigated fraction. The performance is evaluated by computing the confidence interval for MF using the criteria of coverage probability. That is the proportion of time the true parameter value of the mitigated fraction contained within the confidence interval. All confidence intervals were calculated using bootstrap method described in Section 2.6. Details of this simulation study are given below.

First define a response variable that comes from a multinomial distribution, i.e.

$n_1, n_2, \dots, n_J \sim Multinomial(N, \tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_J)$.

- n_j denotes the number of observations in category j ($j = 1, \dots, J$) with $\sum_{j=1}^{J} n_j = N$, where N is the total number of observations for each treatment.
- $\tilde{\pi}_j$ is the probability a subject falls into the j^{th} category with $\sum_{i=1}^{J} \tilde{\pi}_j = 1$.
- For this simulations study, five ordinal categories (c_0, c_1, c_2, c_3 and c_4) were considered in the order of increasing severity, i.e. c_0 is the lowest level of the categories and c_4 is the highest level of the categories.
- A single explanatory variable, x_1, was used as a group indicator with $x_1 = 1$ for the control group and $x_1 = 0$ for the vaccine group.

The simulation study was conducted using the following model. Let π_j be the cumulative probabilities such that $\pi_j = P(Y \leq j|x_1) = \tilde{\pi}_1 + \cdots + \tilde{\pi}_j$ with $\sum_{j=1}^{J} \tilde{\pi}_j = 1$.

The cumulative probability model will be:

$\pi_j = G(\eta_j), \eta_j = \gamma_j - \beta_1 x_1$ for $j = 1, \dots, J - 1$,

where η_j is the linear predictor, G is inverse of an arbitrary link function.

The simulation consists of the following steps:

1. For given values of $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, and β_1, the linear predictor, η_j was calculated. The values for $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, and β_1 used in the simulation study are given in the description of the Table 3.1.

2. Probabilities $(\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \tilde{\pi}_4, \tilde{\pi}_5)$ for vaccine and control groups were found based on the following model with given a link function.

   $$P(Y \leq j|x_1) = G(\eta_j) \text{ for } j = 1,2,3,4.$$

   Note: The cell probabilities $(\tilde{\pi}_1, \tilde{\pi}_2, \dots, \tilde{\pi}_J)$ can be found in such a way that $\tilde{\pi}_j = \pi_j - \pi_{j-1}$ for $j = 2, .., J$ with $\tilde{\pi}_1 = \pi_1$ and $\pi_J = 1$.

3. Using the probabilities, i.e. $\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \tilde{\pi}_4, \tilde{\pi}_5$, for vaccine and control, independent samples containing 100 observations each were generated for the two treatment groups respectively (i.e.100 observations for control group and 100 observations for vaccine group.

A total of 10,000 simulated data sets were generated for each parameter setting for a given link function. For each data set, a 95% confidence interval for the mitigated fraction was found using the bootstrap technique discussed in Section 2.6. The simulation results for all link functions are presented in Table 3 where the true value of MF has been calculated using the method discussed in Section 3.5. For a true coverage rate of 0.95, the Monte Carlo simulation standard error is approximately 0.0022.

Table 3.1. Results for simulation study of 95% confidence interval

Link		Estimated coverage rate	95% CI for the true coverage rate
log-log		0.9524	(0.94813, 0.95667)
probit		0.9521	(0.94783, 0.95637)
clog-log (Monte Carlo simulation)		0.9511	(0.94683, 0.95537)
Logit (Monte Carlo simulation)		0.9504	(0.94613, 0.95467)
Logit (Normal approximation)	$\delta = \left(\frac{15}{16}\right)\left(\frac{\pi}{\sqrt{3}}\right)$	0.9533	(0.94903, 0.95757)
	$\delta = 1.749$	0.9546	(0.95033, 0.95887)
	$\delta = 1.702$	0.9526	(0.94833, 0.95687)
	Agresti (2017)	0.952	(0.94773, 0.95627)

The values for $\gamma_1 = -1.5, \gamma_2 = -0.3, \gamma_3 = 1, \gamma_4 = 2$, and $\beta_1 = 0.3$ are used for this simulation study.

From Table 3.1, it is evident that the latent variable approach to calculate MF yields promising results. All, but one, of the 95% CI for the true coverage rate contains the true coverage rate (0.95) based on the 10,000 simulations.

3.7 Model comparison and model selection

Different link functions such as probit, logit, loglog and cloglog can be used to estimate the mitigated fraction. Based on the simulation study, all link functions give promising results to estimate the mitigated fraction. In the light of selecting the best link function to estimate the mitigated fraction, a simulation study was conducted. The best link function will be selected based on the percentage of bias-to-true value of mitigated fraction.

The simulation consists of the following steps:

1. Using the probabilities, i.e. $\tilde{\pi}_1, \tilde{\pi}_2, \tilde{\pi}_3, \tilde{\pi}_4, \tilde{\pi}_5$, as described in Section 3.6, for vaccine and control, independent samples containing 100 observations each were generated for the two treatment groups respectively (i.e.100 observations for control group and 100 observations for vaccine group.
2. Step 1 was repeated and a total of 10,000 simulated data sets were randomly generated for each parameter setting for a given link function. For each data set, the mitigated fraction was found using the latent variable approach, that is 10,000 estimated mitigated fraction values were found using the latent variable approach.
3. The mean value of the 10,000 estimated mitigated fraction values was found for each link function.
4. The true value of the mitigated fraction was found for each link function using the method discussed in Section 3.5.
5. The bias was found for each link function as follows

$$Bias = \ mean\ value\ of\ the\ 10{,}000\ estimated\ mitigated\ fraction\ values$$
$$-\ true\ value\ of\ the\ migtaged\ fraction.$$

The simulation results for all link functions are presented in Table 3.2.

Table 3.2. Percentage of Bias-True value of mitigated fraction comparison

Link	Mean Estimated MF	True Value	Bias	Bias-to-true value percentage
Probit	0.09634182	0.08563379	0.01070803	12.50%
Logit (Monte Carlo simulation)	0.08948851	0.0766	0.01289	16.83%
loglog	0.08744318	0.07650079	0.01094	14.30%
cloglog	0.1443742	0.1244831	0.01989	15.98%

3.8 Real data analysis

This data set is from Poon (2004), and the origin of this data set was found in Whitehead et al. (2001). This data set was used as a benchmark for this analysis and was collected in a vaccine trial on administering the anti-cholinesterase drug tacrine to patients with Alzheimer's disease. Patients receiving tacrine (T) and placebo (P) are, respectively, classified into five categories: 'very much improved' (c_0), 'minimally improved' (c_1), 'no change' (c_2), 'minimally worse' (c_3) and 'very much worse' (c_4). Here 'c_0' is the clinically best response and 'c_4' is the worst (order of increasing disease severity). Whitehead et al. (2001) and Poon (2004) have suggested that there is a tendency for patients receiving the treatment 'Tacrine' to have better responses than those receiving placebo. The data is given in Table 3.3 and the analysis results are presented in the Table 3.4. As discussed in the Section 3.5, the mitigated fraction can be estimated without using the latent variable, and the analysis results are presented in the Table 3.5.

Table 3.3. Patients with Alzheimer's disease for the tacrine study

Severity	P	T
c_0	2	4
c_1	22	23
c_2	54	45
c_3	29	22
c_4	3	2
Total	110	96

Data source: Poon, W. (2004). A latent normal distribution model for analyzing ordinal responses with applications in meta-analysis. Statistics in medicine, 23:2155–2172.

Table 3.4. Results of estimated values for the mitigated fraction for the Tacrine study

Link		$\boldsymbol{T = P(Y_1 > Y2)}$	$\boldsymbol{MF = 2T - 1}$	**95% bootstrap Confidence interval**
probit		0.55	0.10	(-0.08, 0.26)
log-log		0.55	0.10	(-0.06, 0.24)
clog-log (Monte Carlo simulation)		0.54	0.09	(-0.05, 0.26)
logit (Monte Carlo simulation)		0.55	0.10	(-0.08, 0.24)
Logit (Normal approximation)	$\delta = \left(\frac{15}{16}\right)\left(\frac{\pi}{\sqrt{3}}\right)$	0.55	0.09	(-0.08, 0.27)
	$\delta = 1.749$	0.55	0.09	(-0.07, 0.25)
*Siev (2005)		0.54	0.08	(-0.07,0.23)
*Poon (2004)		0.55	0.10	(-0.11, 0.30)

*These values have been directly taken from Siev (2005).

Table 3.5. Results of estimated values of MF for the Tacrine study without latent variable

Link	$\boldsymbol{T = P(Y_1 > Y2)}$	$\boldsymbol{MF = 2T - 1}$	**95% bootstrap Confidence interval**
probit	0.54	0.09	(-0.06, 0.22)
Logit	0.54	0.08	(-0.07, 0.23)
log-log	0.54	0.09	(-0.05, 0.22)
clog-log	0.53	0.07	(-0.07, 0.19)
*Siev (2005)	0.54	0.08	(-0.07, 0.23)
*Poon (2004)	0.55	0.10	(-0.11, 0.30)

*These values have been directly taken from Siev (2005).

3.9 Interpretation of T and MF

Recall that the mitigated fraction was calculated using $MF = P(Y_1 > Y_2) - P(Y_2 > Y_1)$, where Y_1 denotes the non-vaccinated group and Y_2 denotes the vaccinated group. The interpretation of the mitigated fraction is somewhat not straight forward from a technical standpoint. For instance, when the probit model is used, from Table 3.4, $T = 0.55$ (55%) means 55% of the non-vaccinates are expected to be more severely affected than the vaccinates. Again, under the probit model, from Table 3.4, $MF = 0.1$ (10%) indicates that there is a 10% increase in the chance that a vaccinate's disease will be less severe than a non-vaccinate's disease over the chance a non-vaccinate's disease being less severe than a vaccinate's.

Chapter 4 - Mitigated fraction for clustered data

4.1 Theoretical setup for mitigated fraction for clustered data

Blocking or stratified designs are common in many veterinary and epidemiological studies. In many studies, it is impossible to select homogeneous experimental units. When experiments are not homogeneous, one method is to group experimental units into sets of nearly alike experimental units. Generally, groups of similar experimental units are called blocks (Milliken and Johnson, 2009). Each block has a set of experimental units which are similar, such as pens of cows, litters of mice, or litters of piglets. The variability between units within a block will be less than that of units from different blocks. Moreover, clustering of observations arises in a repeated measures study where each subject on which the measurements are collected is treated as a cluster.

First, let us define a generalized linear mixed model that accounts for a cluster effect. Let Y_{it} denote the response for observation t in block i. Let $\boldsymbol{X}_{it} = \left(x_{1it}, x_{2it}, \dots, x_{pit}\right)^T$ denote the values of the p-dimensional vector of explanatory variables for that observation. Let $\boldsymbol{\beta}$ be a p-dimensional column vector of parameter coefficients. Also, let $\boldsymbol{z}_{it}$ be a q-dimensional vector of covariates associated with a q-dimensional random column vector of parameter coefficients, $\boldsymbol{b_i}$. Let $\pi_{itj}(\boldsymbol{b_i}) = P(Y_{it} \leq j|\boldsymbol{b_i})$ be the conditional cumulative probability given cluster effect $\boldsymbol{b_i}$ and the cumulative probability is typically modeled as,

$$G^{-1}\left(\pi_{itj}(\boldsymbol{b_i})\right) = \gamma_j - \boldsymbol{X}'_{it}\boldsymbol{\beta} - \boldsymbol{z}'_{it}\boldsymbol{b_i} \text{ for } j = 1, \dots, J-1 \tag{4.1}$$

where G^{-1} is an arbitrary link function.

The above model is typically called as **cluster-specific** model. Following Agresti (2010), a conditional latent variable could be written for this model as,

$$Y_{it}^{*}|\boldsymbol{b_i} = \boldsymbol{X}_{it}'\boldsymbol{\beta}+ \boldsymbol{z}_{it}'\boldsymbol{b_i} + \epsilon_{it},$$

The above latent variable is a conditional variable based on cluster effect $\boldsymbol{b_i}$ and the estimation of the mitigated fraction requires a marginal latent variable Y_{it}^{*}. Y_{it}^{*} could easily be derived from a marginal model which is typically called a **population-averaged** model (Agresti and Natarajan, 2001).

Let $\pi_{itj} = P(Y_{it} \leq j)$ be marginal cumulative probability which could be derived as follows,

$$\pi_{itj} = E\big[G(\gamma_j - \boldsymbol{X}_{it}'\boldsymbol{\beta} - \boldsymbol{z}_{it}'\boldsymbol{b_i})\big],$$

where the expectation is with respect to the distribution of random effect $\boldsymbol{b_i}$**.** It could similarly be expressed as,

$$\pi_{itj} = \int G(\gamma_j - \boldsymbol{X}_{it}'\boldsymbol{\beta} - \boldsymbol{z}_{it}'\boldsymbol{b_i})\, \boldsymbol{dF}(\boldsymbol{b_i}).$$

Here, $\boldsymbol{F}$ is the distribution of random effects $\boldsymbol{b_i}$ and $\boldsymbol{F}$ is generally assumed to be a Gaussian distribution with mean $\boldsymbol{0}$ and variance-covariance matrix $\boldsymbol{D}$.

Now to model the marginal cumulative probability π_{it}, we can assume the same link such that

$$G^{-1}(\pi_{it}) = G^{-1}\big[\int G(\gamma_j - \boldsymbol{X}_{it}'\boldsymbol{\beta} - \boldsymbol{z}_{it}'\boldsymbol{b_i})\, \boldsymbol{dF}(\boldsymbol{b_i})\big].$$

Note that we do not normally expect for the right-hand side of the above equation to be $\gamma_j - \boldsymbol{X}_{it}'\boldsymbol{\beta}$ with same parameter values as in cluster-specific model. But when identity link, $G^{-1} = I$, is used, it is trivial and $\pi_{it}(\boldsymbol{b_i}) = \gamma_j - \boldsymbol{X}_{it}'\boldsymbol{\beta}- \boldsymbol{z}_{it}'\boldsymbol{b_i}$. In general, if the conditional cumulative probabilities in the cluster-specific model hold with an arbitrary link G^{-1} then the marginal cumulative probabilities in marginal model will not hold with same link G^{-1} (Agresti et al., 2001).

The probit link used for the cluster-specific model, $\Phi^{-1}(\pi_{it}(\boldsymbol{b_i}))$, with Gaussian distribution for $\boldsymbol{b_i}$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$**,** gives a marginal probit model with different parameters as given as,

$$\pi_{it} = \Phi\left[\left(\gamma_j - \boldsymbol{X}'_{it}\boldsymbol{\beta}\right) * a_p(\boldsymbol{D})\right], \qquad \text{for } j = 1,2,\dots,J-1 \tag{4.2}$$

where $a_p(\boldsymbol{D}) = |\boldsymbol{D}\boldsymbol{z}_{it}\boldsymbol{z}'_{it} + I|^{-\boldsymbol{q}/\boldsymbol{2}}$ with q as the dimension of $\boldsymbol{b_i}$ (Zeger, Liang and Albert, 1988).

When the logit link is used for the cluster-specific model, $logit(\pi_{it}(\boldsymbol{b_i}))$, with Gaussian distribution for $\boldsymbol{b_i}$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$, the marginal cumulative probabilities π_{it} do not have a closed-form expression. An approximate expression was discussed in Zeger et al. (1988) with the use of Gaussian approximation to logistics function given by,

$$logit(\pi_{it}) \approx \left(\gamma_j - \boldsymbol{X}'_{it}\boldsymbol{\beta}\right) * a_l(\boldsymbol{D}), \qquad \text{for } j = 1,2,\dots,J-1 \tag{4.3}$$

where $a_l(\boldsymbol{D}) = |c^2\boldsymbol{D}\boldsymbol{z}_{it}\boldsymbol{z}'_{it} + I|^{-\boldsymbol{q}/\boldsymbol{2}}$ with $c = 16\sqrt{3}/15\pi$.

Having the above general set up for clustered models, let us discuss a simple clustered data analysis for a vaccine study with a (generalized) randomized complete block design where two treatments (either vaccine or placebo) are assigned to experimental units completely at random within each block. Here there is a single factor of treatment type with two levels (vaccine and placebo) as the explanatory variable. Let x_{it} denote the explanatory variable as a group indicator for an observation where $x_{it} = 1$ for the placebo group and $x_{it} = 0$ for the vaccine group. For this simple setup $p = 1$ for the single explanatory variable resulting in $\boldsymbol{\beta} = \beta_1$, $\boldsymbol{X_{it}} = x_{it}$ and $q = 1$ for single random block effect with $\boldsymbol{b_i} = b_i$, and $\boldsymbol{z}'_{it} = 1$.

For the probit link, the cluster specific model becomes

$$\Phi^{-1}\left(\pi_{itj}(b_i)\right) = \gamma_j - x_{it}\beta_1 - b_i \text{ for } j = 1,\dots,J-1.$$

Now the marginal model equation (4.2) becomes

$$\pi_{itj} = P(Y_{it} \leq j) = \Phi\left[\left(\gamma_j - x_{it}\beta_1\right) * (1 + \sigma_b^2)^{-1/2}\right].$$

An underlying normal latent variable Y_{it}^* will be given as $Y_{it}^* \sim N\left((1 + \sigma_b^2)^{-\frac{1}{2}}\beta_1 x_{it}, 1\right)$. Let Y_1^* and Y_2^* denote independent underlying latent variables such that when $x_{it} = 1$ then Y_1^* for the placebo group and when $x_{it} = 0$ then Y_2^* for the vaccine group.

Thus, $Y_1^* \sim N\left((1 + \sigma_b^2)^{-\frac{1}{2}}\beta_1, 1\right)$ and $Y_2^* \sim (0,1)$.

And it immediately follows that $MF = 2T - 1 = 2 * \Phi\left(\frac{(1+\sigma_b^2)^{-\frac{1}{2}}\beta_1}{\sqrt{2}}\right) - 1$.

Relationship between mitigated fraction and block variance under probit model

The expression for estimated MF for probit model is $\left\{2 * \Phi\left(\frac{(1+\sigma_b^2)^{-\frac{1}{2}}\beta_1}{\sqrt{2}}\right) - 1\right\}$. It can clearly be seen from this expression that the argument for the standard normal cdf, Φ, depends on the value of the block variance. For a fixed value of β_1, when the value of the block variance increases, the value for the Φ tends to 0.5, and thus the estimated MF tends to 0. When the value for MF is 0, then there is no severity reduction advantage due to vaccine and thus there will be no difference between vaccine and placebo. The figure 1 shows a relationship between mitigated fraction and block variance for $\beta_1 = 0.3$. Care must be taken when choosing value for block variance compared to a given value for β_1 for the simulation study. Table 8 show the central coverage probability for different values of block variance with a fixed value of β_1.

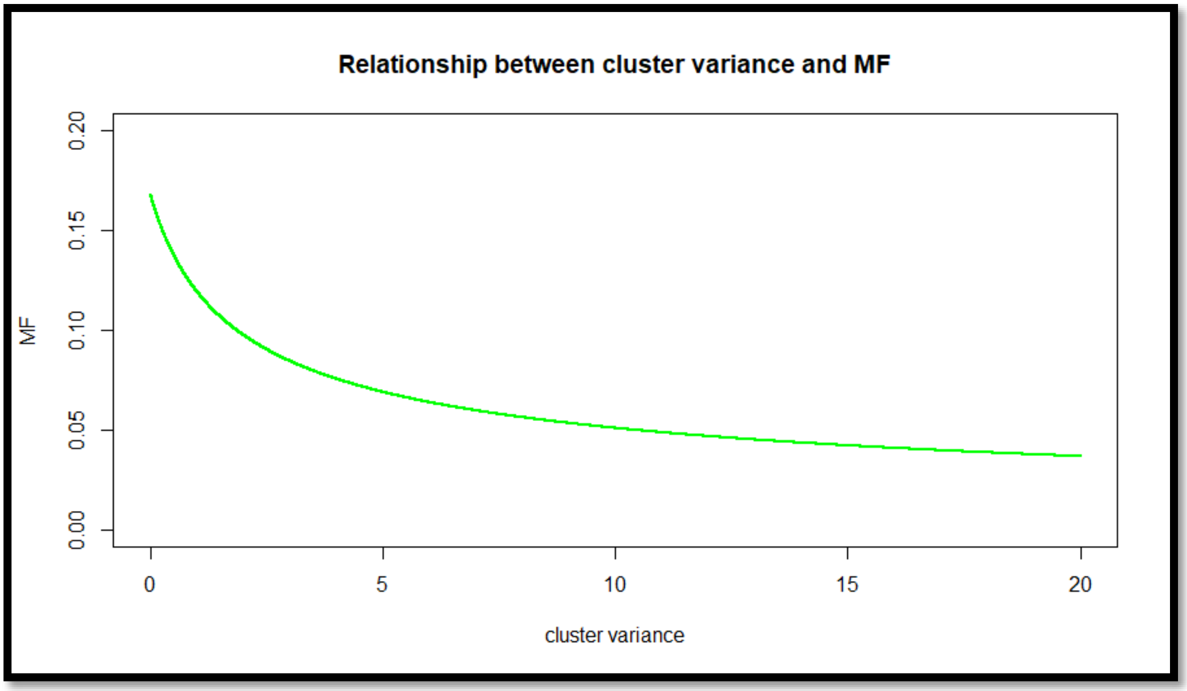

Figure 4.1. Relationship between mitigated fraction and block variance

For the logit link, the cluster specific model becomes

$$logit\left(\pi_{itj}(b_i)\right) = \gamma_j - x_{it}\beta_1 - b_i$$

Now the marginal model comes directly from equation (4.3)

$$logit(\pi_{itj}) = logit[P(Y_{it} \leq j)] = \left[\left(\gamma_j - x_{it}\beta_1\right) * (1 + c^2\sigma_b^2)^{-1/2}\right].$$

Again, under logit model the underlying latent variable will be given as $Y_{it}^{*} \sim logistic\left((1 + c^2\sigma_b^2)^{-\frac{1}{2}}\beta_1 x_{it}, 1\right)$ and thus $Y_1^{*} \sim logistic((1 + c^2\sigma_b^2)^{-\frac{1}{2}}\beta_1, 1)$ and $Y_2^{*} \sim logistic(0,1)$.

Again, a numerical solution using Monte Carlo simulation methods, discussed in Chapter 3, was used for the logistic distribution as the distribution of $(Y_1^{*} - Y_2^{*})$ does not follow a closed form distribution. The value of $P(Y_1^{*} > Y_2^{*})$ was estimated using Monte Carlo simulation method and therefore it immediately follows that the estimation for the MF will be equal to $2 * \frac{1}{n}\sum_{i=1}^{n} I_{(Y_1^{*} > Y_2^{*})} - 1$.

4.2 True value for mitigated fraction

Following the discussion in Section 4.1, the marginal probabilities can be found under probit and logit link models. Let π_{ij} be the marginal probability for the *j*th category for *i*th group (either placebo or vaccine) with $j = 1,2, \dots, J$ and $i = 1,2$ (The equation (4.1) and equation (4.2) can be used to find the marginal probabilities for probit and logit link models respectively). Now $T = P(Y_1 > Y_2) + \frac{1}{2}P(Y_1 = Y_2)$ can be calculated by

$$T = \sum\sum\nolimits_{j>k} \pi_{1j}\pi_{2k} + \frac{1}{2}\sum_{j=k} \pi_{1j}\pi_{2k}$$

for $j, k = 1,2, \dots, J$ (Agresti et al., 2017).

If the true values for $\pi_{ij}{}'$s are known, then $MF = 2T - 1$.

If the true values for $\pi_{ij}{}'$s are not known, then the estimated mitigated fraction equals $2\hat{T} - 1$ with

$$\hat{T} = \sum\sum\nolimits_{j>k} \hat{\pi}_{1j}\hat{\pi}_{2k} + \frac{1}{2}\sum_{j=k} \hat{\pi}_{1j}\hat{\pi}_{2k}$$

where $\hat{\pi}_{1j}$ and $\hat{\pi}_{2k}$ are fitted values from the corresponding link function of cumulative probability model. And hence the mitigated fraction can be estimated without latent variable assumption using the above measure and estimated MF equal to $2\hat{T} - 1$.

4.3 Simulation study

First define a response variable that comes from the multinomial distribution with a random blocking factor, i.e.

$n_{it1}, n_{it2}, \dots, n_{itJ} | b_i \sim Multinomial(N_{it}, \pi'_{it1}, \pi'_{it2}, \dots, \pi'_{itJ})$.

- n_{itj} denotes number of observations in category j for treatment t and block i with $\sum_{j=1}^{J} n_{itj} = N_{it}$, where the N_{it} is the total number of observations for treatment t in block i.

- π'_{itj} is the probability of j^{th} category for treatment t in cluster i with $\sum_{j=1}^{J} \pi'_{itj} = 1$.
- b_i denotes the block effect for block i with $b_i \sim N(0, \sigma_b^2)$; the number of blocks was set to 8 for the simulation study.
- For this simulations study, five ordinal categories (c_0, c_1, c_2, c_3 and c_4) were considered in the order of increasing severity, i.e. c_0 is the lowest level of the categories and c_4 is the highest level of the categories.
- A single explanatory variable, x_{it}, was used as a group indicator with $x_{it} = 1$ for the control group and $x_{it} = 0$ for the vaccine group.

The simulation study was conducted using the following model. Let π_{itj} be the cumulative probabilities such that $\pi_{itj} = P(Y_{it} \leq j | x_{it}) = \pi'_{it1} + \pi'_{it2} + \cdots + \pi'_{itj}$ with $\sum_{j=1}^{J} \pi'_{itj} = 1$. The cumulative probability model will be:

$$\pi_{itj} = G(\eta_j), \eta_j = \gamma_j - x_{it}\beta_1 - b_i \text{ for } j = 1,2,3,4.$$

where η_j is the linear predictor, G is inverse of an arbitrary link function.

The simulation consists of the following steps:

1. First, simulate the block effects. Block effects were randomly drawn from a normal distribution with a mean of 0 and a particular variance. Different values for the block variance were chosen and the values are given in the description under Table 8 and Table 9. Since there were 8 blocks, 8 block effects were drawn for each simulated data set, with each value repeated for each observation within the same block.

2. For each block, given value of $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, β_1, and σ_b^2, the linear predictor, η_j was found. The values for $\gamma_1, \gamma_2, \gamma_3, \gamma_4$, β_1, and σ_b^2 used in the simulation study are given in the description of the Table 8 and Table 9.
3. Probabilities $(\pi'_{it1}, \pi'_{it2}, \pi'_{it3}, \pi'_{it4}, \pi'_{it5})$ for vaccine and control groups were found based on the following model with given a link function.

   $$P(Y_{it} \leq j|x_{it}) = G(\eta_j) \text{ for } j = 1,2,3,4.$$

   Note: The cell probabilities $(\pi'_{it1}, \pi'_{it2}, \pi'_{it3}, \pi'_{it4}, \pi'_{it5})$ can be found in such a way that $\pi'_{itj} = \pi_{itj} - \pi_{itj-1}$ for $j = 2, . . , J$ with $\pi'_{it1} = \pi_{it1}$ and $\pi_{itJ} = 1$.
4. Using the probabilities, i.e. $\pi'_{it1}, \pi'_{it2}, \pi'_{it3}, \pi'_{it4}, \pi'_{it5}$, for vaccine and control, independent samples containing 100 observations each were generated for the two treatment groups respectively (i.e.100 observations for control group and 100 observations for vaccine group.

A total of 10,000 simulated data sets were generated for each parameter setting for a given link function. For each data set, a 95% confidence interval for the mitigate fraction was found using the bootstrap technique. The simulation results for probit link and logit link function are presented in Table 4.1 and Table 4.2 respectively where the true value of MF has been calculated using the method discussed in Section 4.2. For a true coverage rate of 0.95, the Monte Carlo simulation standard error is approximately 0.0022.

Table 4.1. Results for simulation study of 95% confidence interval from the probit link

β_1	σ_b^2	σ	Estimated coverage rate	95% CI for the true coverage rate
0.3	0.01	0.10	0.950	(0.9457, 0.9543)
0.3	0.1	0.32	0.949	(0.9447, 0.9533)
0.3	0.2	0.45	0.953	(0.9487, 0.9573)
0.3	0.3	0.55	0.948	(0.9437, 0.9523)
0.3	0.6	0.77	0.947	(0.9427, 0.9513)
0.3	0.9	0.95	0.944	(0.9397, 0.9483)

The values of (0.01,0.1,0.2,0.3,0.6,0.9) for σ_b^2 and $\gamma_1 = -1.5, \gamma_2 = -0.3, \gamma_3 = 1, \gamma_4 = 2$, and $\beta_1 = 0.3$ are used for this simulation study.

From Table 4.1, it is evident that the latent variable approach using the probit link to calculate MF yields promising results. All, but one, of the 95% CI for the true coverage rate contains the true coverage rate (0.95) based on the 10,000 simulations.

Table 4.2. Results for simulation study of 95% confidence interval from the logit link

β_1	σ_b^2	σ	Estimated coverage rate	95% CI for the true coverage rate
0.5	0.1	0.32	0.952	(0.9477, 0.9563)
0.5	0.5	0.71	0.947	(0.9447, 0.9533)
0.5	1.5	1.22	0.946	(0.9417, 0.9503)

The values of (0.1,0.5, 1.5) for σ_b^2 and $\gamma_1 = -1.5, \gamma_2 = -0.3, \gamma_3 = 1, \gamma_4 = 2$, and $\beta_1 = 0.5$ are used for this simulation study.

From Table 4.2, it is evident that the latent variable approach using the logit link to calculate MF yields promising results. All the 95% CI for the true coverage rate contains the true coverage rate (0.95) based on the 10,000 simulations.

4.4 Model comparison and model selection.

Based on the simulation study for clustered data, the probit and logit functions give promising results to estimate the mitigated fraction. In order to select the best link function to estimate the mitigated fraction, a simulation study was conducted. The best link function will be selected based on the percentage of bias-to-true value of the mitigated fraction.

The simulation consists of the following steps:

1. Using the probabilities, i.e. $\pi'_{it1}, \pi'_{it2}, \pi'_{it3}, \pi'_{it4}, \pi'_{it5}$, as described in Section 4.3, for vaccine and control, independent samples containing 20 observations each were generated for the two treatment groups respectively (i.e. 20 observations for control group and 20 observations for vaccine group).
2. Step 1 was repeated and a total of 10,000 simulated data sets were generated for each parameter setting for a given link function. For each data set, the mitigated fraction was found using the latent variable approach, that is 10,000 estimated mitigated fraction values were found using the latent variable approach.
3. The mean value of the 10,000 estimated mitigated fraction values was found for each link function.
4. The true value of the mitigated fraction was found for each link function using the method discussed in Section 4.2.
5. The bias was found for each link function as follows.

$$Bias = \ mean\ value\ of\ the\ 10{,}000\ estimated\ mitigated\ fraction\ values \\ - true\ value\ of\ the\ migtaged\ fraction.$$

The simulation results for all link functions are presented in Table 4.3.

Table 4.3. Percentage of Bias -True value of mitigated fraction comparison

Link	Mean Estimated MF	True Value	Bias	Bias-to-true value percentage
probit	0.25099	0.236504	0.014487	6.13%
logit (Monte Carlo simulation)	0.164088	0.154322	0.009766	6.33%

4.5 Real data analysis

This example is again from Poon (2004) where five randomized vaccine trials on administering the anti-cholinesterase drug tacrine to patients with Alzheimer's disease were considered. Patients receiving tacrine (T) and placebo (P) are, respectively, classified into five categories: 'very much improved' (c_0), 'minimally improved' (c_1), 'no change' (c_2), 'minimally worse' (c_3) and 'very much worse' (c_4). Here 'c_0' is the clinically best response and 'c_4' is the worst (order of increasing disease severity) and the trials are treated as blocks/clusters. The data is given in Table 4.4 and the results are presented in Table 4.5. Whitehead et al. (2001) and Poon (2004) have suggested that in each trial, there is a tendency for patients receiving the treatment 'Tacrine' to have better responses than those receiving placebo.

As discussed in the Section 4.2, the mitigated fraction can be estimated without using the latent variable. The analysis results are presented in the Table 4.6.

Table 4.4. The data for the clustered Tacrine study

	Trial 1		**Trial 2**		**Trial 3**		**Trial 4**		**Trial 5**	
Severity	P	T	P	T	P	T	P	T	P	T
$\boldsymbol{c_0}$	2	4	1	14	7	13	8	21	2	3
$\boldsymbol{c_1}$	22	23	22	119	16	20	24	106	13	14
$\boldsymbol{c_2}$	54	45	35	180	17	24	73	175	18	19
$\boldsymbol{c_3}$	29	22	11	54	10	10	52	62	7	3
$\boldsymbol{c_4}$	3	2	3	6	3	1	13	17	1	0
Total	110	96	72	373	53	68	170	381	41	39

Data source: Poon, W. (2004). A latent normal distribution model for analyzing ordinal responses with applications in meta-analysis. Statistics in medicine, 23:2155–2172.

Table 4.5. Results of the mitigated fraction for the clustered Tacrine study

Link	$\boldsymbol{MF = 2T - 1}$	**95% bootstrap Confidence interval**
probit	0.1610275	(0.0900, 0.2297)
Logit (Monte Carlo simulation)	0.1720648	(0.0958, 0.2464)

Table 4.6. Results of estimated values for the MF for the Tacrine study without latent variable

Link	$\boldsymbol{MF = 2T - 1}$	**95% bootstrap Confidence interval**
probit	0.1434804	(0.0808, 0.2032)
Logit	0.1461774	(0.0824, 0.2074)

Chapter 5 - Bayesian approach for mitigated fraction

In this chapter, a Bayesian statistical approach will be discussed to estimate the mitigated fraction from ordinal response variables. A Bayesian approach seems more natural in medical diagnosis studies and pharmaceutical industries. Compared to frequentist methods, the Bayesian approach is rather straightforward in model parameter estimation and results interpretation, especially for models with random effects and non-normal responses (Generalized linear mixed models). Agresti et al. (2001) stated that the Bayesian approach is advantageous specifically for clustered ordinal responses since the frequentist approach appears to suffer with cluster-specific and population-average models when it comes to the generalized linear mixed models. The parameter estimation through the Bayesian approach is computationally advantageous as the estimation directly coming from posterior samples when compared to the maximum likelihood estimation which leads to computational awkwardness for high-dimensional maximization.

Albert and Chib (1993) presented a Bayesian approach for binary and ordinal response data models using the data augmentation approach via an underlying latent continuous variable. Albert and Chib (1993) briefly mentioned possible extensions from binary data to ordered categorical data. They proposed an ordinal probit model in a Bayesian framework using the Gibbs sampling technique for generating posterior samples for parameters. Cowles (1996) discussed a hybrid Gibbs/Metropolis–Hasting (MH) sampling scheme for accelerating MCMC convergence for models like the ordered probit model. Johnson et al. (1999) discussed a detailed Bayesian framework for modeling ordinal data using the algorithm from Cowles (1996).

5.1 Some notations related to Bayesian analysis

In this chapter of the Bayesian analysis, most of terms and notation follow closely to those found in Gelman, Carlin, Stern, Dunson, Vehtari, and Rubin (2013). Consistent with Gelman et al. (2013), the terms 'distribution' and 'density' are interchangeably used in this chapter. The same notation will be used for continuous density functions and discrete probability mass functions.

Let $p(\cdot \mid \cdot)$ be a function which denotes a conditional probability density with the arguments determined by the context; $p(\cdot)$ denotes a marginal probability distribution. To avoid confusion, in this chapter the notation $Pr(\cdot)$ will be used for the probability of an event whereas the standard notation $P(\cdot)$ has been used for the probability of an event in all other chapters Gelman et al. (2013).

Bayes' rule

With reference to Gelman et al. (2013), the joint probability density (mass) function of θ and y can be written as a product of two density (mass) functions that are often referred to as the prior distribution $p(\theta)$ and the sampling distribution (or data distribution) $p(y|\theta)$, respectively. Simply conditioning on the known value of the data y, using the Bayes' rule of conditional probability, yields the posterior density $p(\theta|y)$ will be:

$$p(\theta|y) = \frac{p(y,\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)}, \tag{5.1}$$

where $p(y) = \int p(\theta)p(y|\theta)d\theta$ (or $p(y) = \sum_{\theta} p(\theta)p(y|\theta)$ in case of discrete θ, and the sum is over all possible values of θ). The factor $p(y)$ in the expression (5.1) does not depend on θ and, with fixed y, can thus be considered a constant, yielding the unnormalized posterior density, which is the right side of the expression (5.2):

$$p(\theta|y) \propto p(\theta)p(y|\theta). \tag{5.2}$$

The second term in the expression (5.2), $p(y|\theta)$, is considered as a function of θ, not y (Gelman et al., 2013).

Likelihood

Likelihood contributes significantly to the posterior inferences. Following the Bayes' rule, the data y has the contribution to the posterior inference (5.2) only through $p(y|\theta)$, which, when regarded as a function of θ, for fixed y, is called the likelihood function (Gelman et al., 2013).

5.2 The Gibbs Sampler

The Gibbs sampler is a Markov chain Monte Carlo (MCMC) algorithm. The Gibbs sampler is used as a computational technique in the calculation of marginal distribution of the parameters of interest, given a set of conditional distributions. In the Bayesian context, it is used to obtain marginal posterior distribution of the parameters of interest.

Let $\theta = (\theta_1, \theta_2, \dots, \theta_d)$ be the vector of posterior distribution parameters of interest. One is interested in generating samples from the posterior distribution $p(\theta|y)$. The primary goal is to determine the marginal distributions, $p(\theta_1), p(\theta_2), \dots, p(\theta_d)$ (Chipman and Hamada, 1996). When the full conditional distributions $p\big(\theta_k|\{\theta_j, j \neq k, j = 1,2, \dots, d\}\big)$ are available, then the Gibbs sampler can be used to generate sample from $p(\theta|y)$.

To generate sample from $p(\theta|y)$, the Gibbs sampler can be implemented as follows:

Set the initial values $\theta^{(0)} = \left(\theta_1^{(0)}, \theta_2^{(0)}, \dots, \theta_d^{(0)}\right)$.

Repeat the following procedure until convergence for $r = 1, 2, . \dots$

Generate $\theta_1^{(r)}$ from $p\left(\theta_1 |\, \theta_2^{(r-1)}, \dots, \theta_d^{(r-1)}, y\right)$

Generate $\theta_2^{(r)}$ from $p\left(\theta_2 |\, \theta_1^{(r-1)}, \theta_3^{(r-1)}, \dots, \theta_d^{(r-1)}, y\right)$

.

.

.

Generate $\theta_d^{(r)}$ from $p\left(\theta_d |\, \theta_1^{(r-1)}, \theta_2^{(r-1)}, \dots, \theta_{d-1}^{(r-1)}, y\right)$.

Following Geman and Geman (1984) and Chipman et al. (1996), when r approaches infinity, the joint distribution of the random variables $\theta^{(r)} = \left(\theta_1^{(r)}, \theta_2^{(r)}, \dots, \theta_d^{(r)}\right)$ converges in distribution to a sample from the joint distribution. Consequently, any subset of the random variables, $\theta = (\theta_1, \theta_2, \dots, \theta_d)$, can be viewed as a sample from the appropriate marginal distribution, i.e. $p(\theta_1), p(\theta_2), \dots, p(\theta_d)$ (Chipman and Hamada, 1996).

Following Chipman et al. (1996), there are several different strategies for implementing the Gibbs sampler for obtaining posterior samples of parameters. All the different strategies are centralized to obtain a representative sample of the posterior. One of the important factors is the number of iterations required to remove the effect of starting values of the parameters of interest. The number of iterations required to remove the effect of starting values are called **burn-in time** (Chipman et al., 1996). Generally, the burn-in time is quite large in such a way that it is most efficient to use many values taken from a single long run of the Gibbs sampler. The length of the burn-in time can be determined by comparing several chains using different starting values and checking the convergency of the simulated sequence. Several other issues of obtaining a representative sample can be found in Chipman et al. (1996).

5.3 Bayesian probit model

First, a Bayesian probit model is considered to estimate the mitigated fraction. The implementation of this Bayesian probit model for ordinal data follows closely to those found in Albert et al. (1993) and Chipman et al. (1996). Let $\boldsymbol{Y} = (Y_{1,}, \dots, Y_n)$ denote the observed vector of the responses for all individuals, $\boldsymbol{X_i'} = (x_{i1}, x_{i2}, \dots, x_{ip})$ is a p-dimensional vector of explanatory variables, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)'$ is a p-dimensional vector of parameters.

The likelihood function for $\boldsymbol{Y}$ given $\boldsymbol{\beta}, \boldsymbol{\gamma}$ will be

$$L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{i=1}^{n} \sum_{j=1}^{J} I(Y_i = j) \left[\Phi\left(\gamma_j - \boldsymbol{X_i'\beta}\right) - \Phi\left(\gamma_{j-1} - \boldsymbol{X_i'\beta}\right)\right],$$

where Φ is the standard norm cdf, and $I(\cdot)$ indicates an indicator function.
The unknown parameters are the regression vector $\boldsymbol{\beta}$ and the vector of cutoffs $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{J-1})$.

As discussed in Chapter 2, under a probit model, a continuous latent variable Y_i^* satisfies an ordinary regression model $Y_i^* = \boldsymbol{X_i'\beta} + \epsilon_i$ where ϵ_i follows a standard normal cdf. The latent normal continuous variable Y_i^* is distributed as $N(\boldsymbol{X_i'\beta}, 1)$. If the model is parameterized using the latent data $\boldsymbol{Y}^* = (Y_1^*, \dots, Y_n^*)$ along with the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, then the likelihood, as a function of this entire set of unknown parameters and latent data is expressed as

$$L(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*) = \prod_{i=1}^{n} \Phi(Y_i^* - \boldsymbol{X_i'\beta}) * \left\{\sum_{j=1}^{J} I(Y_i = j)\, I(\gamma_{j-1} < Y_i^* \leq \gamma_j)\right\}.$$

For the rest of this chapter, the vector notations will used with $\boldsymbol{Y}, \boldsymbol{Y}^*, \boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ as vectors of lengths n, n, p, and $J - 1$, respectively, and $\boldsymbol{X} = (\boldsymbol{X}_1', \boldsymbol{X}_2', \dots, \boldsymbol{X}_n')'$ denoting a full rank $n \times p$ design matrix.

Independent priors can be assumed for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. The joint prior distribution can be written as $p\ (\boldsymbol{\beta}, \boldsymbol{\gamma}) = p\ (\boldsymbol{\gamma})\ p\ (\boldsymbol{\beta})$. The priors can be vague (flat) priors for both $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Also, normal priors can be used for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ as follows: a prior for $\boldsymbol{\beta}$ will be given as $\boldsymbol{\beta} \sim N\left(\boldsymbol{0}, \Sigma_\beta\right)$ and for $\boldsymbol{\gamma}$ will be given as $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal and $\boldsymbol{\gamma}$ is restricted so that $\gamma_1 < \gamma_2 <, \dots, \gamma_{J-1}$. Typically, $\Sigma_{\boldsymbol{\beta}}$ also diagonal, with $\Sigma_{\boldsymbol{\beta}} = \sigma_{\boldsymbol{\beta}}^2 I$.

The joint posterior density of $\boldsymbol{\beta}, \boldsymbol{\gamma},$ and $\mathbf{Y}^*$ will be written as follows:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^* | \boldsymbol{Y}) = C\ p(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*) p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*)$$

where C is a proportionality constant.

Following the expression (5.2), the joint posterior density is given, up to a proportionality constant, by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^* | \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*) p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*)$$

Using the fact $p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*) = p(\boldsymbol{Y}^* | \boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\gamma})$,

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^* | \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*) p(\boldsymbol{Y}^* | \boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\gamma}).$$

For any given $\boldsymbol{Y}^*$**,** $\boldsymbol{Y}$ is independent of $\boldsymbol{\beta}$**,** and hence $p(\boldsymbol{Y} | \boldsymbol{\beta}, \gamma, \boldsymbol{Y}^*) = p(\boldsymbol{Y} | \boldsymbol{\gamma}, \boldsymbol{Y}^*)$,

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^* | \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\gamma}, \boldsymbol{Y}^*) p(\boldsymbol{Y}^* | \boldsymbol{\beta}, \boldsymbol{\gamma}) p(\boldsymbol{\beta}, \boldsymbol{\gamma})$$

The distribution of $\boldsymbol{Y}^*$ given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ does not depend on $\boldsymbol{\gamma}$, and hence $p(\boldsymbol{Y}^* | \boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{Y}^* | \boldsymbol{\beta})$. This is follows that the only dependence of $\boldsymbol{Y}^*$ on $\boldsymbol{\gamma}$ is through $\boldsymbol{Y}$.

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^* | \boldsymbol{Y}) \propto p(\boldsymbol{Y} | \boldsymbol{\gamma}, \boldsymbol{Y}^*) p(\boldsymbol{Y}^* | \boldsymbol{\beta}) p(\boldsymbol{\beta}, \boldsymbol{\gamma}).$$

Having assumed independent priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, $p\ (\boldsymbol{\beta}, \boldsymbol{\gamma}) = p\ (\boldsymbol{\gamma}) p\ (\boldsymbol{\beta})$.

The joint posterior density of $\boldsymbol{\beta}, \boldsymbol{\gamma},$ and $\boldsymbol{Y}^*$ given by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*|\boldsymbol{Y}) \propto\ p(\boldsymbol{Y}|\boldsymbol{\gamma}, \boldsymbol{Y}^*)p(\boldsymbol{Y}^*|\boldsymbol{\beta})p\ (\boldsymbol{\gamma})\ p\ (\boldsymbol{\beta})$$

Here the form of $p(\boldsymbol{Y}^*|\boldsymbol{\beta})$ will be determined by the link function. Under a probit link function,

$$p(\boldsymbol{Y}^*|\boldsymbol{\beta}) = N(\boldsymbol{X\beta}, \boldsymbol{I}).$$

Now the joint posterior distribution under the probit model becomes:

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*|\boldsymbol{Y}) \propto\ p\ (\boldsymbol{\gamma})\ p\ (\boldsymbol{\beta}) \prod_{i=1}^{n} \Phi(Y_i^* - \boldsymbol{X_i'\beta}) * \left\{\sum_{j=1}^{J} I(Y_i = j)\, I(\gamma_{j-1} < Y_i^*\ \leq \gamma_j)\right\}$$

with the form of ϕ, i.e. the standard normal cdf, and then joint posterior distribution under the probit model will be

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*|\boldsymbol{Y}) \propto\ p(\boldsymbol{\beta})p(\boldsymbol{\gamma}) \prod_{i=1}^{n} \left[\sqrt{1/2\pi}\ \exp\ (-(Y_i^* - \boldsymbol{X_i'\beta})^2/2)\right.$$

$$\left. * \left\{\sum_{j=1}^{J} I(Y_i = j)\, I\big(\gamma_{j-1} < Y_i^*\ \leq \gamma_j\big)\right\}\right].$$

Using flat (highly diffuse) priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, with $p(\boldsymbol{\beta}) \propto 1$ and $p(\boldsymbol{\gamma}) \propto 1$, the joint posterior density is given, up to a proportionality constant, by

$$p(\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{Y}^*|\boldsymbol{Y}) \propto \prod_{i=1}^{n} \left[\sqrt{\frac{1}{2\pi}}\ \exp\ (-\frac{(Y_i^* - \boldsymbol{X_i'\beta})^2}{2}) * \left\{\sum_{j=1}^{J} I(Y_i = j)\, I\big(\gamma_{j-1} < Y_i^*\ \leq \gamma_j\big)\right\}\right] \qquad (5.3)$$

Gibbs sampler will be used to estimate the parameters since full conditional distributions $p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y})$, $p(\boldsymbol{Y}^*|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Y})$ and $p(\boldsymbol{\gamma}|\boldsymbol{Y}^*, \boldsymbol{\beta}, \boldsymbol{Y})$ can be constructed.

The full conditional distributions of $\boldsymbol{\beta}$,$\boldsymbol{\gamma}$ and $\boldsymbol{Y}^*$ can be expressed as following.

1. The full conditional posterior distribution of $\boldsymbol{\beta}$ given on $\boldsymbol{\gamma}, \boldsymbol{Y}^*$ **and** $\boldsymbol{Y}$ is

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto\ p(\boldsymbol{\beta})p(\boldsymbol{Y}^*|\boldsymbol{\beta}).$$

With $p(\boldsymbol{\beta}) \propto 1$ and $p(\boldsymbol{Y}^*|\boldsymbol{\beta}) = N(\boldsymbol{X\beta}, \boldsymbol{I})$, this becomes

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto\ 1 * N(\boldsymbol{X\beta}, \boldsymbol{I}).$$

Here, the right hand side of the above expression needed to be modified in such a way that the distribution of $\boldsymbol{\beta}$ given $\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}$ will be in closed form.

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto \exp\left(\frac{-1}{2} \parallel \boldsymbol{Y}^* - \boldsymbol{X\beta} \parallel^2\right)$$

Let $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^*$, and now

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto \exp\left(\frac{-1}{2} \parallel \boldsymbol{Y}^* - \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X\beta} \parallel^2\right)$$

Since $\parallel \boldsymbol{Y}^* - \boldsymbol{X}'\widehat{\boldsymbol{\beta}} \parallel^2$ does not depend on $\boldsymbol{\beta}$**,** and having zero for cross of product of

$\left(\boldsymbol{Y}^* - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'$ and $(\boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X\beta})$**,** then

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto \exp\left(\frac{-1}{2} \parallel \boldsymbol{X}\widehat{\boldsymbol{\beta}} - \boldsymbol{X\beta} \parallel^2\right)$$

$$\propto \exp\left(\frac{-1}{2}\left[\left(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\right)'(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right]\right)$$

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma}, \boldsymbol{Y}^*, \boldsymbol{Y}) \propto MVN_p\left(\widehat{\boldsymbol{\beta}}, (\boldsymbol{X}'\boldsymbol{X})^{-1}\right)$$

where $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}^*$**.**

2. The full conditional posterior distribution of $\boldsymbol{Y}^*$ given on $\boldsymbol{\gamma}$, $\boldsymbol{\beta}$ and $\boldsymbol{Y}$ is

$$p(\boldsymbol{Y}^*|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Y}) \propto p(\boldsymbol{Y}^*|\boldsymbol{\beta})p(\boldsymbol{Y}|\boldsymbol{Y}^*, \boldsymbol{\gamma})$$

$$p(\boldsymbol{Y}^*|\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{Y}) \propto N(\boldsymbol{X\beta}, \boldsymbol{I}) * p(\boldsymbol{Y}|\boldsymbol{Y}^*, \boldsymbol{\gamma})$$

The full conditional posterior distribution of $Y_1^*, \dots, Y_n^*$ are independent with

$$p(Y_i^*|\boldsymbol{\gamma}, \boldsymbol{\beta}, Y_i = j) \propto \prod_{i=1}^{n} N(\boldsymbol{X_i'\beta}, 1)I(\gamma_{j-1} < Y_i^* \leq \gamma_j)$$

It is an independent truncated normal distribution with current values of the category cutoffs.

3. The posterior distribution of $\boldsymbol{\gamma}$ given on $\boldsymbol{Y}^*$, $\boldsymbol{\beta}$ and $\boldsymbol{Y}$ is

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}^*, \boldsymbol{\beta}, \boldsymbol{Y}) \propto f(\boldsymbol{\gamma})p(\boldsymbol{Y}|\boldsymbol{Y}^*, \boldsymbol{\gamma})$$

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}^{*},\boldsymbol{\beta},\boldsymbol{Y}) \propto \prod_{i=1}^{n}\left\{\sum_{j=1}^{J} I(Y_i = j)\, I(\gamma_{j-1} < Y_i^{*} \ \le \gamma_j)\right\}$$

It is to be noted that the above expression is a function of γ_j and the summation comes with only two components of γ_j as $1(Y_i = j)I\big(\gamma_{j-1} < Y_i^{*} \ \le \gamma_j\big)$ and $1(Y_i = j+1)I(\gamma_j < Y_i^{*} \ \le \gamma_{j+1})$. Thus, the full conditional posterior distribution of γ_j given $\boldsymbol{Y}^{*}$, $\boldsymbol{\beta}, \boldsymbol{Y}$ and $\{\gamma_k, k \ne j\}$ is given, up to a proportionality constant, by

$$p\big(\gamma_j|\boldsymbol{Y}^{*},\gamma_k,\boldsymbol{\beta},\boldsymbol{Y}\big) \propto \prod_{i=1}^{n}\big[1(Y_i = j)I\big(\gamma_{j-1} < Y_i^{*} \ \le \gamma_j\big) + 1(Y_i = j+1)I(\gamma_j < Y_i^{*} \ \le \gamma_{j+1})\big]$$

This conditional distribution can be seen to be uniform distribution on the interval,

$$\Big[max\Big\{\max_i\{Y_i^{*}\colon Y_i = j\}, \gamma_{j-1}\Big\}, min\Big\{\min_i\{Y_i^{*}\colon Y_i = j+1\}, \gamma_{j+1}\Big\}\Big]$$

5.4 Alternative priors for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ - Normal prior for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$

5.4.1 A normal prior for $\boldsymbol{\gamma}$

Alternatively, a normal prior for $\boldsymbol{\gamma}$ can be assumed as, $\boldsymbol{\gamma} \sim N(\boldsymbol{0},\boldsymbol{D})$, where $\boldsymbol{D}$ is a diagonal and $\boldsymbol{\gamma}$ is restricted so that $\gamma_1 < \gamma_2 <, \dots, \gamma_{J-1}$. Appropriate values will be chosen for the hyper parameters $\boldsymbol{D}$.

The full conditional posterior distribution of $\boldsymbol{\gamma}$ given $\boldsymbol{Y}^{*}$, $\boldsymbol{\beta},$ and $\boldsymbol{Y}$ is given, up to a proportionality constant, by

$$p(\boldsymbol{\gamma}|\boldsymbol{Y}^{*},\boldsymbol{\beta},\boldsymbol{Y}) \propto \ p(\boldsymbol{\gamma})p(\boldsymbol{Y}|\boldsymbol{Y}^{*},\boldsymbol{\gamma})$$

The full conditional posterior distribution of γ_j given $\boldsymbol{Y}^{*}$, $\boldsymbol{\beta}, \boldsymbol{Y}$ and $\{\gamma_k, k \ne j\}$ is given, up to a proportionality constant, by

$$p\left(\gamma_j|\boldsymbol{Y}^*,\gamma_k,\boldsymbol{\beta},\boldsymbol{Y}\right) \propto N\left(0,\sigma_{\gamma_j}^2\right)\prod_{i=1}^{n}\left[1(Y_i=j)I\left(\gamma_{j-1}<Y_i^* \le \gamma_j\right)+1(Y_i=j+1)I(\gamma_j < Y_i^* \le \gamma_{j+1})\right]$$

$$p\left(\gamma_j|\boldsymbol{Y}^*,\gamma_k,\boldsymbol{\beta},\boldsymbol{Y}\right) \propto N\left(0,\sigma_{\gamma_j}^2\right)\left[max\left\{\max_i\{Y_i^*:Y_i=j\},\gamma_{j-1}\right\},min\left\{\min_i\{Y_i^*:Y_i=j+1\},\gamma_{j+1}\right\}\right]$$

Therefor the full conditional posterior distribution of γ_j given $\boldsymbol{Y}^*$, $\boldsymbol{\beta}, \boldsymbol{Y}$ and $\{\gamma_k, k \neq j\}$ distributed as truncated normal distribution with appropriate interval $\left[max\left\{\underset{i}{max}\{Y_i^*:Y_i=j\},\gamma_{j-1}\right\},min\left\{\underset{i}{min}\{Y_i^*:Y_i=j+1\},\gamma_{j+1}\right\}\right]$.

5.4.2 A normal prior for $\boldsymbol{\beta}$

Alternatively, a normal prior for $\boldsymbol{\beta}$, $\boldsymbol{\beta}\sim N\left(\boldsymbol{0},\ \Sigma_\beta\right)$, typically, $\Sigma_{\boldsymbol{\beta}}$ also diagonal, with $\Sigma_{\boldsymbol{\beta}} = \sigma_{\boldsymbol{\beta}}^2 I$.

The full conditional posterior distribution of $\boldsymbol{\beta}$ given $\boldsymbol{Y}^*$, $\boldsymbol{\beta}, \boldsymbol{\gamma}$ and $\boldsymbol{Y}$ is given, up to a proportionality constant, by

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{Y}^*,\boldsymbol{Y}) \propto\ p(\boldsymbol{\beta})p(\boldsymbol{Y}^*|\boldsymbol{\beta}).$$

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{Y}^*,\boldsymbol{Y}) \propto\ N\left(\boldsymbol{0},\ \Sigma_\beta\right)N(\boldsymbol{X\beta},\ \boldsymbol{I}).$$

$$p(\boldsymbol{\beta}|\boldsymbol{\gamma},\boldsymbol{Y}^*,\boldsymbol{Y}) \propto\ MVN_p\left(\widehat{\boldsymbol{\beta}},\left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\boldsymbol{1}}+\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\right)$$

where $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-\boldsymbol{1}}+\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{Y}^*$**.**

5.5 Estimating the mitigated fraction for two independent groups using Bayesian estimation

As described in Chapter 3 for estimating the mitigated fraction for two independent groups, there is a single factor of treatment type with two levels (vaccine and control) as an explanatory variable. Let x_1 denote the explanatory variable as a group indicator for an observation, where $x_1 = 1$ for the control group and $x_1 = 0$ for the vaccine group. Also, let Y denote an independent ordinal random variable with J categories. The cumulative probability model will be:

$\pi_j = G(\eta_j), \eta_j = \gamma_j - \beta_1 x_1$ for $j = 1, \dots, J - 1$,

where η_j is the linear predictor**,** G is inverse of an arbitrary link function, and

$\pi_j = P(Y \leq j|x_1) = \tilde{\pi}_1 + \cdots + \tilde{\pi}_j$ with $\sum_{j=1}^{J} \tilde{\pi}_j = 1$ are cumulative probabilities.

For this analysis of estimating the mitigated fraction using the Bayesian approach, a non-informative normal prior for $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$**.** Since there is only one explanatory variable, $\boldsymbol{\beta} = (\beta_0, \beta_1)$. Since the inclusion of β_0 in the regression model, appropriate identification constraint has been set as $\gamma_1 = 0$ and $\sigma_\epsilon^2 = 1$. For each γ_j, a non-informative normal prior such that $\gamma_j \sim N(0,10^6), j = 2,3,4$ was used. For each β_k, a non-informative normal prior such that $\beta_k \sim N(0,10^6)$, $k = 1,2$ was used. Markov chain Monte Carlo (MCMC) methods of Gibbs sampling was implemented to make the inferences for $(\gamma_2, \gamma_3, \gamma_4, \beta_1)$ from the posterior distribution. Once the $\gamma_2, \gamma_3, \gamma_4, \boldsymbol{\beta}$ are found from Bayesian approach, estimated MF will be as:

- with latent variable: $MF = 2T - 1 = 2 * \Phi\left(\frac{\beta_1}{\sqrt{2}}\right) - 1$.

- without latent variable: $2\hat{T} - 1 = 2\sum\sum_{j>k} \hat{\pi}_{1j}\hat{\pi}_{2k} + \frac{1}{2}\sum_{j=k} \hat{\pi}_{1j}\hat{\pi}_{2k} - 1$.

where $\hat{\pi}_{1j}$ and $\hat{\pi}_{2k}$ - fitted values can be from the corresponding link function of cumulative probability model.

All of the Bayesian analyses were performed using chains of 50 000 samples, of which the first half (i.e. 25,000) was treated as burn-in time and were removed to find a representative samples for all parameters. Also, convergence of parameters has been assessed by plotting the simulated sequences of the parameters.

5.6 Bayesian approach for the Tacrine study

The Bayesian approach was applied to estimate the mitigated for the Tacrine study. The data is used from Table 5 in Chapter 3. Mitigated fraction was estimated using latent variable approach discussed in Section 5.4, and without using the latent variable discussed in Section 3.5. The analysis results are presented in the Table 5.1.

Table 5.1. Bayesian results of estimated $\boldsymbol{MF}$ for the Tacrine study

Link	$\boldsymbol{T = P(Y_1 > Y2)}$	$\boldsymbol{MF = 2T - 1}$
Probit (with latent variable)	0.55	0.1
Probit (without latent variable)	0.54	0.08

Chapter 6 - Conclusion and Discussion

This dissertation study may be of great interest to animal health companies and associated government agencies. This work is really filling the gap of estimation of the mitigated fraction for disease severity when the outcome variable is ordinal, especially when observations are clustered. This work focuses on developing methods to estimate the mitigated fraction for ordinal data with a parametric approach assuming a generalized linear mixed model (GLMM). Ordinal outcomes in terms of measuring disease severity are widely popular in the areas of veterinary medicines and epidemiological fields. Currently, the USDA's Center for Veterinary Biologics (CVB) recommends a form of the mitigated fraction which can be calculated when the disease severity is graded by some continuous measure or by some discrete assessment resulting in unambiguous ranks. Considering the growing interest in the measure of disease severity for ordinal outcomes, the development of methods to estimate the mitigated from ordinal data is of utmost importance. Two different approaches were discussed to estimate the mitigated fraction: one with the frequentist approach and another one with the Bayesian approach. Based on our simulation results in Chapters 3 and Chapter 4, our latent variable approach gives promising results to estimate the mitigated fraction for ordinal data under different link functions. Especially, the probit link gives the best option to estimate the mitigated fraction based on the bias-to-true-value comparison. Also, the probit link gives a straightforward (i.e. closed form) solution for estimating the mitigated fraction in all situations of CRD and RCBD.

In Chapter 3 and Chapter 4, methods were discussed to estimate the mitigated fraction for two independent groups in a CRD and RCBD setup, respectively. Our study focused on settings where an experiment is performed to estimate the mitigated fraction involving two treatments such as vaccine and placebo. Under our proposed method, an extension can easily be made to the cases where more than two independent treatment groups are simultaneously compared to a placebo using appropriate usage of dummy variables as group indicators. Another extension that can easily be made is the inclusion of multiple random effects which can simply be included in the linear predictor of the cumulative link model (4.1).

The proposed method using the latent variable approach can be used to estimate the mitigated fraction while adjusting for other explanatory variables such as age, gender, and weight of the experimental units. Since many experimental studies on vaccine trials have their interest in measuring the efficacy of vaccine based on explanatory variables, our proposed method is a powerful approach to estimate the mitigated fraction while accounting for other explanatory variables.

It is very important to note that under the latent variable construction, it was assumed that the model holds when an underlying continuous response has the usual regression model structure with a constant variance. In that case, the latent response at different explanatory variable values differs in terms of location but not dispersion of the latent variables. But, by using the latent variable approach, the dispersion of the latent variable still can be handled in terms of estimating the mitigated fraction. If the distribution of the latent response at different predictor values differs in terms of both location and dispersion, following McCuullagh (1980), a general version of the cumulative link model can be written as

$$G^{-1}[P(Y_i \leq j|\boldsymbol{X}_i)] = \frac{\alpha_j - \boldsymbol{\beta}'\boldsymbol{X}_i}{\exp\,(\gamma'\boldsymbol{X}_i)} \tag{6.1}$$

The denominator in the right-hand side of the model (6.1) contains the scale parameters γ that describes how the dispersion depends on $\boldsymbol{X}$**.** The latent variable for the above model (6.1) can be given as $Y_i^* = \boldsymbol{\beta}'\boldsymbol{X}_i + \exp\,(\gamma'\boldsymbol{X}_i)\epsilon_i$. In our ordinary cumulative link model used in this dissertation is a special case of the model (6.1) with $\gamma = 0$ (Agresti, 2010). Let us consider the situation from Chapter 3 where an experiment involving two treatments such as vaccine and placebo to estimate the mitigated fraction. For this situation, the model (6.1) reduces to $G^{-1}[P(Y \leq j|x_1)] = \frac{\alpha_j - \beta_1 x_1}{\exp(\gamma' x_1)}$, and the latent variable with dispersion effect γ for this situation can be written as $Y^* = \beta_1 x_1 + \exp(\gamma x_1)\,\epsilon$. Under the probit model, Y^* follows a normal distribution, i.e. $Y^* \sim N(\beta_1 x_1, \exp(\gamma x_1))$. Letting Y_1^* and Y_2^* are the two independent continues latent random variables such that Y_1^* for control group when $x_1 = 1$ and Y_2^* for vaccine group when $x_1 = 0$. Then Y_1^* and Y_2^* will be following a normal distribution as $Y_1^* \sim N(\beta_1, \exp(\gamma)^2)$**,** and $Y_2^* \sim N(0,1)$. And hence it immediately follows that the mitigated fraction will be equal to $2 * \Phi\left(\frac{\beta_1}{\sqrt{1+\exp(\gamma)^2}}\right) - 1$, i.e. $MF = 2 * \Phi\left(\frac{\beta_1}{\sqrt{1+\exp(\gamma)^2}}\right) - 1$.

Future work

One of the problems in vaccine studies is that usually most of the subjects do not become infected. In this situation, the resulting data contain many observations that are zero, and the outcomes are generally referred to as the zero-inflation problem. In this situation, a different approach will be recommended where it will allow cumulative models of ordinal data to model the distribution of ordinal outcomes more accurately when most of the subjects do not become infected. In the event of zero-inflation problem, a zero-inflated model needs to be considered to explicitly model the zero inflation, and to model the cumulative probabilities with the relationship of possible predictors.

Let Y_i denote the ordinal response for observation i, $i = 1,2,\dots,n$ with levels $0,1,2,\dots,J$. A multinomial distribution for Y_i will be

$$Y_i \sim Multinomial(1, \tilde{\pi}_0, \tilde{\pi}_1, \dots, \tilde{\pi}_J)$$

where $\tilde{\pi}_j$ is the probability of a response falls into the j^{th} category with $\sum_{i=1}^{J} \tilde{\pi}_j = 1$.

Let $\theta = \tilde{\pi}_0, \tilde{\pi}_1, \dots, \tilde{\pi}_J$, and a zero-inflated model describes a latent mixture of two populations,

$$p(Y_i|\theta, z_i) = z_i \mathbb{I}\{y_i = 0\} + (1 - z_i) Multinomial(1, \theta)$$

where $z_i = 0$ or 1 and $z_i|\omega \sim iid\ Bernolli(\omega)$ with ω as the population mixture parameter.

When it comes to estimating the mitigated fraction for clustered ordinal data, the marginal model is easily be tractable for the probit model and the logit model. However, it is to be noted that there is no clear way to get a marginal model when log-log or clog-log link is used for *cluster-specific* model with Gaussian distribution with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{D}$ for random effects $\boldsymbol{b_i}$**.** However, the clog-log link used for *cluster-specific* model with log-gamma distribution for random effects $\boldsymbol{b_i}$, the resulting marginal likelihood has a closed form (Thomas, 1966).

In this dissertation, the Bayesian approach was discussed only under the probit model. A comprehensive Bayesian approach needs to be implemented to estimate the mitigated fraction with other link functions such as logit, log-log, and clog-log. Under the probit model, Gibbs sampling was used to estimate the parameters from the posterior distribution since the full conditional distributions of the parameters were in closed form distributions. But under the logit, log-log, and clog-log link function, the full conditional distributions of the parameters will not be in closed form distributions and thus Gibbs sampling will not be able to apply. In this situation, the Metropolis-Hastings algorithm needs to be used to estimate the parameters.

References

Agresti, A. (2010). *Analysis of ordinal categorical data* (2nd ed.). Wiley Series in Probability and Statistics, New Jersey.

Agresti, A. (2013). *Categorical data analysis* (3rd ed.). Wiley Series in Probability and Statistics, New Jersey.

Agresti, A., & Kateri, M. (2017). Ordinal Probability Effect Measures for Group Comparisons in Multinomial Cumulative Link Models. *Biometrics* 73, 214–219.

Agresti, A, & Natarajan, A. (2001). Modeling Clustered Ordered Categorical Data: A Survey. *International Statistical Institute (ISI)*.

Albert, J. H. (1992). Bayesian estimation of the polychoric correlation coefficient. *Journal of Statistical Computation and Simulation*, Vol. 44, pp. 47- 61.

Albert, J. H., & Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, Vol. 88, No. 422, pp. 669- 679.

Anderson, J. A., & Philips, P. R. (1981). Regression, Discrimination and Measurement Models for Ordered Categorical Variables. *Journal of the Royal Statistical Society*. Series C (Applied Statistics), Vol. 30, No. 1, pp. 22-31.

Bross, I. D. J. (1958). How to use ridit analysis. *Biometrics*, 14, 18–38.

Camilli, G. (1994). Origin of the Scaling Constant in Item Response Theory. *Journal of Educational and Behavioral Statistics*.

CDC (Centers for Disease Control and Prevention). Lesson 3: Measures of Risk, Section 6: Measures of Public Health Impact https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section6.html (Last accessed 07/09/2021).

Chipman, H., & Hamada, M. (1996). Bayesian Analysis of Ordered Categorical Data from Industrial Experiments. *Technometrics*, Feb., Vol. 38, No. 1, pp. 1-10.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence Diagnostics: A Comparative Review. *Journal of the American Statistical Association*, Vol. 91, No. 434, pp. 883-904.

Cox, C. (1995). Location-scale cumulative odds models for ordinal data: a generalized non-linear model approach. *Statistics In Medicine*, Vol. 14, 1191-1203.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Chapman & Hall: New York.

Fleiss, J. L, Levin, B., & Paik, M. C. (2003). *Statistical Methods for Rates and Proportions* (3rd ed.). Wiley Series in Probability and Statistics.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis* (3rd ed.). Taylor & Francis Group, Florida.

Geman, S., & Geman, D. (1984). Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-6, NO. 6.

Haley, D.C. (1952). Estimation of the Dosage Mortality Relationship When the Dose Is Subject to Error Technical Report No. 15 (Office of Naval Research Contract No. 25140, NR-342-022). Applied Mathematics and Statistics Laboratory, Stanford University.

Halloran, M. V., Struchiner, C. J, & Longini, I. M. (1997). Study Designs for Evaluating Different Efficacy and Effectiveness Aspects of Vaccines. *American Journal of Epidemiology*. Volume 146, Number 10.

Halloran, M. E., Struchiner, C. J., & Longini, I. M. (2009). *Design and Analysis of Vaccine Studies*. Springer, New York.

Jackman, S. (2000). Models for Ordered Outcomes**.** *Political Science* 200C.

Johnson, V. E., & Albert, H. (1999). *Ordinal Data Modeling* (Statistics for Social and Public Policy). Springer, New York.

Johnson, N.J., & Kotz, S. (1970). *Continuous Univariate Distributions-2*. Houghton Mifflin, Boston.

Kruskal, W. H. (1957). Historical Notes on the Wilcoxon Unpaired Two-Sample Test. *Journal of the American Statistical Association*.

McCullagh, P. (1980). Regression Models for Ordinal Data. *Journal of the Royal Statistical Society*. Series B (Methodological).

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *In Frontiers in Econometrics,* P. Zarembka (ed), 105–142. New York: Academic Press.

McKelvey, R. D., & and Zavoina. (1975). A statistical model for the analysis of ordinal level dependent variables. *Journal of mathematical sociology*.

Mehrotra, D. V. (2006). Vaccine clinical trials: A statistical primer. *Journal of Biopharmaceutical Statistics,* 16: 403–414, 2006.

Milliken, A., & Johnson, D. E. (2009). *Analysis of Messy Data* (2nd ed.). Volume 1: Designed Experiments. Taylor & Francis Group, Florida.

Polack, F. P., Thomas, S. J., Kitchin, N., Absalon, J., Gurtman, A., Lockhart, S., Perez, J. L., Marc, J. P., Moreira, E. D., Zerbini, C., Bailey, R., Swanson, K. A., Roychoudhury, S., Koury, K., Li, P., Kalina, W. V., Cooper, D., Frenck, R. W., Hammitt, L. L., Tureci, O., Nell, H., Schaefer, A., Unal, S., Tresnan, D. B., Mather, S., Dormitzer, P. R., Sahin, U., Jansen, K. U., & Gruber, W. C. (2020). Safety and Efficacy of the BNT162b2 mRNA Covid-19 Vaccine. *The new England journal of medicine*.

Poon, W. (2004). A latent normal distribution model for analyzing ordinal responses with applications in meta-analysis. *Statistics in medicine*, 23:2155–2172.

Rosenthal, J.S. (2006). *At rigorous probability theory*. World Scientific Publishing Co. Pte. Ltd. New Jersey.

Savalei, V. (2006). Logistic Approximation to the Normal: The KL Rationale. *Psychometrika*.

Selvin, S. (1977). A further note on the interpretation of ridit analysis. *America Journal of Epidemiology*.

Shao, X., Ma, X., Chen, F., Song, M., Pan, X., & You, K. (2020). A random parameter ordered probit analysis of injury severity in truck involved rear-end collisions. *International Journal of Environmental Research and Public Health*.

Siev. D. (2005). An Estimator of Intervention Effect on Disease Severity. *Journal of Modern Applied Statistical Methods*, Vol. 4: Iss. 2, Article 14.DOI: 10.22237/jmasm/1130803980.

Thomas, R. T. H. (1966). A Mixed Effects Model for Multivariate Ordinal Response Data Including Correlated Discrete Failure Times with Ordinal Responses. Biometrics, Vol. 52, No. 2, pp. 473-491

Tugwell, B. D., Lee, L. E., Gillette, H., Lorber, E. M., Hedberg, K., & Cieslak, P. R. (2004). Chickenpox outbreak in a highly vaccinated school population. *Pediatrics*. 114 (4) 1130-1131; DOI: https://doi.org/10.1542/peds.2004-1110.

Vargha, A., & Delaney, H. D. (1998). The Kruskal-Wallis test and stochastic homogeneity. *Journal of Educational and Behavioral Statistics*.

Whitehead, A., Omar, R. Z., Higgins, J. P. T., Savaluny, E., Turner, R. M., & Thompson, S. G. (2001). Meta-analysis of ordinal outcomes using individual patient data. *Statistics In Medicine*: 20:2243–2260. Doi: 10.1002/Sim.919.

Wikipedia. http://en.wikipedia.org/wiki/Gumbel_distribution (Last accessed 07/09/2021).

Wolfe, D. A., & Hogg, A. R. V. (1971). On Constructing Statistics and Reporting Data. *The American Statistician*, Vol. 25, No. 4. https://www.jstor.org/stable/2682922

Zeger, S.L., Liang, K.-Y., & Albert, P.S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, Vol. 44, No. 4, pp. 1049-1060.

Appendix A - R code

The following R-codes are some of the selected R codes used in the simulation study.

```
library(ordinal)

library(bootstrap)

library(boot)

library(MASS)

nu_trt=2

b1=0.1725

size = 100

alpha=c(-1.5,-0.3,1,2,Inf)

alpha1=rep(alpha,nu_trt)

c=5

mlog_odds=rep(0,nu_trt*c)

treatment = rep( c(rep("trt", times = c), rep("con", times = c)))

dat = data.frame(treatment)

#View(dat)

for(j in 1:(nu_trt*c)){

  mlog_odds[j] =  (alpha1[j]-b1*(dat$treatment[j] == "con"))

}

mcumproprobit=pnorm(mlog_odds)

#cumproplogit = plogis(log_odds)

mcell_pro=mcumproprobit

for(j in 2:(nu_trt*c)){
  mcell_pro[j]=mcumproprobit[j]-mcumproprobit[j-1]
}
mcell_pro[seq(1,nu_trt*c,by=c)]=mcumproprobit[seq(1,nu_trt*c,by=c)]
dat$mcumpro=mcumproprobit
dat$mcelpro=mcell_pro
#View(dat)
#MF true value calculation
alpha=0
#p=c(0.02,0.28,0.1,0.2,0.4)
#v=c(0.3,0.25,0.15,0.25,0.05)
p=dat$mcelpro[6:10]
v=dat$mcelpro[1:5]
for(j in 1:length(p)){
  for(k in 1:length(v)){
    if(j==k){
      alpha=alpha+0.5*p[j]*v[k]
    }
    else if(j>k){
      alpha=alpha+p[j]*v[k]
    }
  }
}
```

```
true=2*alpha-1
#print(alpha)
#print(mf)
#true=0.081439394
count.probit=0
count.logit=0
count.loglog=0
count.sim=0
simN=1
simCI.probit=matrix(0,simN,2)
simCI.logit=matrix(0,simN,2)
simCI.loglog=matrix(0,simN,2)
simCI.cloglog=matrix(0,simN,2)
average_MF_over_5000.probit=rep(0,simN)
average_MF_over_5000.logit=rep(0,simN)
average_MF_over_5000.loglog=rep(0,simN)
average_MF_over_5000.cloglog=rep(0,simN)
log_odds=rep(0,nu_trt*c)
for(k in 1:simN){
  ###probit model confidence interval
  for(j in 1:(nu_trt*c)){
    log_odds[j] =  alpha1[j]-b1*(dat$treatment[j] == "con")
  }
```

```
 cumproprobit=pnorm(log_odds)
 #cumproplogit = plogis(log_odds)
 cell_pro=cumproprobit
 for(j in 2:(nu_trt*c)){
   cell_pro[j]=cumproprobit[j]-cumproprobit[j-1]
 }
 cell_pro[seq(1,nu_trt*c,by=c)]=cumproprobit[seq(1,nu_trt*c,by=c)]
 dat$cumpro=cumproprobit
 dat$celpro=cell_pro
 #View(dat)
fi=seq(1,(nu_trt*c-c+1), by=c)
 y=rep(0,nu_trt*c)
 for(i in fi ){
   #rmultinom(1, size = 110, prob = dat$celpro[i:(i+4)])
   #checkcell[i:(i+4)]=dat$celpro[i:(i+4)]
   y[i:(i+4)]= rmultinom(1, size = size, prob = dat$celpro[i:(i+4)])
 }
 dat$y=y
 #View(dat)
 #write.csv(dat, "ordinalsimulation2.csv", col.names = TRUE, row.names = FALSE)
 #size = 20#
 #ya=rep(0,size*nu_trt)
 trt_t=rep(0,size*nu_trt)
 #m=c(1,11)

 xt=1

 #i=11

   severity <- c(1,2,3,4,5)

   trt=dat$y[1:c]

   con=dat$y[(c+1):(c*nu_trt)]

   yb <- factor(c(rep(severity, con),

                 rep(severity, trt)), levels = severity)

   trt_t[xt:(xt+(size-1))]=1

   trt_t[((xt+(size-1))+1):(((xt+(size-1))+1)+(size-1))]=0

 newdata=data.frame(trt_t,yb)

 #View(newdata)

 y=factor(newdata$yb)

 trt=newdata$trt_t

# blk=newdata$blk

 ## Cumulative link mixed model with two random terms:

 #probit.m <- clmm(y ~ trt + (1|blk),  link = "probit",threshold = "equidistant")

 probit.mm <- clm(y ~ trt,  link = "probit")

 summary(probit.mm)

 betap=tail(coef(probit.mm),n=1)

 alpha=pnorm(betap/sqrt(2),0,1)

 MF=2*alpha-1

 print(MF)
```

```
      ###probit model confidence interval
      #library(bootstrap)
      #library(boot)
      theta.hat=function(d,i)
      {
        yn<<-factor(d$yb[i])
        xn<<-d$trt_t[i]
        probit.m <- clm(yn ~ xn ,  link = "probit")
        betap=tail(coef(probit.m),n=1)
        MF=2*pnorm(betap/sqrt(2),0,1)-1
      }
      oboot=boot(data=newdata,statistic=theta.hat, R=5000)
      BCa.probit=boot.ci(oboot,conf=.95,type="bca")
      simCI.probit[k,]=BCa.probit$bca[4:5]
      if(true>=simCI.probit[k,1] && true<=simCI.probit[k,2]){
        count.probit=count.probit+1}
      average_MF_over_5000.probit[k]=oboot$t0
    count.sim=count.sim+1
    }
print(count.probit)
```

```
#for logit

        library(ordinal)

        library(bootstrap)

        library(boot)

        library(MASS)

        nu_trt=2

        b1=0.2842

        size = 100

        alpha=c(-1.5,-0.3,1,2,Inf)

        alpha1=rep(alpha,nu_trt)

        c=5

        mlog_odds=rep(0,nu_trt*c)

        treatment = rep( c(rep("trt", times = c), rep("con", times = c)))

        dat = data.frame(treatment)

        #View(dat)

        for(j in 1:(nu_trt*c)){

          mlog_odds[j] =  (alpha1[j]-b1*(dat$treatment[j] == "con"))

        }

        mcumproprobit=plogis(mlog_odds)

        #cumproplogit = plogis(log_odds)

        mcell_pro=mcumproprobit

        for(j in 2:(nu_trt*c)){

          mcell_pro[j]=mcumproprobit[j]-mcumproprobit[j-1]
        }

        mcell_pro[seq(1,nu_trt*c,by=c)]=mcumproprobit[seq(1,nu_trt*c,by=c)]

        dat$mcumpro=mcumproprobit

        dat$mcelpro=mcell_pro

        #View(dat)

        #MF true value calculation

        alpha=0

        #p=c(0.02,0.28,0.1,0.2,0.4)

        #v=c(0.3,0.25,0.15,0.25,0.05)

        p=dat$mcelpro[6:10]

        v=dat$mcelpro[1:5]

        for(j in 1:length(p)){

          for(k in 1:length(v)){

            if(j==k){

              alpha=alpha+0.5*p[j]*v[k]

            }

            else if(j>k){

              alpha=alpha+p[j]*v[k]

            }

          }

        }

        true=2*alpha-1

        print(alpha)
```

```
#print(mf)

#true=0.081439394

count.probit=0

count.logit=0

count.loglog=0

count.sim=0

simN=1

simCI.probit=matrix(0,simN,2)

simCI.logit=matrix(0,simN,2)

simCI.loglog=matrix(0,simN,2)

simCI.cloglog=matrix(0,simN,2)

average_MF_over_5000.probit=rep(0,simN)

average_MF_over_5000.logit=rep(0,simN)

average_MF_over_5000.loglog=rep(0,simN)

average_MF_over_5000.cloglog=rep(0,simN)

log_odds=rep(0,nu_trt*c)

for(k in 1:simN){

  ###probit model confidence interval

  for(j in 1:(nu_trt*c)){

    log_odds[j] =  alpha1[j]-b1*(dat$treatment[j] == "con")

  }

  cumproprobit=plogis(log_odds)

  #cumproplogit = plogis(log_odds)
```

```
cell_pro=cumproprobit

for(j in 2:(nu_trt*c)){

  cell_pro[j]=cumproprobit[j]-cumproprobit[j-1]

}

cell_pro[seq(1,nu_trt*c,by=c)]=cumproprobit[seq(1,nu_trt*c,by=c)]

dat$cumpro=cumproprobit

dat$celpro=cell_pro

#View(dat)

fi=seq(1,(nu_trt*c-c+1), by=c)

y=rep(0,nu_trt*c)

for(i in fi ){

  #rmultinom(1, size = 110, prob = dat$celpro[i:(i+4)])

  #checkcell[i:(i+4)]=dat$celpro[i:(i+4)]

  y[i:(i+4)]= rmultinom(1, size = size, prob = dat$celpro[i:(i+4)])

}

dat$y=y

#View(dat)

#write.csv(dat, "ordinalsimulation2.csv", col.names = TRUE, row.names = FALSE)

#size = 20#

#ya=rep(0,size*nu_trt)

trt_t=rep(0,size*nu_trt)

#m=c(1,11)

xt=1
```

```
#i=11

severity <- c(1,2,3,4,5)

trt=dat$y[1:c]

con=dat$y[(c+1):(c*nu_trt)]

yb <- factor(c(rep(severity, con),

            rep(severity, trt)), levels = severity)


trt_t[xt:(xt+(size-1))]=1

trt_t[((xt+(size-1))+1):(((xt+(size-1))+1)+(size-1))]=0

newdata=data.frame(trt_t,yb)

#View(newdata)

y=factor(newdata$yb)

trt=newdata$trt_t

# blk=newdata$blk

## Cumulative link mixed model with two random terms:

#probit.m <- clmm(y ~ trt + (1|blk),  link = "probit",threshold = "equidistant")

probit.mm <- clm(y ~ trt,  link = "logit")

#summary(probit.mm)

#n1=10000

#b_log=tail(coef(probit.mm),n=1)

# y1=rlogis(n1, location = b_log, scale = 1)

# y2=rlogis(n1, location = 0, scale = 1)

#2*mean(y1>y2)-1
```

```
# betap=tail(coef(probit.mm),n=1)
 #alpha=pnorm(betap/sqrt(2),0,1)
 #MF=2*alpha-1
 #print(MF)
 ###probit model confidence interval
 #library(bootstrap)
 #library(boot)
 theta.hat=function(d,i)
 {
   yn<<-factor(d$yb[i])
   xn<<-d$trt_t[i]
   probit.m <- clm(yn ~ xn ,  link = "logit")
   betap=tail(coef(probit.m),n=1)
   n1=10000
   y1=rlogis(n1, location = betap, scale = 1)
   y2=rlogis(n1, location = 0, scale = 1)
   MF=2*mean(y1>y2)-1
 }
 oboot=boot(data=newdata,statistic=theta.hat, R=5000)
 BCa.probit=boot.ci(oboot,conf=.95,type="bca")
 simCI.probit[k,]=BCa.probit$bca[4:5]
 if(true>=simCI.probit[k,1] && true<=simCI.probit[k,2]){
   count.probit=count.probit+1}
```

```
                average_MF_over_5000.probit[k]=oboot$t0
                count.sim=count.sim+1
              }
print(count.probit)


#Bias probit
              library(ordinal)
              library(bootstrap)
              library(boot)
              library(MASS)
              nu_trt=2
              b1=0.1725
              size = 100
              alpha=c(-1.5,-0.3,1,2,Inf)
              alpha1=rep(alpha,nu_trt)
              c=5
              mlog_odds=rep(0,nu_trt*c)
              treatment = rep( c(rep("trt", times = c), rep("con", times = c)))
              dat = data.frame(treatment)
              #View(dat)
              for(j in 1:(nu_trt*c)){
                mlog_odds[j] =  (alpha1[j]-b1*(dat$treatment[j] == "con"))
              }
```

```
mcumproprobit=pnorm(mlog_odds)

#cumproplogit = plogis(log_odds)

mcell_pro=mcumproprobit

for(j in 2:(nu_trt*c)){

  mcell_pro[j]=mcumproprobit[j]-mcumproprobit[j-1]

}

mcell_pro[seq(1,nu_trt*c,by=c)]=mcumproprobit[seq(1,nu_trt*c,by=c)]

dat$mcumpro=mcumproprobit

dat$mcelpro=mcell_pro

#View(dat)

#MF true value calculation

alpha=0

#p=c(0.02,0.28,0.1,0.2,0.4)

#v=c(0.3,0.25,0.15,0.25,0.05)

p=dat$mcelpro[6:10]

v=dat$mcelpro[1:5]

for(j in 1:length(p)){

  for(k in 1:length(v)){

    if(j==k){

      alpha=alpha+0.5*p[j]*v[k]

    }

    else if(j>k){

      alpha=alpha+p[j]*v[k]
    }
  }
}
true=2*alpha-1
#print(alpha)
#print(mf)
#true=0.081439394
count.probit=0
count.logit=0
count.loglog=0
count.sim=0
simN=10000
simCI.probit=matrix(0,simN,2)
simCI.logit=matrix(0,simN,2)
simCI.loglog=matrix(0,simN,2)
simCI.cloglog=matrix(0,simN,2)
average_MF_over_5000.probit=rep(0,simN)
average_MF_over_5000.logit=rep(0,simN)
average_MF_over_5000.loglog=rep(0,simN)
average_MF_over_5000.cloglog=rep(0,simN)
log_odds=rep(0,nu_trt*c)
for(k in 1:simN){
  ###probit model confidence interval
```

```
for(j in 1:(nu_trt*c)){
  log_odds[j] =  alpha1[j]-b1*(dat$treatment[j] == "con")
}
cumproprobit=pnorm(log_odds)
#cumproplogit = plogis(log_odds)
cell_pro=cumproprobit
for(j in 2:(nu_trt*c)){
  cell_pro[j]=cumproprobit[j]-cumproprobit[j-1]
}
cell_pro[seq(1,nu_trt*c,by=c)]=cumproprobit[seq(1,nu_trt*c,by=c)]
dat$cumpro=cumproprobit
dat$celpro=cell_pro
#View(dat)
fi=seq(1,(nu_trt*c-c+1), by=c)
y=rep(0,nu_trt*c)
for(i in fi ){
  #rmultinom(1, size = 110, prob = dat$celpro[i:(i+4)])
  #checkcell[i:(i+4)]=dat$celpro[i:(i+4)]
  y[i:(i+4)]= rmultinom(1, size = size, prob = dat$celpro[i:(i+4)])
}
dat$y=y
#View(dat)
#write.csv(dat, "ordinalsimulation2.csv", col.names = TRUE, row.names = FALSE)
```

```
#size = 20#

#ya=rep(0,size*nu_trt)

trt_t=rep(0,size*nu_trt)

#m=c(1,11)

xt=1

#i=11

severity <- c(1,2,3,4,5)

trt=dat$y[1:c]

con=dat$y[(c+1):(c*nu_trt)]

yb <- factor(c(rep(severity, con),

             rep(severity, trt)), levels = severity)


trt_t[xt:(xt+(size-1))]=1

trt_t[((xt+(size-1))+1):(((xt+(size-1))+1)+(size-1))]=0

newdata=data.frame(trt_t,yb)

#View(newdata)

y=factor(newdata$yb)

trt=newdata$trt_t

# blk=newdata$blk

## Cumulative link mixed model with two random terms:

#probit.m <- clmm(y ~ trt + (1|blk),  link = "probit",threshold = "equidistant")

probit.mm <- clm(y ~ trt,  link = "probit")

summary(probit.mm)
```

```
          betap=tail(coef(probit.mm),n=1)
          alpha=pnorm(betap/sqrt(2),0,1)
          MF=2*alpha-1
          average_MF_over_5000.probit[k]=MF
}

#Bias Logit
        library(ordinal)
        library(bootstrap)
        library(boot)
        library(MASS)
        nu_trt=2
        b1=0.2842
        size = 100
        alpha=c(-1.5,-0.3,1,2,Inf)
        alpha1=rep(alpha,nu_trt)
        c=5
        mlog_odds=rep(0,nu_trt*c)
        treatment = rep( c(rep("trt", times = c), rep("con", times = c)))
        dat = data.frame(treatment)
        #View(dat)
        for(j in 1:(nu_trt*c)){
          mlog_odds[j] =  (alpha1[j]-b1*(dat$treatment[j] == "con"))
```

```
}
mcumproprobit=plogis(mlog_odds)
#cumproplogit = plogis(log_odds)
mcell_pro=mcumproprobit
for(j in 2:(nu_trt*c)){
  mcell_pro[j]=mcumproprobit[j]-mcumproprobit[j-1]
}
mcell_pro[seq(1,nu_trt*c,by=c)]=mcumproprobit[seq(1,nu_trt*c,by=c)]
dat$mcumpro=mcumproprobit
dat$mcelpro=mcell_pro
#View(dat)
#MF true value calculation
alpha=0
#p=c(0.02,0.28,0.1,0.2,0.4)
#v=c(0.3,0.25,0.15,0.25,0.05)
p=dat$mcelpro[6:10]
v=dat$mcelpro[1:5]
for(j in 1:length(p)){
  for(k in 1:length(v)){
    if(j==k){
      alpha=alpha+0.5*p[j]*v[k]
    }
    else if(j>k){
      alpha=alpha+p[j]*v[k]
    }
  }
}
true=2*alpha-1
print(true)
#print(mf)
#true=0.081439394
simN=10000
average_MF_over_5000.logit=rep(0,simN)
log_odds=rep(0,nu_trt*c)
for(k in 1:simN){
  ###probit model confidence interval
  for(j in 1:(nu_trt*c)){
    log_odds[j] =  alpha1[j]-b1*(dat$treatment[j] == "con")
  }
  cumproprobit=plogis(log_odds)
  #cumproplogit = plogis(log_odds)
  cell_pro=cumproprobit
  for(j in 2:(nu_trt*c)){
    cell_pro[j]=cumproprobit[j]-cumproprobit[j-1]
  }
  cell_pro[seq(1,nu_trt*c,by=c)]=cumproprobit[seq(1,nu_trt*c,by=c)]
```

```
dat$cumpro=cumproprobit
dat$celpro=cell_pro
#View(dat)
fi=seq(1,(nu_trt*c-c+1), by=c)
y=rep(0,nu_trt*c)
for(i in fi ){
  #rmultinom(1, size = 110, prob = dat$celpro[i:(i+4)])
  #checkcell[i:(i+4)]=dat$celpro[i:(i+4)]
  y[i:(i+4)]= rmultinom(1, size = size, prob = dat$celpro[i:(i+4)])
}
dat$y=y
#View(dat)
#write.csv(dat, "ordinalsimulation2.csv", col.names = TRUE, row.names = FALSE)
#size = 20#
#ya=rep(0,size*nu_trt)
trt_t=rep(0,size*nu_trt)
#m=c(1,11)
xt=1
#i=11
severity <- c(1,2,3,4,5)
trt=dat$y[1:c]
con=dat$y[(c+1):(c*nu_trt)]
yb <- factor(c(rep(severity, con),
            rep(severity, trt)), levels = severity)
trt_t[xt:(xt+(size-1))]=1
trt_t[((xt+(size-1))+1):(((xt+(size-1))+1)+(size-1))]=0
newdata=data.frame(trt_t,yb)
#View(newdata)
y=factor(newdata$yb)
trt=newdata$trt_t
# blk=newdata$blk
## Cumulative link mixed model with two random terms:
#probit.m <- clmm(y ~ trt + (1|blk),  link = "probit",threshold = "equidistant")
probit.m <- clm(y ~ trt,  link = "logit")
betap=tail(coef(probit.m),n=1)
n1=10000
y1=rlogis(n1, location = betap, scale = 1)
y2=rlogis(n1, location = 0, scale = 1)
MF=2*mean(y1>y2)-1
average_MF_over_5000.logit[k]=MF}
```