8-2024

An Application of an In-Depth Advanced Statistical Analysis in Exploring the Dynamics of Depression, Sleep Deprivation, and Self-Esteem

Muslihat Gaffari
East Tennessee State University

Recommended Citation

Gaffari, Muslihat, "An Application of an In-Depth Advanced Statistical Analysis in Exploring the Dynamics of Depression, Sleep Deprivation, and Self-Esteem" (2024). *Electronic Theses and Dissertations.* Paper 4428. https://dc.etsu.edu/etd/4428

An Application of an In-Depth Advanced Statistical Analysis in Exploring the Dynamics of Depression, Sleep Deprivation, and Self-Esteem

---

A thesis

presented to

the faculty of the Department of Mathematics and Statistics

East Tennessee State University

In partial fulfillment

of the requirements for the degree

Master of Science in Mathematical Sciences

---

by

Muslihat Adejoke Gaffari

August 2024

Mostafa Zahed, Ph.D., Chair

Robert Price, Ph.D.

Michele Joyner, Ph.D.

Keywords: Depression, Sleep Deprivation, Self-Esteem, Mental Health, Log-Linear Model, Multinomial Logistic Regression Model, Generalized Linear Model.

ABSTRACT

An Application of an In-Depth Advanced Statistical Analysis in Exploring the Dynamics of Depression, Sleep Deprivation, and Self-Esteem

by

Muslihat Adejoke Gaffari

Depression, intertwined with sleep deprivation and self-esteem, presents a significant challenge to mental health worldwide. The research shown in this paper employs advanced statistical methodologies to unravel the complex interactions among these factors. The study scrutinizes large datasets through log-linear homogeneous association, multinomial logistic regression, and generalized linear models to uncover nuanced patterns and relationships. By elucidating how depression, sleep disturbances, and self-esteem intersect, the research aims to deepen understanding of mental health phenomena. It clarifies the relationship between these variables and explores reasons for prioritizing depression research. It evaluates how statistical models, such as log-linear, multinomial logistic regression, and generalized linear models, shed light on their intricate dynamics. Findings offer insights into risk and protective factors associated with these variables, guiding tailored interventions for individuals in psychological distress. Additionally, policymakers can utilize these insights to develop comprehensive strategies promoting mental health and well-being at a societal level.

DEDICATION

This research work is solely dedicated to Almighty Allah, the sustainer and the cherisher of the universe, who in HIS infinite mercy, allowed me to complete this thesis efficaciously and to attain this zenith of success.

ACKNOWLEDGMENTS

All praises, glorification, and adoration are due to Almighty Allah, the cherisher and the sustainer of the universe, who has been my constant companion and unfailing source of inspiration, who in His infinite mercy guided me throughout my academic pursuit in East Tennessee State University.

I want to express my sincere appreciation to my supervisor, Dr. Mostafa Zahed, for his invaluable guidance and support throughout this research. Your beneficial criticism and treasured contributions have been instrumental in the success of this thesis. I am truly grateful for your unwavering commitment and dedication to my academic growth and achievement.

I sincerely appreciate the esteemed committee members, Dr. Robert Price and Dr. Michele Joyner, for their insightful contributions and support of this research.

I am greatly indebted to the Department of Mathematics and Statistics at East Tennessee State University for their financial support, which enabled me to complete my graduate studies and thesis successfully.

My profound gratitude goes to my support system and God-sent, Anifowose Abdul Afeez Olanrenwaju, for his ceaseless prayers, love, care, support, and encouragement throughout my academic years.

I am greatly obligated to my family for their care, support, encouragement, and sacrifices to ensure my academic success.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

1.1 Background of the Study

Sadness is a main part of the human condition. Feeling down and a little blue is sometimes a temporary experience connected to specific events. A sense of sadness or hopelessness can be steadier, which is well-known as depression. Depression is a serious problem that can affect people's lives. What they think and feel will affect their ability to sleep [1].

Depression, also known as Major Depressive Disorder (MDD) or clinical depression, is a common and severe mental health condition characterized by persistent and pervasive feelings of sadness, hopelessness, and a lack of interest or pleasure in most activities. It goes beyond the regular ups and downs that people experience in life. Depression can affect a person's thoughts, feelings, and physical well-being, often leading to a range of emotional and physical symptoms [2].

There are different types of depression, some of which develop due to specific circumstances, which include symptoms of depressed mood or loss of interest, most of the time for at least two weeks. For example, *Persistent Depressive Disorder*, also called Dysthymia or Dysthymic disorder, consists of less severe symptoms than other Major Depressive Disorders. Still, it is longer lasting, usually up to 2 years. *Perinatal Depression* occurs during or after pregnancy. *Seasonal Affective Disorder* comes and goes with the seasons. *Depression with Symptoms of Psychosis* is a severe form of depression in which a person experiences psychosis symptoms such as delusions or hallucinations [2].

There are some common causes of depression, such as a brain chemistry imbalance, drug abuse, stress, female sex hormones, and physical health problems [3]. Depression symptoms can include suicidal thoughts or attempts, trouble concentrating, a loss of interest in things that previously brought joy, insomnia, and irritability [4]. Depression also affects the body physically. For instance, an increased risk of heart disease, a weakened immune system, impaired digestion, and the nervous system, a lowered libido, and a lowered tolerance for pain are possible signs of depression [5].

The degrees of depression can vary from moderate to severe. Whether it is mild or severe, it can be treated with either medication, psychotherapy, or a combination of the two. Antidepressants are used to treat depression, which works by changing how the brain produces or processes certain chemicals involved in mood or stress. Psychotherapy or talk therapy is sometimes used alone for the treatment of mild depression. For moderate to severe depression, psychotherapy is often implored along with antidepressant medications. If medication or psychotherapy does not reduce the symptoms of depression, brain stimulation therapies can be tried. There are different types of brain stimulation therapies, such as Electro-Convulsive Therapy (ECT), Repetitive Transcranial Magnetic Stimulation (RTMS), Vagus Nerve Stimulation (VNS), Magnetic Seizure Therapy (MST), and Deep Brain Stimulation (DBS). ECT and RTMS are the most widely used brain stimulation therapies [2].

This research examines two potential causes of depression, which are trouble sleeping and feeling bad about oneself. Trouble sleeping, also known as insomnia, is most often described as a subjective complaint of poor sleep quality or quantity despite adequate time for sleep, resulting in daytime fatigue, irritability, and decreased concentration. Insomnia is classified as idiopathic or comorbid. Comorbid insomnias are associated with psychiatric disorders, medical disorders, substance abuse, and specific sleep disorders. Idiopathic insomnia is essentially a diagnosis of exclusion. There are some causes of insomnia, such as poor sleep habits, stress, and anxiety [6].

On the other hand, sleeping too much or hypersomnia is the inability to stay awake and alert during the day despite having more than an adequate amount of nighttime sleep [7]. The treatment of sleep disorders (insomnia or hypersomnia) varies from moderate to severe. Medications, lifestyle changes, and addressing underlying medical conditions can be used to treat mild sleep disorders. At the same time, Cognitive Behavioral Therapy (CBT) is a therapy used to treat severe sleep disorders [8].

Feeling bad about oneself, commonly called low self-esteem or poor self-image, is a psychological state characterized by negative perceptions and beliefs about one's worth, abilities, and overall value. This state can manifest in various ways and profoundly impact an individual's emotions, thoughts, behaviors, and relationships. Early life experiences such as negative criticism and abuse or bullying, traumatic events such as loss, failure, or rejection, social comparison, and interpersonal relationships can cause low self-esteem. Low self-esteem can be treated by having healthy relationships, setting realistic goals, engaging in positive affirmations, seeking support, and practicing self-compassion [9].

Depression, sleep deprivation, and self-esteem are integral components of human experience, each exerting a profound influence on mental well-being and quality of life. These psychological constructs, individually and collectively, play pivotal roles in shaping individuals' emotional states, cognitive functioning, and interpersonal relationships. Understanding the complex interplay among these variables is essential for elucidating the mechanisms underlying mental health disorders and developing targeted interventions to mitigate their impact.

The relationship between trouble sleeping, depression, and feeling bad about yourself is complicated. Some people find they cannot sleep at all, while others find they cannot stop sleeping. Sleep problems and depression may also be innate biological factors that cause people to feel bad about themselves. Sleep problems are also associated with more severe depressive illness [10].

The primary purpose of this research is to clarify the relationship between depression, trouble sleeping or sleeping too much, and feeling bad about yourself. We will address that by answering the following questions:

RQ 1: What are the critical reasons for prioritizing the study of depression, and how does a deeper understanding of this mental health condition contribute to improved prevention, intervention, and overall well-being in individuals and society?

RQ 2: How can Log-Linear Models, Multinomial Logistic Regression, and Generalized Linear Models(GLM) be employed to analyze the association between depression, sleep disturbances, and self-esteem, shedding light on the intricate relationships within mental health?

RQ 3: How can the performance of proposed statistical models be effectively compared using statistical measures such as the Likelihood ratio test, Pearson chi-square, Mean Square Error, Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC)?

1.2 Significance of the Study

Investigating the relationship between depression, sleep deprivation, and self-esteem has important implications for mental health interventions and public policy. These interconnected aspects worsen symptoms and lower well-being, emphasizing the importance of comprehending their intricate relationships. To lessen psychological distress, specific therapies can be created by identifying changeable risk and protective variables. Furthermore, the knowledge gained from this study helps shape evidence-based policies supporting mental health and fair access to care. Promoting social justice and well-being in society requires addressing the socioeconomic determinants of mental health inequities, particularly among vulnerable communities. This study contributes to the field's understanding and helps develop solutions for better mental health outcomes.

1.3 Terminologies

For easy understanding and readability, these are terms used in this research:

- MDD - Major Depressive Disorder
- ECT - ElectroConvulsive Therapy
- RTMS - Repetitive Transcranial Magnetic Stimulation
- VNS - Vagus Nerve Stimulation
- MST - Magnetic Seizure Therapy
- DBS - Deep Brain Stimulation

- GLM - Generalized Linear Models
- AIC - Akaike Information Criterion
- BIC - Bayesian Information Criterion
- MSE - Mean Square Error
- SLE - Systemic Lupus Erythematosus
- MEC - Mobile Examination Center
- CMH - Cochran-Mantel-Haenszel
- GEE - Generalized Estimation Equations
- MLE - Maximum Likelihood Estimation
- NCHS - National Center for Health Statistics
- MEC - Mobile Examination Center
- NHANES - National Health and Nutrition Survey

2 LITERATURE REVIEW

Depression, sleep disturbances, and self-esteem are interconnected aspects of mental health that significantly impact individuals' well-being. Understanding the complex relationships among these factors is crucial for developing effective interventions and promoting mental health. In this literature review, we synthesize findings from various studies to explore the bidirectional relationships between depression, sleep disturbances, and self-esteem. The review encompasses research spanning different populations and methodologies, providing insights into the nuanced dynamics of these psychological phenomena.

A systematic review of the relationship between sleep disturbances and mental health assessed the bidirectionality between sleep disturbances, anxiety, and depression, indicating a one-way relationship where anxiety predicted excessive daytime sleepiness, highlighting the intricate interplay between sleep disturbances and mental health problems [11]. Postpartum sleep disturbance and postpartum depression revealed a strong relationship between the two among predominantly educated, middle-class, older, white participants [12].

The study of depression is carried out on specific populations, such as the quality of sleep and depression in college students, shedding light on the unique challenges faced by a demographic group that highlighted the prevalence of depression among college students and its association with sleep quality, emphasizing the importance of addressing mental health issues in educational settings and the effect of anxiety and depression on sleep quality in individuals at high risk for insomnia revealed the complex relationship between psychiatric comorbidity and sleep disturbance [13, 14].

The meta-analysis examination of the relationship between sleep disorders and suicidal behavior in patients with depression indicated a significant association between sleep disorders and various manifestations of suicidal behavior and estimated more specific categories, including insomnia, nightmares, hypersomnia, suicidal ideation, suicide attempt, and completed suicide underscoring the importance of addressing sleep disturbances in suicide prevention efforts and highlighting the critical role of sleep in the context of depression and suicidal thoughts [15].

The investigation of the relationship between sex hormones, sleep problems, and depression emphasized the complex interplay between biological factors, sleep disturbances, and depression, which provides insights into potential mechanisms underlying these relationships and the examination of the relationship between depression and sleep quality, diseases, and general characteristics elucidated the multifaceted nature of depression and its associations with various demographic and clinical factors [16, 18].

The association between depression and sleep quality in patients with Systemic Lupus Erythematosus (SLE) reveals the significant impact of chronic illness on mental health outcomes and highlights the need for targeted interventions addressing both physical and psychological aspects of chronic conditions [17].

In conclusion, the findings provide valuable insights into the complex dynamics of depression, sleep disturbances, and self-esteem, underscoring the importance of holistic approaches to mental health care guided by advanced statistical methodologies to tailor interventions and promote well-being across diverse populations.

3 RESEARCH AND METHODOLOGY

This chapter examines the methodologies aimed at addressing the purpose of this research. These included methods and models such as the Log-linear, Multinomial Logistic Regression, and Generalized Linear Models used to analyze this research.

3.1 Categorical Variables

Categorical variables, also called qualitative variables, place individuals into one of several groups. Categorical variables have measurement scales consisting of a set of categories. For instance, political philosophy is often measured as liberal, moderate, or conservative. Diagnoses regarding breast cancer are based on a mammogram in the categories normal, benign, probably benign, suspicious, and malignant [28]. In the data used for this research, mental health is measured as “little interest in doing things”, “depression,” “trouble sleeping”, “feeling tired,” “overeating,” and “feeling bad about yourself.”

There are three types of categorical variables: nominal, ordinal, and interval. The nominal categorical variable makes classifications without order, but the ordinal categorical variable creates classifications with an order that possibly varies between groups. The interval categorical variable makes classifications with order and equal distances between groups. For example, education is a nominal categorical variable when measured as learning in a private or public school; It becomes an ordinal categorical variable when measured by the highest degree attained using categories such as high school, bachelor’s, master’s, and doctorate. It is an interval categorical variable when measured by the years required to attain a specific level of education, using the integers $0, 1, 2, \cdots$ [28].

3.1.1 Distributions of Categorical Data

In this section, three key distribution categories will be reviewed.

- Bernoulli or Binomial Distribution

  The Bernoulli distribution represents a single trial with a fixed total and unknown time that has two responses, a success or a failure (0 or 1). The distribution has a simple form of a random variable:

  $$Z = \begin{cases} \pi & \text{if } z = 1 \text{ (success)} \\ \\ 1 - \pi & \text{if } z = 0 \text{ (failure)} \end{cases}$$

  Suppose π is the true probability of success. Then, the Bernoulli probability mass function is

  $$P(Z = z) = \pi^z (1 - \pi)^{(1-z)}, \tag{1}$$

  where π is the probability of success and $1 - \pi$ is the probability of failure [20].

  The binomial distribution involves n independent and identical trials such that each trial can result in one of the two possible outcomes: success or failure. This distribution is often used to estimate or determine the proportion of individuals with a particular attribute in a large population. If π is the probability of observing success in each trial, then the random variable Y, which is the number of successes, can be observed from these n trials. The random variable Y is defined as follows:

  $$Y = \sum_{i=1}^{n} Z_i \tag{2}$$

Then, the probability of observing y successes out of these n trials is given by the Binomial probability mass function:

$$P(Y = y) = \binom{n}{y} \pi^y (1 - \pi)^{(n-y)}, \tag{3}$$

where n is a fixed number of independent trials [21].

- Multinomial Distribution is a type of distribution with a fixed total and unknown time in addition to multiple categories and responses. It models the probability of selecting one of k mutually exclusive categories. This is also called a multivariate distribution. The distribution is a generalization of the Binomial distribution ($k = 2$) to many dimensions where, instead of two groups, the N elements are divided into k groups, each with a probability π_i ranging from 1 to k. The Multinomial probability mass function is given by:

$$P(Y_1 = y_1, \ldots, Y_k = y_k) = \frac{N!}{y_1! \ldots y_k!} \pi_1^{y_1} \ldots \pi_k^{y_k}, \tag{4}$$

where Y_i represents the random variable of the count for each category and π_i is the probability of an individual trial resulting in category i [20].

- Poisson Distribution is a type of distribution with a fixed time and unknown total. This distribution gives the probability of observing y events in a given time, assuming that events occur independently at a constant rate. Let Y denote the number of events in a unit interval of time or a unit distance. Then, Y is called the Poisson random variable with a mean number of events λ in a unit interval. The Poisson probability density function with mean λ is given by:

$$P(Y = y) = \frac{e^{-\lambda} \lambda^y}{y!}, y = 0, 1, 2, ... \tag{5}$$

where Y is the random variable representing the count and λ is the mean or the rate.

3.2 Descriptive Table Calculations

We will consider four main types of descriptive table calculations in this research. For simplicity, we will start with 2 × 2 tables:

Table 1: 2 × 2 tables

Success	Failure
π_{11}	π_{12}
π_{21}	π_{22}

where π_{11} is the probability of success for the first row and first column, π_{21} is the probability of success for the second row and first column, π_{12} is the probability of failure for the first row and second column, and π_{21} is the probability of failure of the second row and second column.

With only success or failure as options, we often focus only on the success probability for each row:

π_{11} is written as $\pi_{(1|1)}$ or simply as π_1.

π_{21} is written as $\pi_{(1|2)}$ or simply as π_2.

3.2.1 Relative Risk (RR)

In the Test of Proportions, we often compare success probabilities for two samples. The statistics

$$\pi_{(1|1)} - \pi_{(1|2)} = \pi_1 - \pi_2 \tag{6}$$

is called the difference in proportions. Of course, a ratio better accounts for the relative magnitudes of success probabilities.

The formula for relative risk is given by:

$$\frac{\pi_{(1|1)}}{\pi_{(1|2)}} = \frac{\pi_1}{\pi_2}, \tag{7}$$

where π_1 is the probability of success for the first row and π_2 is the probability of success of the second row.

Therefore, relative risk is the ratio of success probabilities for two rows. With more than two rows, we have multiple relative risks comparing all rows pairwise.

3.2.2 Properties of Relative Risk (RR)

(i) $RR \in [0, \infty)$. This means relative risk encompasses all possible scenarios, from no risk to infinitely higher risk.

(ii) If $RR = 1$, then rows are independent or homogeneous. This indicates that the risk of the event occurring is the same in both groups and suggests that the exposure or treatment being studied does not affect the risk of the event.

(iii) If $RR > 1$, then the first group is at greater risk than the second group. This indicates that the likelihood of occurrence is increased in comparison to the other group by exposure or characteristics associated with the First Group.

(iv) If $RR < 1$, then the first group is at lower risk than the second group. This suggests that, compared to the second group, the exposure or characteristic associated with the first group decreases the likelihood of occurrence of the event.

(v) While the difference in proportions will be the same regardless of how "success" is labeled, the relative risk will not. The illustration is shown below:

$$\pi_1 - \pi_2 = (1 - \pi_1) - (1 - \pi_2)$$

$$\frac{\pi_1}{\pi_2} \neq \frac{(1 - \pi_1)}{(1 - \pi_2)}$$

In other words, this mean that while the difference in proportions remains constant, the relative risk changes when the labeling of "success" changes due to the different effect on the numerator and denominator of the relative risk ratio [28].

3.2.3 Odds(Ω)

In general, for success probability π, the odds denoted as Ω are defined as the ratio of success probability to failure probability.

The formula for odds is given by:

$$\Omega = \frac{\pi}{1 - \pi}, \tag{8}$$

where Ω is the odds and π is the success probability.

We noticed that while relative risk compares success probabilities across groups, odds compare probabilities for the same group.

3.2.4 Properties of Odds(Ω)

(i) $\Omega \in [0, \infty)$. A value of 0 means that the event is impossible, while higher values indicate increasing likelihood.

(ii) $\Omega = 1$ means that success and failure are equally likely. When the odds are exactly 1, it signifies that the event is equally likely to occur as it is not to occur.

(iii) $\Omega > 1$ means that success is more likely than failure. It indicates that the event is more likely to happen than not to happen.

(iv) $\Omega < 1$ means that success is less likely than failure. It suggests that the event is less likely to happen than not to happen.

(v) Switching the designation of "success" inverts Ω. This property arises because switching the labels of success and failure also switches their probabilities. Therefore, if the probability of success was initially reduced compared to failure, it would be more likely after a switch. This deviation ensures that the relative probability of the identified events is accurately reflected in the odds ratio, regardless of their designation [28].

3.2.5 Odds Ratio(θ)

It is often interesting to compare odds across rows. It is the combination of odds and relative risk. Therefore, the odds ratio is the ratio of the odds of success in one group to the odds of success in another group.

The formula for odds ratio is given by:

$$\theta = \frac{\Omega_1}{\Omega_2}, \tag{9}$$

where θ is the odds ratio and Ω is the odds.

The odds ratio is written in many ways. For example, the odds for a 2 by 2 table is given by:

$$\theta = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}, \tag{10}$$

where π_1 is the success probability of the first row and π_2 is the success probability of second row.

The odds ratio can also be written in terms of cell values as follows:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \tag{11}$$

3.2.6 Properties of Odds Ratio(θ)

(i) $\theta \in [0, \infty)$. As the odds ratio approaches ∞, it suggests increasingly higher odds of success in the first group compared to the second group.

(ii) $\theta = 1$ means subjects in both rows are equally likely to have success (independent or homogeneous).

(iii) $\theta > 1$ means the odds of success are greater for row 1 than row 2. Subjects of row 1 are more likely to have success than those of row 2.

(iv) $\theta < 1$ means the odds of success are lower for row 1 than row 2. Subjects of row 1 are less likely to have success than those of row 2.

(v) Transposing rows (designation of success) and columns or groups does not change the value of θ. This property makes the odds ratio preferable over relative risk in certain scenarios because it remains consistent regardless of how the data is arranged or labeled.

3.2.7 Log-odds ($\ln(\theta)$)

All the previous descriptive live on skewed distributions. Applying a logarithm can reduce this skewness. The log odds, also known as the natural logarithm of the odds ratio, is a transformation of the odds ratio (θ).

The formula of the log-odds is given as:

$$\ln(\theta) = \ln\left(\frac{\Omega_1}{\Omega_2}\right), \tag{12}$$

where θ is the odd ratio, Ω_1 is the odds of row 1 and Ω_2 is the odds of row 2.

3.2.8 Properties of Log-Odds ($\ln(\theta)$)

(i) $\ln(\theta) \in (-\infty, \infty)$. If the log odds are negative, the odds ratio is less than 1, suggesting lower odds of success. If the log odds are positive, the odds ratio is greater than 1, indicating higher odds of success.

(ii) A log-odds of 0 equals an odds ratio less than 1. This means the odds of success are equal to the odds of failure.

(iii) A log-odds less than 0 equals an odds ratio less than 1. This means the odds of success are lower than the odds of failure.

(iv) A log-odds greater than 0 equals an odds ratio greater than 1. This means the odds of success are higher than the odds of failure.

(v) Log-odds does not have an intuitive interpretation. It is typically used in making inferences about odds ratios [28].

3.3 Contingency Table

A contingency table, also known as a cross-tabulation or a two-way table, is a statistical table used to display the frequency distribution of two or more categorical variables. Each cell in the table represents the frequency count or percentage of cases that fall into a particular combination of categories for the studied variables. The primary purpose of cross-tabulation is to uncover patterns, associations, and dependencies between categorical variables in a dataset [24].

Considering the case of two categorical variables X and Y, X with I categories and Y with J categories, let n_{ij} represent the count of the number of responses that fall into level i of X and level j of Y. This gives rise to a table of the following form:

Table 2: Contingency Table

	1	2	$\cdots$	$\cdots$	$\cdots$	J	
1	n_{11}	n_{12}	$\cdots$	$\cdots$	$\cdots$	n_{1J}	n_{1+}
2	n_{21}	n_{22}	$\cdots$	$\cdots$	$\cdots$	n_{2J}	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$			$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$		$\ddots$		$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$			$\ddots$	$\vdots$	$\vdots$
I	n_{I1}	n_{I2}	$\cdots$	$\cdots$	$\cdots$	n_{IJ}	n_{I+}
	n_{+1}	n_{+2}	$\cdots$	$\cdots$	$\cdots$	n_{+J}	n_{++}

where I is the number of rows or number of levels of X, J is the number of columns or number of levels of Y, n_{ij} is the count of cell i, j or number of responses in level i of X and level j of Y, n_{i+} is the sum over row i or number of responses in level i of X, n_{+j} is the sum over column j or number of responses in level j of Y, and n_{++} is the total number of counts [24].

It was introduced by Karl Pearson in 1904. Classifications of subjects on both variables have IJ possible combinations. A contingency table with I rows and J columns is called an $I \times J$ table [28].

Considering two cases of calculations in a contingency table with more than two levels of variables and more than two variables, it is important to be specific about what is being compared in either of these cases.

3.3.1 More than two levels of variables

To compare across populations, both relative risk and odds ratios are used because if there are more than two populations, there is more than one possible comparison to make. For example, the table below shows the cross-classification between college graduation (graduated or not graduated) and three populations. We recorded 50 white individuals, 50 black or African-American individuals, and 50 Hispanic or Latino individuals.

Table 3: Table of Graduation Status by Race

	Graduated	Not Graduated
White	38	12
Black or African American	32	18
Hispanic or Latino	29	21

The estimation of the odds of graduation for each row is calculated as follows:

- $\hat{\Omega}_1 = \frac{38}{12} \approx 3.17$
- $\hat{\Omega}_2 = \frac{32}{18} \approx 1.78$
- $\hat{\Omega}_3 = \frac{29}{21} \approx 1.38$

The estimation of odds ratios of graduation is calculated as follows:

- $\theta_{1,2} = \frac{\Omega_1}{\Omega_2} = \frac{3.17}{1.78} \approx 1.78$
- $\theta_{1,3} = \frac{\Omega_1}{\Omega_3} = \frac{3.17}{1.38} \approx 2.30$
- $\theta_{2,3} = \frac{\Omega_2}{\Omega_3} = \frac{1.78}{1.38} \approx 1.29$

Three distinct odds ratios can be calculated as follows:

(i) The odds of graduation for Whites are $\frac{3.17}{1.78} = 1.78$ times that of African Americans.

(ii) The odds of graduation for Whites are $\frac{3.17}{1.38} = 2.30$ times that of Hispanics or Latinos.

(iii) The odds of graduation for African Americans are $\frac{1.78}{1.38} = 1.29$ times that of Hispanics or Latinos.

The following example will consider more than two options for both variables by comparing the three races with high school, some college, and college graduation.

Table 4: Table of Education Level by Race

	High School	Some College	College Graduation
White	15	20	15
Black or African American	25	15	10
Hispanic or Latino	30	15	5

In each row, three distinct odds can be calculated. To calculate these odds, there are three pairings of populations to consider: high school and college graduation, some college and college graduation, and high school and some college, meaning nine odds ratios can be estimated. The odds of high school versus college graduation are calculated as follows:

- $\Omega_{13,1} = \frac{15}{15} = 1$
- $\Omega_{13,2} = \frac{25}{10} = 2.5$
- $\Omega_{13,3} = \frac{30}{5} = 6$

The odds ratio for Hispanic versus white individuals is calculated as:

$$\begin{aligned}\theta_{13,31} &= \frac{\Omega_{13,3}}{\Omega_{13,1}} \\ &= \frac{6}{1} \\ &= 6\end{aligned}$$

The odds ratio for Hispanic versus white individuals is 6. This means that for Hispanic individuals, the odds of high school versus college graduation are six times that of white individuals.

3.3.2 More Than Two Variables

For more than two variables, controlling for a third variable Z is often useful in assessing the relationship between two other variables X and Y. There are two options:

- A marginal contingency table provides the marginal distributions of the variables X and Y, showing the total frequencies or percentages for each variable category. In other words, it displays each row and column totals in the contingency table [24]. Let π_{ij} denote the probability that X, Y occurs in the cell in row i and column j. The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y. The marginal distributions are the row and column totals that result from summing the joint probabilities. We denote these by $\{\pi_{i+}\}$ for the row variable and $\{\pi_{+j}\}$ for the column variable, where the subscript "+" denotes the sum over that index; that is, $\pi_{i+} = \sum_j \pi_{ij}$ and $\pi_{+j} = \sum_i \pi_{ij}$ [28].

- A conditional contingency table examines the relationship between two variables X and Y while holding one variable constant. It provides the frequency distribution of one variable for each category of the other variable. It is also known as cell percentages or proportions, and it is calculated by dividing the frequency in each contingency table cell by the corresponding marginal total. Conditional frequencies express the proportion of observations in each cell relative to the total number of observations for that row or column [24]. Given that a subject is classified in row i of X, $\pi_{j|i}$ denotes the probability of classification in column j of Y, $j = 1, \ldots, J$.The probabilities $\{\pi_{1|i}, \ldots, \pi_{J|i}\}$ form the conditional distribution of Y at category i of X where $\sum_j \pi_{j|i} = 1$ [28].

For example, suppose at clinic 1, 18 of 30 males show evidence of staph infection and 12 of 20 females show the same. At clinic 2, 2 of 10 males show evidence of staph, and 8 of 40 females show the same.

The marginal contingency table for staph infection evidence for both clinics is shown below:

Table 5: Marginal table of staph infection versus for both clinics

	Male	Female
Staph	20	20
No Staph	20	40

The marginal odds ratio is calculated as follows:

$$\begin{aligned}\hat{\theta}_{XY} &= \frac{n_{11+} \cdot n_{22+}}{n_{21} \cdot n_{12+}} \\ &= \frac{20 \times 40}{20 \times 20} \\ &= 2\end{aligned}$$

The conditional contingency table for staph infection evidence for each of the clinic is shown below:

Table 6: Conditional Table for staph versus gender for Clinic 1 and Clinic 2

	Male	Female
Staph	18	12
No Staph	12	8

(a) Clinic 1

	Male	Female
Staph	2	8
No Staph	8	32

(b) Clinic 2

The conditional odds ratio for each clinic is calculated as follows:

$$\hat{\theta}_{XY|k} = \frac{n_{11k}n_{22k}}{n_{21k}n_{12k}},$$

where k is the clinics.

Clinic 1 : $\hat{\theta}_{XY|1} = \dfrac{18 \times 8}{12 \times 12} = 1$ Clinic 2 : $\hat{\theta}_{XY|2} = \dfrac{2 \times 32}{8 \times 8} = 1$

At each clinic, the odds of staph are equal for males and females. Overall, the odds of staph for males are twice that of females.

3.3.3 Hypothesis Testing of Contingency Table

Hypothesis testing for contingency tables involves examining the association of two categorical variables by comparing observed frequencies to expected frequencies under the assumption of independence. There are common tests used to test the hypothesis for contingency tables.

- Linear trend test is used to determine whether there is an increasing or decreasing trend in the levels observed of one variable as the levels of the other "increase". The test assumes both variables X and Y are ordinal. When the variables are ordinal, a trend association is common. As the level of X increases, responses on Y tend to increase toward higher levels, or responses on Y tend to decrease toward lower levels.

  The test statistic, sensitive to positive or negative linear trends, utilizes correlation information in the data. The test statistic is calculated as follows:

  $$M^2 = (n_{++} - 1)r_{uv}^2 | n_{ij}, \tag{13}$$

  where n_{++} is the total number of observations in the dataset, r_{uv} is the Pearson correlation coefficient, n_{ij} is the frequency of observations.

  The Pearson correlation coefficient (r) between X and Y equals the covariance divided by the product of the sample standard deviations of X and Y. That is,

  $$r = \frac{\sum_{i,j}(u_i - \bar{u})(v_j - \bar{v})p_{ij}}{\sqrt{\left(\sum_i (u_i - \bar{u})^2 p_{i+}\right)\left(\sum_j (v_j - \bar{v})^2 p_{+j}\right)}}, \tag{14}$$

  where u_i is the row scores, v_j is the column scores, $\bar{u}$ is the sample mean of the row scores, $\bar{v}$ is the sample mean of the column scores [28].

- The Cochran-Mantel-Haenszel (C M H) test is an alternative test of XY conditional independence in $2 \times 2 \times K$ contingency tables. This test conditions on the row totals and the column totals in each partial table. In the partial table k, the row totals are n_{1+k}, n_{2+k}, and the columns totals are n_{+1k}, n_{+2k}. Given these totals,

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}, \tag{15}$$

where μ_{11k} represents the expected frequency of observations falling into the category defined by the intersection of the first level of variable 1, the first level of variable 2, and level k of variable 3, n_{1+k} denotes the total count of observations that have the first level of variable 1 and level k of variable 3, n_{+1k} represents the total count of observations that have the first level of variable 2 and level k of variable 3 and n_{++k} denotes the total count of observations that have level k of variable 3.

$$\text{Var}(n_{11k}) = \frac{n_{1+k}n_{2+k}n_{+1k}n_{+2k}}{n_{2++k}(n_{++k} - 1)}, \tag{16}$$

where $\text{Var}(n_{11k})$ denotes the variance of the count of observations falling into the category defined by the intersection of the first level of variable 1, the first level of variable 2, and level k of variable 3, n_{1+k} represents the total count of observations that have the first level of variable 1 and level k of variable 3,n_{2+k} denotes the total count of observations that have the second level of variable 1 and level k of variable 3, n_{+1k} represents the total count of observations that have the first level of variable 2 and level k of variable 3, n_{+2k} denotes the total count of observations that have the second level of variable 2 and level k of variable 3, n_{2++k} represents the total count of observations that have level k of variable 3 and the second level of variable 1, n_{++k} denotes the total count of observations that have level k of variable 3.

Hence, the Cochran-Mantel-Haenszel (C M H) Test Statistic is stated as follows:

$$\text{CMH} = \frac{\left(\sum_k (n_{11k} - \mu_{11k})\right)^2}{\sum_k \text{Var}(n_{11k})} \tag{17}$$

The associated hypotheses for the Cochran-Mantel-Haenszel (C M H) Test for Conditional Independence are stated as follows:

Null Hypothesis (H_0): X, Y are conditionally independent.

Alternative Hypothesis (H_a): X, Y are not conditionally independent [28].

3.4 Categorical Variable Models

Categorical variable models are statistical models designed to analyze data with categorical or qualitative response variables. Categorical variable models are used when the outcome of interest is not numerical but falls into distinct categories [28].

3.4.1 Log-Linear Models

Log-linear Models are a class of statistical models used to analyze the relationship between two or more categorical variables. Log-linear models model cell counts regarding the row and column variables. The counts in the cells are treated as a third variable, as the response. These models evaluate associations that can be visualized using contingency tables [28].

3.4.2 Independence Model

The independence model analyzes contingency tables to assess the association between two or more categorical variables. This model assumes two categorical variables, X and Y, are independent, meaning that the occurrence of X is not influenced by Y. Under the assumption of independence, the independence model is given as follows:

$$\mu_{ij} = n_{++}\pi_{i+}\pi_{+j}, \tag{18}$$

where μ_{ij} is the expected frequency for cell i and j, n_{++} is the total number of observations, π_{i+} is the total count for row i, and π_{+j} is the total count for column j.

The model in equation 18 can be transformed into an additive model by applying a logarithm which is given by:

$$\ln(\mu_{ij}) = \ln(n_{++}) + \ln(\pi_{i+}) + \ln(\pi_{+j}). \tag{19}$$

Alternatively, by considering $\lambda = \ln(n_{++})$, $\lambda_i^X = \ln(\pi_{i+})$, and $\lambda_j^Y = \ln(\pi_{+j})$, the model in equation 19 can be written as:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y, \tag{20}$$

where λ_i^X is the row effect and λ_j^Y is the column effect [28].

3.4.3 Estimation of the Independence Model

The Independence Model is estimated using the Maximum Likelihood Estimation(MLE). The model assumes that the observed counts in each cell follow a Poisson probability distribution function. The Poisson probability distribution function is stated as follows:

$$f(n_{ij}|\mu_{ij}) = \frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!}, \tag{21}$$

where μ_{ij} is the mean rate of occurrence for cell i, j, and n_{ij} is the observed count.

The likelihood function for μ_{ij} is the product of the Poisson Probability Distribution Functions (PDF) for each observed count in the contingency table. It represents the probability of observing the observed cell counts given the parameter μ_{ij}. The likelihood for μ_{ij} is given by:

$$l_{ij}(\mu_{ij}|n_{ij}) = \prod_{i=1}^{I}\prod_{j=1}^{J}\frac{e^{-\mu_{ij}}\mu_{ij}^{n_{ij}}}{n_{ij}!}, \tag{22}$$

where I and J are the number of rows and columns in the contingency table, μ_{ij} is the mean rate of occurrence for cell $\{i, j\}$, and n_{ij} is the observed count [28].

3.4.4 Saturated Model

A saturated model is a statistical model that perfectly fits the observed data, meaning it achieves zero discrepancy between the observed and expected values. It represents the maximum possible degree of fit to the data, capturing all the variability present in the observed data without any loss of information. It is often used as a reference point for comparing the fit of the other, more complex models. This model includes corrections to individual cells to account for deviations from independence. The saturated model is given by:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{23}$$

The log-odds of the saturated model is stated as follows:

$$\ln(\theta) = \ln\left(\frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}\right) = \ln\left(\frac{\mu_{11}\mu_{22}}{\mu_{21}\mu_{12}}\right) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{21}^{XY} - \lambda_{12}^{XY} \tag{24}$$

The parameters λ_{ij}^{XY} determine the log odds ratio. When these parameters equal zero, the log odds ratio is zero, and X and Y are independent.

The odds ratio of the saturated model is stated as follows:

$$\theta = e^{\left(\lambda_{11}^{XY}+\lambda_{22}^{XY}-\lambda_{21}^{XY}-\lambda_{12}^{XY}\right)} = \frac{e^{\lambda_{11}^{XY}}e^{\lambda_{22}^{XY}}}{e^{\lambda_{21}^{XY}}e^{\lambda_{12}^{XY}}} \tag{25}$$

The Independence Model states the log odds as follows [28].:

$$\lambda_{11}^{XY}+\lambda_{22}^{XY} = \lambda_{21}^{XY}+\lambda_{12}^{XY} \tag{26}$$

3.5 Inference of log-linear models

Inference for log-linear models involves conducting statistical inferences about the parameters and structure of the model, checking the goodness of fit, and extending log-linear models to higher dimensions.

3.5.1 Model Fit Statistics for Log-Linear Model

Both Pearson Chi-square (χ^2) and the Likelihood Ratio (G^2) statistics are used as models for goodness of fit statistics for the log-linear model. The associated hypotheses for the goodness of fit are stated as follows:

Null Hypothesis (H_0): Model fit is good.

Alternative Hypothesis (H_a): Model fit is not good.

The Pearson χ^2 statistic is stated as follows:

$$\chi^2 = \sum_{ij}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \tag{27}$$

where n_{ij} is the observed frequency for cell i, j and $\hat{\mu}_{ij}$ is the mean rate of occurrence for cell $\{i,j\}$.

For the independence model,

$$\hat{\mu}_{ij} = \frac{n_{i+} \cdot n_{+j}}{n_{++}}, \tag{28}$$

where $\hat{\mu}_{ij}$ represents the expected cell count for the cell in the contingency table corresponding to the ith level of the row variable and the jth level of the column variable, n_{i+} represents the total frequency or count of observations in the ith row, n_{+j} represents the total frequency or count of observations in the jth column, and n_{++} represents the total frequency or count of observations in the entire contingency table.

The test value of the independence model is compared to a critical χ^2 with degrees of freedom of (I - 1)(J - 1). The degree of freedom is calculated as the number of model parameters subtracted from the number of cell counts. In the saturated model, the Chi-square (χ^2) is zero because there is a lack of discrepancy between the observed and expected frequencies.

In the likelihood ratio (G^2) statistic, the model is used to calculate the value of the likelihood using predicted values $\hat{\mu}_{ij}$ under the null hypothesis (H_0), where:

$$l_{H_0} = l(\hat{\mu}_{ij}), \tag{29}$$

where $l(\hat{\mu}_{ij})$ represents the log-likelihood based on the observed cell counts.

Under the alternative hypothesis (H_a), the data is used to calculate the value of the likelihood using the observed counts n_{ij}, where:

$$l_{H_a} = l(n_{ij}). \tag{30}$$

where $l(n_{ij})$ represents the log-likelihood based on the observed cell counts. Then, the likelihood ratio is calculated as:

$$G^2 = -2\ln\left(\frac{l_{H_0}}{l_{H_a}}\right) = -2\ln\left(\frac{l(\hat{\mu}_{ij})}{l(n_{ij})}\right) \tag{31}$$

By simplifying the likelihood ratio ($\frac{l_{H_0}}{l_{H_a}}$), we have:

$$\frac{l_{H_0}}{l_{H_a}} = \prod_{ij}\left(\frac{(n_{i+}n_{+j})}{(n_{++}n_{ij})}\right)^{n_{ij}}, \tag{32}$$

where n_{i+} represents the total frequency or count of observations in the ith row, n_{+j} represents the total frequency or count of observations in the jth column, and n_{++} represents the total frequency or count of observations in the entire contingency table.

The log of equation 32 is equivalent to the likelihood ratio (G^2) test of independence statistic, which is compared to a critical chi-square (χ^2) value with ($IJ -$ Number of Parameters) degrees of freedom [28].

3.5.2 Hypothesis Testing

Understanding the log-linear model hypothesis testing is crucial as it allows us to test the validity of our model in two key ways. The first phase involves testing the overall hypotheses about the model, while the second phase focuses on the individual parameters. This comprehensive approach ensures the robustness and reliability of our statistical analysis.

In model comparisons, the deviance of a model compares its likelihood value to that of the saturated model. The deviance statistic of the model is stated as follows:

$$D(n_{ij}|\hat{\mu}_{ij}) = -2(L(\hat{\mu}_{ij}; n_{ij}) - L(n_{ij}; n_{ij})) = G^2(M), \tag{33}$$

where $L(\hat{\mu}_{ij}$ represents the log-likelihood of the model based on the expected cell counts ($\hat{\mu}_{ij}$) and the observed cell counts (n_{ij}), $L(n_{ij}; n_{ij})$ represents the log-likelihood of the saturated model based on the observed cell counts (n_{ij}), and $G^2(M)$ represents the likelihood ratio test statistic.

The difference in Deviance values for the independence and saturated models is given by:

$$G^2(M_0; M_1) = G^2(M_0) - G^2(M_1) = -2(L(\hat{\mu}_0) - L(\hat{\mu}_1)), \tag{34}$$

where $G^2(M_0; M_1)$ represents the likelihood ratio test statistic for comparing the goodness of fit between two nested models M_0, and M_1, $G^2(M_0)$ represents the deviance statistics for the models M_0, $G^2(M_1)$ represents the deviance statistics for the models M_1, $L(\hat{\mu}_0)$ represent the log-likelihoods of the expected cell counts ($\hat{\mu}$) for models M_0, and $L(\hat{\mu}_1)$ represent the log-likelihoods of the expected cell counts ($\hat{\mu}$) for models M_1.

$G^2(M_0; M_1)$ is distributed as a chi-square (χ^2) with degrees of freedom which is equal to the difference in the number of model parameters and also used to test the associated hypothesis stated as follows:

Null Hypothesis (H_0): Extra model parameters are zero (not significant)

Alternative Hypothesis (H_a): Extra model parameters are non-zero
at least one is significant)

The individual parameter significance tests can be performed using Wald Statistics. For any parameter λ in a log-linear model, we want to test the following hypotheses:

Null Hypothesis $(H_0) : \lambda = 0$

Alternative Hypothesis $(H_a) : \lambda \neq 0$

Using the maximum likelihood estimation, the normal statistics can be used, which is given by:

$$z = \frac{\hat{\lambda}}{\mathrm{SE}(\hat{\lambda})}, \tag{35}$$

where $\hat{\lambda}$ represents the estimated parameter value and SE($\hat{\lambda}$) represents the standard error of the parameter estimate.

However, the χ^2 statistics is the square of the normal statistic which is more commonly used, is stated as follows:

$$\chi^2 = \left(\frac{\hat{\lambda}}{\mathrm{SE}(\hat{\lambda})} \right)^2, \tag{36}$$

where $\hat{\lambda}$ represents the estimated parameter value and SE($\hat{\lambda}$) represents the standard error of the parameter estimate [28].

3.6 Three-Way Table Models

When considering the modeling of cell counts using three independent classification variables X, Y, and Z, several models can be used.

- The mutual independence model assumes that all three variables X, Y, and Z are independent. The mutual independence model is of the form:

  $$\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}, \tag{37}$$

  where π_{ijk} is the observed probability of observing the combination of values (i,j,k) for X, Y, and Z, π_{i++} is the marginal probability of observing level i of variable X across all levels of variables Y and Z, π_{+j+} is the marginal probability of observing level j of variable Y across all levels of variables X and Z and π_{++k} is the marginal probability of observing level k of variable Z across all levels of variables X and Y.

By multiplying n_{++} and applying a logarithm, this corresponds to:

$$\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z \tag{38}$$

where μ_{ijk} is the expected count or mean response for the combination of values (i,j,k) for X,Y, and Z, λ is the intercept term, which is the expected count when all predictor variables are at their reference levels, λ_i^X is the effect of level i of variable X on the expected count, relative to the reference level of X, λ_j^Y is the effect of level j of variable Y on the expected count, relative to the reference level of Y, λ_k^Z is the effect of level k of variable Z on the expected count, relative to the reference level of Z.

The mutual independence model has an independent correction to λ for each variable.

- In the joint independence model, one of the variables X is to be independent of the joint distribution of the other two variables Y and Z. Similarly, the models XY and YZ allow for models with independence of Z and Y, respectively. The joint independence model is given by:

$$\pi_{ijk} = \pi_{i++}\pi_{+jk}, \tag{39}$$

where π_{ijk} is the observed probability of observing the combination of values $(i,\ j,\ k)$ for X, Y, and Z, π_{i++} is the marginal probability of observing level i of variable X across all levels of variables Y and Z, and π_{+jk} is the conditional probability of observing levels j of variable Y and k of variable Z given level i of variable X.

By multiplying n_{++} and applying a logarithm, this corresponds to:

$$\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}, \tag{40}$$

where μ_{ijk} is the expected count or mean response for the combination of values (i, j, k) for X, Y, and Z, λ is the intercept term, which is the expected count when all predictor variables are at their reference levels, λ_i^X is the effect of level i of variable X on the expected count, relative to the reference level of X, λ_j^Y is the effect of level j of variable Y on the expected count, relative to the reference level of Y, λ_k^Z is the effect of level k of variable Z on the expected count, relative to the reference level of Z, and λ_{jk}^{YZ} is the interaction effect between levels j of variable Y and k of variable Z, indicating how the joint presence of Y and Z affects the expected count, beyond the effects of the individual variables.

- The variables X and Y in the conditional independence model are to be independent and conditional on the value of variable Z. Similarly, (XY, YZ) and (XY, XZ) are conditional independence models.The conditional independence model is given by:

  $$\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}, \tag{41}$$

  where $\pi_{ij|k}$ is the observed probability of observing the combination of values (i, j) for X and Y given level k of variable Z, $\pi_{i+|k}$ is the conditional probability of observing level i of variable X given level k of variable Z, and $\pi_{+j|k}$ is the conditional probability of observing level j of variable Y given level k of variable Z.

By multiplying n_{++} and applying a logarithm, this corresponds to:

$$\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}, \tag{42}$$

where μ_{ijk} is the expected count or mean response for the combination of values (i, j, k) for X, Y, and Z, λ is the intercept term, which is the expected count when all predictor variables are at their reference levels, λ_i^X is the effect of level i of variable X on the expected count, relative to the reference level of X, λ_j^Y is the effect of level j of variable Y on the expected count, relative to the reference level of Y, λ_k^Z is the effect of level k of variable Z on the expected count, relative to the reference level of Z, λ_{ik}^{XZ} is the interaction effect between levels i of variable X and k of variable Z, indicating how the joint presence of X and Z affects the expected count, beyond the effects of the individual variables, and λ_{jk}^{YZ} is the interaction effect between levels j of variable Y and k of variable Z, indicating how the joint presence of Y and Z affects the expected count, beyond the effects of the individual variables.

- The homogeneous association model allows associations between variables X and Y, but assumes the association is independent of k that denotes the number of categories for Z.The model contains the interaction of the variables in pairs:XZ, YZ, XY. This means when there is a homogeneous XY association, there is a homogeneous XZ association and a homogeneous YZ association.

  The homogeneous association model is given by:

$$\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}, \tag{43}$$

where μ_{ijk} is the expected count or mean response for the combination of values $(i,\ j,\ k)$ for X, Y, and Z, λ is the intercept term, which is the expected count when all predictor variables are at their reference levels, λ_i^X is the effect of level i of variable X on the expected count, relative to the reference level of X, λ_j^Y is the effect of level j of variable Y on the expected count, relative to the reference level of Y, λ_k^Z is the effect of level k of variable Z on the expected count, relative to the reference level of Z, λ_{ik}^{XZ} is the interaction effect between levels i of variable X and k of variable Z, indicating how the joint presence of X and Z affects the expected count, beyond the effects of the individual variables, and λ_{jk}^{YZ} is the interaction effect between levels j of variable Y and k of variable Z, indicating how the joint presence of Y and Z affects the expected count, beyond the effects of the individual variables, and λ_{ij}^{XY} is the the interaction effect between levels i of variable X and j of variable Y, indicating how the joint presence of X and Y affects the expected count, beyond the effects of the individual variables.

- The saturated model contains all possible parameters. It includes a three-factor interaction. The saturated model is stated as:

$$\ln(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}, \qquad (44)$$

where μ_{ijk} is the expected count or mean response for the combination of values $(i,\ j,\ k)$ for X, Y, and Z, λ is the intercept term, which is the expected count when all predictor variables are at their reference levels, λ_i^X is the effect of level i of variable X on the expected count, relative to the reference level of X, λ_j^Y is the effect of level j of variable Y on the expected count, relative to the reference level of Y, λ_k^Z is the effect of level k of variable Z on the expected count, relative to the reference level of Z, λ_{ik}^{XZ} is the interaction effect between levels i of variable X and k of variable Z, indicating how the joint presence of X and Z affects the expected count, beyond the effects of the individual variables,λ_{jk}^{YZ} is the interaction effect between levels j of variable Y and k of variable Z, indicating how the joint presence of Y and Z affects the expected count, beyond the effects of the individual variables, and λ_{ij}^{XY} is the the interaction effect between levels i of variable X and j of variable Y, indicating how the joint presence of X and Y affects the expected count, beyond the effects of the individual variables, and λ_{ijk}^{XYZ} is the three-way interaction effect between levels i of variable X, j of variable Y, and k of variable Z, indicating how the joint presence of all three variables affects the expected count, beyond the effects of the individual variables and two-way interactions.

The term λ_{ijk}^{XYZ} allows odds ratios to change across levels of the third factor [28].

3.6.1 Inferential Methods

Two inferential methods are tested for three-way models. One is the Wald statistic, used to test for individual parameters, and the other is the likelihood ratio or deviance statistic, used to test for model comparisons. The deviance statistics test is given by:

$$G^2(M_0|M_1) = G^2(M_0) - G^2(M_1), \tag{45}$$

where $G^2(M_0)$ is the likelihood ratio of the saturated model (full model) and $G^2(M_1)$ is the likelihood ratio of the reduced model.

The deviance statistic test is compared to the chi-square (χ^2) with the degree of freedom of difference in parameters [28].

3.7 Logistic Regression Models

To model the probability associated with each cell, logistic regression is used. It estimates the probability of an event occurring based on a given dataset of independent variables. It is commonly used for tasks like binary classification, where the outcome can be either a yes or no, 0 or 1, or true or false [28].

3.7.1 Logistic Regression Links

The link function relates the linear combination of the predictors to the probability of the outcome variable. These link functions help transform the linear predictor into a probability, allowing logistic regression to model binary outcomes effectively. The standard link functions used in logistic regression are as follows:

- The linear link is the probability of success is equated to parameters:

$$\pi_i = X_i^T \beta, \tag{46}$$

  where π_i is the probability or expected value of the response variable for observation i, X_i^T is the transpose of the predictor variable vector X_i, converting it from a row vector to a column vector or vice versa, and β is the vector of parameters associated with the predictor variables.

- The logit link models the log odds of the probability of an event occurring. It uses the canonical statistic from the binomial distribution:

  $$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = X_i^T\beta, \tag{47}$$

  where $\ln\left(\frac{\pi_i}{1-\pi_i}\right)$ is the logit function, representing the natural logarithm of the odds of the response variable being in the positive category, π_i is the probability or expected value of the response variable for observation i, X_i^T is the transpose of the predictor variable vector X_i, converting it from a row vector to a column vector or vice versa, and β is the vector of parameters associated with the predictor variables.

  The logit transformation is preferred because it allows interpretations using odds ratios and has efficient standard errors.

- The probit link models the inverse cumulative distribution function to the standard normal distribution.

  $$\Phi^{-1}(\pi_i) = X_i^T\beta, \tag{48}$$

  where $\Phi^{-1}(\pi_i)$ is the inverse of the cumulative distribution function (CDF) of the standard normal distribution evaluated at π_i, π_i is the probability or expected value of the response variable for observation i, X_i^T is the transpose of the predictor variable vector X_i, converting it from a row vector to a column vector or vice versa, and β is the vector of parameters associated with the predictor variables.

  This takes advantage of the fact that any continuous cumulative distribution function maps from $(-\infty, \infty)$ to $[0, 1]$.

- The complementary Log-Log models the complementary log-log transformation of the probability of an event occurring. This is another creative transformation from $[0, 1]$ to $(-\infty, \infty)$.

  $$\ln(-\ln(\pi_i)) = X_i^T \beta, \tag{49}$$

  where $\ln(-\ln(\pi_i))$ is the natural logarithm of the negative log-transformed probability $-\ln(\pi_i)$, π_i is the probability or expected value of the response variable for observation i, X_i^T is the transpose of the predictor variable vector X_i, converting it from a row vector to a column vector or vice versa, and β is the vector of parameters associated with the predictor variables [28].

3.7.2 Logistic Regression Models Predictors

The logistic regression models are different combinations of predictors and their associated interpretations.

- One binary predictor model compares probabilities between two groups (Probability of Success versus Level of X).

  The one binary predictor model is presented as follows:

  $$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i, \tag{50}$$

  where π_i is the probability of the response variable being in the positive category for observation i and α_i represents different values of the log odds for each row.

  To calculate the odds of either population A and population B, the formula is given as follows:

Population A:

$$\Omega_A = \frac{\pi_1}{1-\pi_1} = e^{\alpha_1}, \tag{51}$$

where Ω_A is the odds of the response variable being in the positive category of population A.

Population B:

$$\Omega_B = \frac{\pi_2}{1-\pi_2} = e^{\alpha_2}, \tag{52}$$

where Ω_B is the odds of the response variable being in the positive category of population B.

Similarly, the odds ratio of population A and population B is calculated as:

$$\theta = \frac{\Omega_A}{\Omega_B} = \frac{e^{\alpha_1}}{e^{\alpha_2}} = e^{\alpha_1-\alpha_2} \tag{53}$$

If $\alpha_1 = \alpha_2$, then X and Y are independent.

- One categorical predictor model compares probabilities across the many groups the predictor X defines.

  The one categorical predictor model is given by:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \alpha_i, \tag{54}$$

  where π_i is the probability of the response variable being in the positive category for observation i and α_i represents different values of the log odds for each row.

  The log-odds ratio of the model is presented as follows:

$$\ln\left(\theta_{ii'}\right) = \ln\left(\frac{\pi_i(1-\pi_{i'})}{(1-\pi_i)\pi_{i'}}\right) = \text{logit}(\pi_i) - \text{logit}(\pi_{i'}) = \alpha_i - \alpha_{i'} \tag{55}$$

The odds ratio of the model is estimated using the exponential difference in parameters from the logistic regression mode, which is given by:

$$\theta_{ii'} = e^{\alpha_i - \alpha_{i'}} \tag{56}$$

- One continuous predictor model predicts probabilities of success for different populations defined by different values of the continuous predictor. The one continuous predictor model uses regression notation.

  The one continuous predictor model is presented as:

  $$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_i, \tag{57}$$

  where β_0 is the baseline value of the log-odds when $x = 0$ and β_1 is the effect of the predictor which is the expected change in the log-odds for an increase of one unit of x.

  The log - odds ratio of the model is given by:

  $$\begin{aligned} \ln\left(\theta_{(x+1),x}\right) &= \ln\left(\frac{\pi_{(x+1)}(1-\pi_x)}{(1-\pi_{(x+1)})\pi_x}\right) \\ &= \text{logit}(\pi_{(x+1)}) - \text{logit}(\pi_x) \\ &= (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1 x) \\ &= \beta_1 \end{aligned} \tag{58}$$

  The odds ratio of the model is given by:

  $$\theta_{(x+1),x} = e^{\beta_1} \tag{59}$$

- With multiple predictors, it is useful to use regression notation with indicator variables in two binary predictors model.

  Suppose x_1 and x_2 are indicators for two binary predictors of interest; the model is presented as:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i}, \tag{60}$$

  where β_0 is the expected log of the odds when both x_1 and x_2 are zero, β_1 is the adjustment to the log of the odds when $x_1 = 1$, and β_2 is the adjustment to the log of the odds when $x_2 = 1$.

  The log – odd ratio of the model is given by:

$$\ln(\theta_1) = \ln\left(\frac{\pi_{(x_1=1)}(1-\pi_{(x_1=0)})}{\pi_{(x_1=0)}(1-\pi_{(x_1=1)})}\right) = (\beta_0+\beta_1(1)+\beta_2x_2)-(\beta_0+\beta_1(0)+\beta_2x_2) = \beta_1 \tag{61}$$

  The odd ratio for is x_1, with x_2 held fixed which is similar for x_2 is given as follows:

$$\theta_1 = e^{\beta_1} \tag{62}$$

- Models with continuous and categorical predictors combine continuous and categorical predictors within one model.

  To start with one binary and one continuous predictor and also consider x_1 as an indicator and x_2 as continuous, the model is presented as:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \tag{63}$$

  where β_0 is the expected log of the odds when $x_1 = 0$ and $x_2 = 0$, β_1 is the adjustment to the odds when $x_1 = 1$ and x_2 is fixed, and β_2 is the adjustment to the log of the odds for a unit increase of x_2 and x_1 is fixed.

The odd ratio for is x_1, with x_2 held fixed which is similar for x_2 is given as follows:

$$\theta_1 = e^{\beta_1}, \tag{64}$$

where e^{β_1} is the odds ratio associated with a unit increase of x_1 when $x_2 = 0$ [28].

3.7.3 Logistic Regression Inference

Logistic regression inference involves making statistical inferences about the parameters of a logistic regression model. These inferences include estimating parameters, hypothesis testing, and assessing model fit.

(i) Parameter Estimation

Logistic regression estimates the parameters or coefficients of the model using Maximum Likelihood Estimation(MLE). The MLE finds the parameter estimates that maximize the likelihood of observing the given data, assuming the data follows a binomial distribution. The likelihood function represents the probability of observing the data given the model parameters, and MLE aims to maximize this likelihood function.

Suppose N binary responses are assumed to be independent and appear as identical Bernoulli random variables. In that case, the joint likelihood function for logistic regression is the product of the probabilities of observing the response for each observation which is given by:

$$L(\pi; \mathbf{y}) = \prod_{i=1}^{N} \pi^{y_i}(1-\pi)^{(1-y_i)}, \tag{65}$$

where $L(\pi; \mathbf{y})$ represents the joint likelihood function, $\prod_{i=1}^{N}$ denotes the product of terms from $i = 1$ to N, N is the total number of observations, π^{y_i} is the probability of observing a success, and $(1-\pi)^{(1-y_i)}$ is the probability of observing a failure.

Additionally, the log-likelihood function is the natural logarithm of the joint likelihood function which is given by:

$$l(\pi; \mathbf{y}) = \sum_{i=1}^{N} (y_i \ln(\pi) + c, \tag{66}$$

where $l(\pi; \mathbf{y})$ represents the log-likelihood function, $\sum_{i=1}^{N}$ denotes the summation of terms from $i = 1$ to N, N is the total number of observations, $y_i \ln(\pi)$ is the log-likelihood contribution from the ith observation when $y_i = 1$, and $y_i \ln(\pi)$ is the log-likelihood contribution from the ith observation when $y_i = 0$ [28].

(ii) Goodness of Fit

The goodness of fit of a logistic regression model evaluates the quality of the model and assesses how well the model fits the observed data. Several methods are commonly used to assess the goodness of fit of logistic regression models, which are as follows:

- The Pearson's Chi-square (χ^2) or Model deviance is used to classify predictors.

The Pearson's Chi-square is given by:

$$\chi^2 = \sum_{i=1}^{I} \left(\frac{(y_i - \hat{y}_i)^2}{\sqrt{\hat{y}_i(1-\hat{y}_i)}} \right)^2, \tag{67}$$

where y_i is the observed response for the ith observation and $\hat{y}_i$ is the predicted response or probability for the ith observation.

The test value is compared to the critical chi-square (χ^2) with a degree of freedom, the number of parameters in the model subtracted from the number of predictor(s) classifications.

The model deviance is given by:

$$\hat{D}(\hat{\pi};\mathbf{y}) = -2(l(\hat{\pi};\mathbf{y}) - l(\mathbf{y};\mathbf{y})) = 2\sum_{i=1}^{N}[y_i \ln\left(\frac{y_i}{\hat{\pi}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\pi}_i}\right)], \tag{68}$$

where $l(\hat{\pi};\mathbf{y})$ is the log-likelihood function evaluated at the fitted model $(\hat{\pi})$, $l(\mathbf{y};\mathbf{y})$ is the log-likelihood function evaluated at the saturated model(y), which is the maximum attainable likelihood, y_i represents the observed response for the ith observation, $\hat{\pi}_i$ represents the predicted probability of success for the ith observation obtained from the logistic regression model.

The model deviance $(\hat{D})$ is compared to a critical chi-square (χ^2) value with a degree of freedom, which is the number of parameters subtracted from the number of groups [28].

- The Hosmer-Lemeshow test divides the data into groups based on predicted probabilities and compares the observed and expected frequencies within each group. It can be used for any predictor.

By applying the χ^2 goodness of fit test:

$$\chi^2 = \sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \hat{\pi}_{ij}\right)^2}{\left(\sum_j \hat{\pi}_{ij}\right)\left(1 - \frac{1}{n_i}\sum_j \hat{\pi}_{ij}\right)}, \tag{69}$$

where $\sum_j y_{ij}$ represents the total observed frequency in the ith category of the variable, $\sum_j \hat{\pi}_{ij}$ represents the total expected frequency in the ith category of the variable under the null hypothesis of independence, n_i represents the total sample size for the ith category of one of the variables. Then, the test value is compared to a critical χ^2 value with a degree of freedom which is g – 2 [28].

For any predictor, we can compare models and not evaluate fit individually using information criteria to measure the information lost between the data and the model. These are measures of information lost between the data and the model. The smaller the value of these measures, the better they are. These measures are explained as follows:

- AIC (Akaike Information Criterion) is a statistical measure that balances the model fit and complexity and estimates the relative amount of information lost by a model. It imposes a penalty on k, the number of parameters. The AIC is calculated as follows:

  $$AIC = -2\ell(\hat{\pi}; y) + 2k, \tag{70}$$

  where $\ell(\hat{\pi}; y)$ represents the maximized log-likelihood of the fitted model, k is the number of estimated parameters in the model [25].

- BIC (Bayesian Information Criterion) is a criterion for model selection among a finite set of models. It also balances the goodness of fit of a model but imposes a stronger penalty than AIC. The BIC is calculated as follows:

$$BIC = -2\ell(\hat{\pi}; y) + k \ln(N), \tag{71}$$

  where $\ell(\hat{\pi}; y)$ represents the maximized log-likelihood of the fitted model, k is the number of estimated parameters in the model and N is the number of observations [25].

- AIC$_c$ (Akaike Information Criterion "corrected") is an adjustment made to the AIC to account for the bias that can occur when the sample sizes N is relatively small compared to the number of parameters k in the model. It imposes a stronger penalty than AIC but converges to AIC as N grows. The AIC$_c$ is calculated as follows:

$$\mathrm{AIC}_c = AIC + \frac{2k(k+1)}{N-k-1}, \tag{72}$$

  where k is the number of estimated parameters in the model and N is the number of observations in the dataset.

  The correction term $\frac{2k(k+1)}{N-k-1}$ penalizes the AIC further when the sample size is small, preventing over-fitting and providing a more accurate measure of model fit [25].

- QIC (Quasi-Likelihood Information Criterion) is an extension of the AIC for models estimated using quasi-likelihood estimation methods, such as Generalized Estimating Equations(GEE). It is used to compare the fit of different models to the same dataset. The QIC is calculated similarly to AIC, but it uses the quasi-likelihood instead of the likelihood. The QIC is calculated as follows:

$$QIC = -2\log(QL) + 2k, \tag{73}$$

where QL is the maximized value of the quasi-likelihood function of the model and k is the number of estimated parameters in the model.

- QIC_u (Quasi-Likelihood Information Criterion with under-dispersion) is an extension of QIC that accounts for potential under-dispersion in the data. It helps to select predictors. It adjusts the QIC to penalize for under-dispersion, providing a more accurate measure of model fit. The QIC_u is calculated as follows:

  $$\text{QIC}_u = QIC + \frac{k(k+1)}{N-k-1}, \tag{74}$$

  where k is the number of estimated parameters in the model and N is the number of observations in the dataset.

  The additional term $\frac{k(k+1)}{N-k-1}$ in QIC_u corrects for potential under-dispersion in the data, providing a more accurate measure of model fit.

  The QIC and the QIC_u are useful for selecting the most appropriate model for longitudinal or clustered data, considering the correlation structure within the data and potential issues such as under-dispersion [28].

- The Mean Square Error (MSE) is a metric used to assess the accuracy of a statistical model's predictions. It quantifies the average squared difference between the observed values and the predicted values by the model. In the context of model evaluation, the Mean Square Error (MSE) is often used to compare the performance of different models or assess a particular model's goodness of fit.

The MSE is calculated as the average of the squared differences between the observed values y_i and the predicted values $\hat{y}_i$ for each observation i. Mathematically, it is defined as:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{75}$$

where n is the number of observations in the dataset,y_i represents the observed value for observation i, $\hat{y}_i$ represents the predicted value for observation i [26].

(iii) Hypothesis Testing

The hypothesis testing involves making inferences about the model's parameters, which assesses the significance of the relationship between the predictor variables and the response variable. There are two methods of hypothesis testing, which are as follows:

- Individual Parameters

  In logistic regression, the significance of individual predictors is often assessed using Wald tests. The Wald test assesses whether the coefficient of each predictor is significantly different from zero.

- Model Comparisons

  The likelihood ratio (deviance) test compares the fit of a model with predictors to the fit of a null model with no predictors. It assesses whether adding predictors significantly improves the model's fit [28].

3.8 Multinomial Logistic Regression Model

The multinomial logistic regression model is an extension of binary logistic regression that allows the prediction of outcomes with more than two categories. It is used when the dependent variable is categorical with more than two levels, but the independent variables are still continuous or categorical. It models a categorical response using any predictor and assumes the total number of responses is fixed. Multinomial logistic regression is widely used in various fields, including social sciences, epidemiology, marketing, and finance, whenever the outcome of interest has multiple unordered categories [28]. There are different situations for multinomial logistic regression models, which are considered below.

3.8.1 Nominal Responses

In the Multinomial Logistic Regression model, a nominal response variable has distinct categories and no inherent order or ranking. Each category is distinct and separate, with no natural ordering between them. It constructs a model that compares each response category with a baseline category. The baseline category is arbitrary.

The multinomial logistic regression considering nominal responses are as follows:

$$\ln\left(\frac{\pi_j}{\pi_B}\right) = \beta_{0j} + \sum_{k=1}^{K} \beta_{kj} x_k, \tag{76}$$

where π_j is the probability of interest, π_B is the baseline probability, β_{0j} is the unique intercept for each response category, β_{kj} is the unique slope for each predictor at each response category, and x_k does not depend on j.

In this case of nominal response, we are constructing separate logistic regression models for each response category except the baseline. When the last category (J) is equal to 2, the model in equation 70 simplifies to a single equation for $\log\left(\frac{\pi_1}{\pi_2}\right) = \text{logit}(\pi_1)$, resulting in ordinary logistic regression for binary responses [28].

3.8.2 Estimation of Multinomial Logistic Regression Models

Estimating a multinomial logistic regression model involves determining the coefficients associated with each predictor variable for each category of the dependent variable. The Maximum Likelihood Estimation is used to estimate the multinomial logistic regression model to find the parameter estimates that maximize the likelihood function, representing the probability of observing the given sample data given the model parameters [28].

3.8.3 Hypothesis Testing of Multinomial Logistic Regression Models

Hypothesis testing in multinomial logistic regression involves assessing the significance of the predictor variables in explaining the variability in the outcome variable, which has multiple unordered categories. The Wald Test assesses whether the coefficients of the predictor variables are significantly different from zero. The Likelihood Ratio Test compares the fit of the full model to a reduced model with fewer predictors [28].

3.8.4 Model fit of Multinomial Logistic Regression Models

The model fit of a multinomial logistic regression model assesses how well the model predicts the observed outcome categories based on the predictor variables. The Deviance or Pearson χ^2 statistic measures the difference between the log-likelihood of the fitted model and the log-likelihood of the saturated model, which is a model with a perfect fit for classification predictors [28]

3.8.5 Ordinal Responses

The ordinal response variable consists of categories with a natural order or ranking. While the categories are distinct, like nominal variables, they also have an inherent order or hierarchy.

In ordinal response, three standard link functions are used, which are as follows:

- Cumulative logit model assumes that the cumulative log - odds of being in the category j or below. It is precisely for situations where the outcome variable has ordered responses, splits categories into half, and models the lower and upper half. It can be expressed as a linear combination of the predictor variables:

  $$\log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) = \ln\left(\frac{\sum_{k=1}^{j} \pi_k}{\sum_{k=j+1}^{J} \pi_k}\right) = \beta_{0j} + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p, \quad (77)$$

  where $P(Y \leq j)$ is the cumulative probability of being in category j or below, β_{0j},$\beta_1 X_1$,$\beta_2 X_2$, ...,β_p are the model coefficients.

- Adjacent categories logit model compares each category to the previous or the following category. The model will be presented as follows:

$$\ln\left(\frac{\pi_j}{\pi_{(j-1)}}\right) = \beta_{0j} + \sum_{k=1}^{K}\beta_k x_k \quad (j = 2, \ldots, J), \tag{78}$$

where $\ln\left(\frac{\pi_j}{\pi_{(j-1)}}\right)$ represents the natural logarithm of the odds of being in category j compared to category $(j-1)$, β_{0j} represents the unique intercept for each response category, and $\sum_{k=1}^{K}\beta_k x_k$ represents the sum of products of the regression coefficients (β_k) and the predictor variables (x_k).

This model is used if response categories are sufficiently far from each other that comparisons beyond adjacent categories do not have meaning.

- Continuation ratio logit model compares each response category to all previous or following categories. The model will be presented as:

$$\ln\left(\frac{\pi_j}{\sum_{l=1}^{j-1}\pi_l}\right) = \beta_{0j} + \sum_{k=1}^{K}\beta_k x_k \quad (j = 2, \ldots, J), \tag{79}$$

where $\ln\left(\frac{\pi_j}{\sum_{l=1}^{j-1}\pi_l}\right)$ represents the natural logarithm of the odds of belonging to category j compared to the cumulative odds of belonging to categories 1 through $(j-1)$, β_{0j} represents the unique intercept for each response category, and $\sum_{k=1}^{K}\beta_k x_k$ represents the sum of products of the regression coefficients (β_k) and the predictor variables (x_k).

This model is often used when each level is associated with the sum of all previous levels [28].

3.9 Generalized Linear Models (GLMs)

Generalized linear models are a class of statistical models that generalize linear regression to handle response variables with error distributions other than the normal distribution. Log-linear, binary, and multinomial logistic regression models all involve the function of response parameter(s) equal to the linear combination of predictors and model parameters.

This approach can be formulated for any response from the exponential family of distributions, such as Normal, Poisson, and Binary distributions. The right-hand side of the equation looks like a standard linear model, but the left-hand side involves a transformation of the mean [28].

3.9.1 Generalized Linear Models (GLMs) Components

Any Generalized Linear Model is made up of three components:

- A random component represents the distributional assumptions about the response variable Y. It accounts for the variability or randomness in the response variable that is not explained by the predictor variables. It is specified by selecting a probability distribution from the exponential family of distributions, such as normal, binomial, Poisson, or gamma distributions.

- A systematic component describes the relationship between predictor variables X and parameters in the model.

- A link component connects the mean of the response to the model's parameters. It links the other two components.

For example, for the logistic regression model, we would have:

Random component: $Y_i \sim \text{Bin}(1, \pi)$,

where Y_i represents the outcome of the ith observation, which can take values of either 0 or 1 and $\text{Bin}(1, \pi)$ indicates that the distribution of Y_i is a Bernoulli distribution with parameter π.

Systematic component: $\eta_i = \mathbf{X}_i^T \beta$,

where η_i represents the linear predictor for the ith observation, X_i represents the vector of predictor variables for the ith observation, and β represents the vector of coefficients or parameters associated with the predictor variables.

Link component: $\eta_i = \ln\left(\frac{\pi}{1-\pi}\right)$,

where η_i represents the linear predictor for the ith observation, π represents the probability of success [28].

3.9.2 Generalized Linear Models (GLMs) Properties

A generalized linear model is appropriate for a response distributed according to any exponential distribution. The exponential distributions are formulated as follows:

$$f(y; \theta, \phi) = e^{\left(\frac{(y\theta - b(\theta))}{(a(\phi))} + c(y,\phi)\right)}, \tag{80}$$

where θ is the canonical parameter,ϕ is the dispersion parameter, $b(\theta)$ is a scalar function, and $c(y, \phi)$ is a normalizing function.

Here, we will show how a Poisson distribution can be classified as an exponential distribution. The Poisson distribution can be modeled as:

$$f(y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!} = e^{\left[y\ln(\lambda) - e^{\ln(\lambda)} - \ln(y!)\right]}, \tag{81}$$

where $\theta = \ln(\lambda), \quad b(\theta)= e^{\theta}, \quad a(\phi) = 1, \quad c(y, \phi) = -\ln(y!)$

3.9.3 Generalized Linear Models (GLMs) Parameter Estimation

In Generalized Linear Models, parameter estimation involves finding the values of the model coefficients that maximize the likelihood of observing the given data. The process often involves iterative algorithms, such as Newton-Raphson or Fisher scoring, to iteratively update the parameter estimates until convergence. The parameter estimation process is the Maximum Likelihood Estimation.

Maximum Likelihood Estimation is used to find the parameter estimates that maximize the log-likelihood function. The goal is to find the values of β that maximize $l(\beta)$.

For an individual response y_i, the log-likelihood is given as:

$$l_i(\theta_i; y_i, \phi) = \frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i, \phi), \tag{82}$$

where $l_i(\theta_i; y_i, \phi)$ represents the log-likelihood function for the ith observation,y_i is the observed value of the response variable for the ith observation, θ_i is the canonical parameter associated with the response variable, ϕ is the dispersion parameter, and $c(y_i, \phi)$ is a constant term that does not depend on the parameters.

3.9.4 Generalized Linear Models (GLMs) Inference

Inference in generalized linear models involves making statistical inferences about the model's parameters and assessing the model's goodness of fit to the data. The components of inference in the generalized linear model are as follows:

- Hypothesis Testing

  In Generalized Linear Models, the hypothesis testing is done by constructing Wald Statistics based on Maximum Likelihood Estimation:

  $$z_i^* = \frac{\hat{\beta}_i - 0}{\sqrt{[I(\hat{\beta})]_{ii}^{-1}}}$$

  where z_i^* represents the standardized coefficient estimate for the ith parameter $(\hat{\beta}_i)$ represents the estimated coefficient for the ith parameter, and I is the Fisher information matrix.

  The Fisher information matrix is given by:

  $$I(\hat{\beta}) = E\left[\left(\frac{\partial l}{\partial \beta}\right)^2\right]_{|\beta=\hat{\beta}},$$

  where $E\left[\left(\frac{\partial l}{\partial \beta}\right)^2\right]_{|\beta=\hat{\beta}}$ is the expected value of the squared score function.

- Model Fit

  The Deviance test is used to assess the goodness of fit by comparing the deviance of the fitted model to the deviance of a saturated model.

  $$\hat{D}(\hat{\mu}; y) = -2(l_{model} - l_{data}) = 2\sum_{i=1}^{N} \frac{1}{a(\phi)}\left(y_i(\theta_{y_i} - \hat{\theta}) - (b(\theta_{y_i}) - b(\hat{\theta}))\right), \quad (83)$$

  where $\hat{D}(\hat{\mu}; y)$ represents the deviance statistics, l_{model} represents the log-likelihood of the fitted model, l_{data} represents the log-likelihood of the saturated model, N represents the number of observations, y_i represents the observed response variable for the ith observation, and $\hat{\theta}$ represents the fitted values of the mean response variable obtained from the model.

This is called the scaled deviance with $a(\phi)$ included in the above equation. Without it, it is simply the deviance. Components are called the deviance residuals [28].

4 ANALYSIS OF RESULTS

This chapter aims to thoroughly discuss the research questions and provide the relevant information and analyses to clarify their significance. The research questions are as follows:

RQ 1: What are the critical reasons for prioritizing the study of depression, and how does a deeper understanding of this mental health condition contribute to improved prevention, intervention, and overall well-being in individuals and society?

RQ 2: How can Log-Linear Models, Multinomial Logistic Regression, and Generalized Linear Models(GLM) be employed to analyze the association between depression, sleep disturbances, and self-esteem, shedding light on the intricate relationships within mental health?

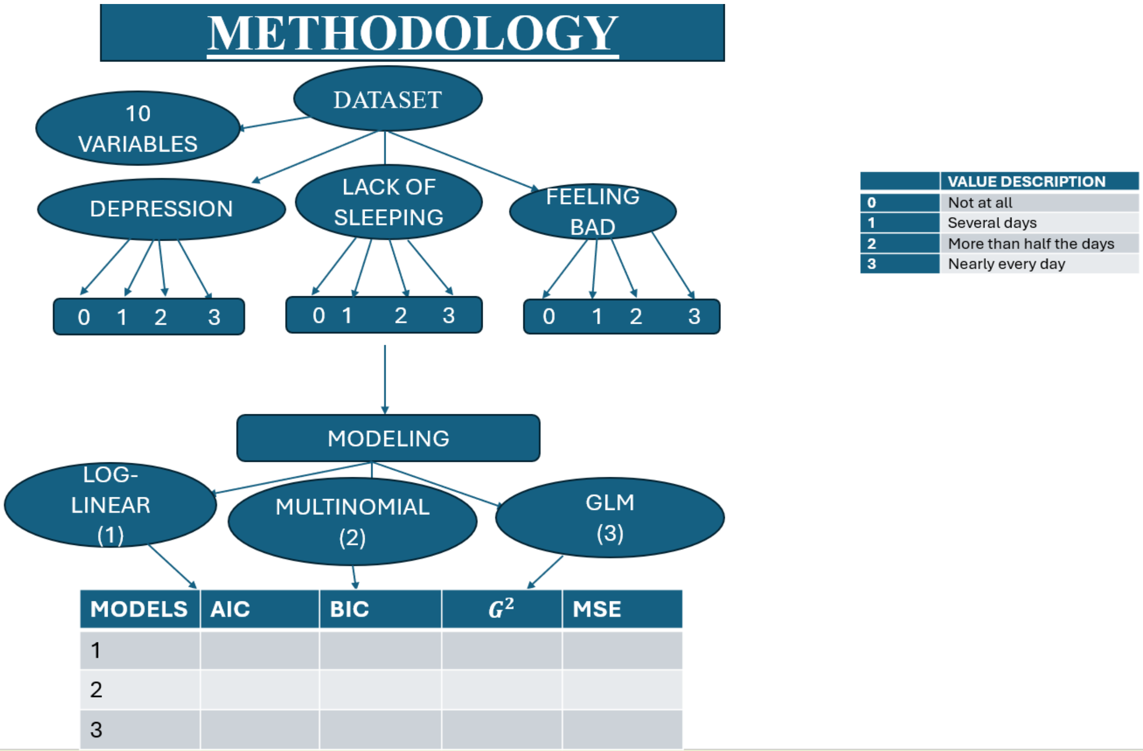
RQ 3: How can the performance of proposed statistical models be effectively compared using statistical measures such as the Likelihood ratio test, Pearson chi-square, Mean Square Error, Bayesian Information Criterion (BIC), and Akaike Information Criterion (AIC)?

4.1 Description of Data

The dataset is obtained from the National Center for Health Statistics(NCHS). The data was downloaded from the National Health and Nutrition Examination Survey (NHANES) website [22]. In this data set, a nine-item depression screening instrument, also called the Patient Health Questionnaire, was administered to determine the frequency of depression symptoms over the past two weeks with a follow-up question to assess the overall impairment of the symptoms [19]. The responses for the nine-item instrument were categorized as “not at all”, “several days,” “more than half the days,” and “nearly every day” and were given a point ranging from 0 to 3. In this data, mental health is measured as “little interest in doing things”, “depression,” “trouble sleeping” or “sleeping too much,” “feeling tired” or “having little energy,” “poor appetite” or “overeating,” “feeling bad about yourself,” “lack of concentration,” “moving or speaking slowly or too fast,” and suicidal thought. The questions are asked at the Mobile Examination Center (MEC) by trained interviewers using the Computer-Assisted Personal Interview system as part of the MEC interview. Both male and female participants aged twelve and older are eligible for the screening, but only data from participants aged eighteen and older are studied in this research. Participants requiring a proxy were not eligible because of the sensitive nature of the questions. The data for youth aged twelve to seventeen is accessible through the NCHS Research Data Center [23]. In this study, three categories “depression”, “trouble sleeping,” and “feeling bad about yourself,” are considered.

4.2 Methodology Flowchart: Analysis Techniques

In the methodology of this thesis, a detailed flowchart has been included to illustrate the progression of techniques utilized. The dataset consists of 10 variables with 5162 observations. Three of the variables are considered for this research. The data analysis is done by checking the descriptive statistics such as mosaic plots and contingency tables, and the relationship of these variables and their effect on mental health is analyzed using the Log-linear Model, Multinomial Logistic Regression Model, and Generalized Linear Models. Each variable will be analyzed as a dependent variable, and the other two will be independent. The Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Deviance Statistic, and Mean Square Error (MSE) of these models are used to obtain the best model for the analysis.



	VALUE DESCRIPTION
0	Not at all
1	Several days
2	More than half the days
3	Nearly every day

MODELS	AIC	BIC	G^2	MSE
1				
2				
3				

Figure 1: Methodology Flowchart

4.3 Analysis of Feeling Bad About Yourself

The feeling bad about yourself variable will be analyzed. At the same time, depression and trouble sleeping are kept constant. To explain multi-dimensional categorical data, it would be better to look at the structure between variables visually.

A mosaic plot is an area-proportional visualization of frequencies composed of the cells created by a rectangle's recursive vertical and horizontal splits [27]. The plot is a graphical display that allows the examination of relationships among two or more categorical variables. It is a special type of stacked bar chart. It provides a visual representation of a contingency table.

The plot in Figure 2 is drawn for this data, and we would like to highlight if there are too many depressed people who have trouble sleeping. The plot depicts relationships between “Depression”, “Trouble Sleeping,” and “Feeling Bad About Yourself.” “Depression” levels are shown on the vertical axis, and “Trouble Sleeping” levels are shown on the horizontal axis, ranging from 0 to 3. On the right side of the plot, there is a vertical bar labeled “Pearson residuals” with a scale from $-$ 12 to 69 which provide a visual presentation of the deviations of observed cell frequencies from the frequencies expected under the assumption of independence between “Depression” and “Trouble Sleeping.” Large positive residuals greater than 4 can be found for “Not At All” levels of “Depression” and “Trouble Sleeping” and are colored in green. These positive residuals (4 to 69) indicate that the observed frequency is higher than expected under the independence assumption. On the other hand, there are large negative residuals, which are less than 4 for “Several Days of Depression” or “Several Days of Trouble Sleeping,” colored orange. These negative residuals indicate that the observed frequency is lower than expected under the independence assumption. Residuals between $-$ 4 and 4, shaded in gray, indicate that the observed frequency is close to the expected frequency under the independence assumption. The p-value less than $2.22e^{-}16$ represents the probability of observing deviations from independence computed from a chi-square distribution with a degree of freedom 9. The degree of freedom of the chi-square (χ^2) test of independence is calculated as follows:

$$\begin{aligned} df &= (I-1)(J-1) \\ &= (4-1)(4-1)) \\ &= 9, \end{aligned}$$

where df is the degree of freedom, I is the number of levels of "Depression" and J is the number of levels of "Trouble Sleeping".

The association between "Depression" and "Trouble Sleeping" will be evaluated statistically with the p-value from the chi-square distribution in the mosaic plot. The associated hypotheses for the chi-square distribution are as follows:

Null Hypothesis (H_0): There is no association between depression and trouble sleeping.

Alternative Hypothesis (H_a): There is an association between depression and trouble sleeping.

The p-value is small enough to reject the null hypothesis of independence and conclude that there is an association between depression and trouble sleeping.

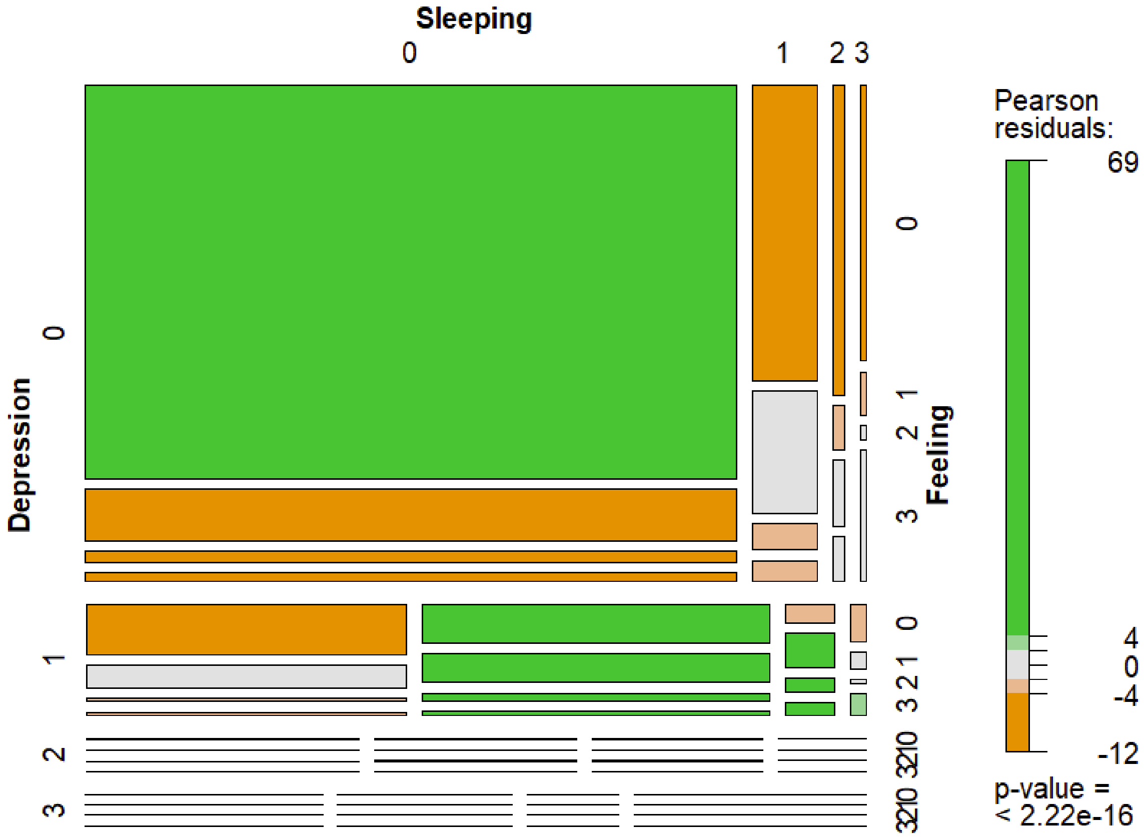


Figure 2: The mosaic plot of feeling bad about yourself

There are two ways to account for the third variable, “Feeling bad about yourself,” which are conditional and marginal tables. Conditional contingency tables can be constructed for each “Feeling bad about yourself” level as follows. Tables 7, 8, 9, and 10 display the conditional contingency table between “Depression” and “Trouble sleeping” for each level of “Feeling bad about yourself. ”.

Table 7 displays the conditional contingency table between “Depression” and “Trouble sleeping” for the ‘Not At All” level of “Feeling bad about yourself.”

Table 7: Conditional contingency table for “Not At All” level of “Feeling Bad About Yourself”.

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	2851	216	42	19
Several	233	198	14	9
More than half	65	26	22	8
Nearly	57	20	8	11

Table 8 displays the conditional contingency table between “Depression” and “Trouble sleeping” for “Several Days” level of “Feeling bad about yourself.”

Table 8: Conditional contingency table for “Several Days” level of “Feeling Bad About Yourself”.

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	374	90	6	3
Several	109	146	25	4
More than half	12	28	14	8
Nearly	5	18	8	16

Table 9 displays the conditional contingency table between “Depression” and “Trouble sleeping” for the “ More Than Half The Days” level of “Feeling bad about yourself.”

Table 9: Conditional contingency table for "More Than Half The Days" level of "Feeling Bad About Yourself"

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	83	19	9	1
Several	16	36	10	1
More than half	8	11	14	8
Nearly	6	8	4	8

Table 10 displays the conditional contingency table between "Depression" and "Trouble sleeping" for the "Nearly Everyday" level of "Feeling bad about yourself."

Table 10: Conditional contingency table for "Nearly Every Day" level of " Feeling Bad About Yourself"

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	59	15	6	9
Several	14	22	9	5
More than half	10	5	9	7
Nearly	8	10	9	40

Since each category has four levels, we can calculate several odds ratios. For simplicity, one level of the "Feeling Bad About Yourself" variable, which is more meaningful, is picked to compute the odds ratio. The calculation of the odds ratio between "Several Days" and "More Than Half The Days" levels of "Feeling Down, Depressed, or Hopeless" and "Trouble Sleeping" for "More Than Half The Days" level of "Feeling Bad About Yourself" is done with the circled values in the conditional contingency table 11.

Table 11: Conditional contingency table for "More Than Half The Days" level of "Feeling Bad About Yourself" for odds ratio calculation.

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	83	19	9	1
Several	16	㊱	⑩	1
More than half	8	⑪	⑭	8
Nearly	6	8	4	8

The conditional probability, $\theta_{XY|Z}$ can be calculated as follows:

$$\begin{aligned}\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\ &= \theta_{XY|4} = \frac{n_{222}n_{322}}{n_{232}n_{332}} \\ &= \frac{36 \times 14}{11 \times 10} \\ &= \frac{504}{110} \\ &= 4.582\end{aligned}$$

It means that for those who have had depression for several days, the odds of "several days of trouble sleeping" will be 4.58 times more than those who have had depression for more than half the days due to more than half the days of feeling bad about themselves.

Also, the calculation of the odds ratio between "Several Days" and "Nearly Every Day of "Feeling Down" and having "Trouble Sleeping" for "More Than Half The Days" of "Feeling Bad About Yourself" is done with the circled figures in the conditional contingency table 12.

Table 12: Conditional contingency table for “More Than Half The Days” level of “Feeling Bad About Yourself” for second odds ratio calculation.

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	83	19	9	1
Several	16	㊱	10	①
More than half	8	11	14	8
Nearly	6	⑧	4	⑧

The conditional probability $\theta_{XY|Z}$ can be calculated as follows:

$$\begin{aligned}
\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\
&= \theta_{XY|4} = \frac{n_{222}n_{442}}{n_{242}n_{422}} \\
&= \frac{36 \times 8}{8 \times 1} \\
&= \frac{288}{8} \\
&= 36
\end{aligned}$$

It means that for those who have depression for several days, the odds of “several days of trouble sleeping” will be 36 times more than those who have depression nearly every day due to “more than half the days feeling bad about themselves”. If the rest of the odds ratio is calculated, it is found that all odd ratios are greater than one. In other words, $\theta_{XY|2} = c > 1$, for all X,Y, where c is a constant.

We conclude that depression has a greater effect on mental health than trouble sleeping, which are independent of each other condition on “more than half the days feeling bad about yourself. ”.

We will determine whether there is an increasing or decreasing trend between the levels of depression and trouble sleeping. The associated hypotheses for the linear trend are given by:

Null Hypothesis (H_0): There is no trend between depression and trouble sleeping.

Alternative Hypothesis (H_a): A positive linear trend exists between depression and trouble sleeping.

The linear trend test statistic is calculated as follows:

$$\begin{aligned} Q &= \sqrt{(n_{++} - 1)r^2} \\ &= \sqrt{(242 - 1)(1)^2} \\ &\approx 15.524, \end{aligned}$$

where n_{++} is the total number of counts in the conditional contingency table for "More Than The Days" level of "Feeling Bad About Yourself" and r is the weighted correlation between "Depression" and "Trouble Sleeping".

The linear trend test statistic value is compared with the critical value of chi-square with a degree of freedom 3 ($\chi^2(3) = 7.815$). The degree of freedom of chi-square is calculated as follows:

$$\begin{aligned} df &= k - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where df is the degree of freedom and k is the number of levels of "Feeling Bad About Yourself".

This comparison suggests the null hypothesis (H_0) should be rejected since the test statistic value is greater than the critical value. The data provide sufficient evidence to conclude there is a positive linear association between trouble sleeping and depression due to more than half the days feeling bad about yourself. This test must be appropriate only if researchers suspect a positive linear association before seeing data.

Then, the marginal contingency table that presents the total counts across all levels of "Feeling Bad About Yourself" levels is shown below in Table 13.

Table 13: Marginal contingency table between "Depression" and "Trouble Sleeping."

	Trouble Sleeping			
Depression	Not at all	Several	More than half	Nearly
Not at all	3367	340	63	32
Several	372	402	58	19
More than half	95	70	59	31
Nearly	76	56	29	75

The odds ratio between "several days" and "more than half the days" of feeling down and having trouble sleeping would be calculated. Therefore, the joint probability (θ_{XY}) between "Depression" (X) and "Trouble Sleeping" (Y) can be calculated as follows:

$$\begin{aligned}\theta_{XY} &= \frac{n_{11+} \cdot n_{22+}}{n_{12+} \cdot n_{21+}} \\ &= \frac{402 \times 59}{70 \times 58} \\ &= \frac{23718}{4060} \\ &= 5.84\end{aligned}$$

It means that for those who have depression for several days, the odds of several days of trouble sleeping will be 5.84 times more than those who have depression for more than half the days. If we calculate all odds ratios for the marginal table, we will find that all are greater than 1. Therefore, we would conclude the odds of success are more significant for those who have depression.

Finally, a test is performed to evaluate the conditional independence (conditional on Feeling Bad About Yourself level). An appropriate test would be the Cochran-Mantel-Haenszel (C M H) Test of conditional independence [28]. The Cochran-Mantel-Haenszel Test assesses the conditional independence of categorical predictors associated with categorical outcomes.

The associated hypotheses of the Cochran-Mantel-Haenszel Test are stated as follows:

Null Hypothesis (H_0): Depression and Trouble sleeping are conditionally independent.

Alternative Hypothesis (H_a): Depression and Trouble sleeping are not conditionally independent.

The Cochran-Mantel-Haenszel (C M H) χ^2 statistic is approximately 1178, which is greater than the chi-square critical value of the degree of freedom 9 ($\chi^2(9) = 16.919$). The degree of freedom of the chi-square test of conditional independence is calculated as follows:

$$\begin{aligned} df &= (I-1)(J-1) \\ &= (4-1)(4-1)) \\ &= 9, \end{aligned}$$

where *df* is the degree of freedom, I is the number of levels of “Depression” and J is the number of levels of “Trouble Sleeping.”

We would reject H_0 since the test statistic value exceeds the critical value. We conclude that “Depression” and “Trouble Sleeping” are not conditionally independent conditional on each “Feeling Bad About Yourself” level. For at least one feeling level, there is a significant association between depression and trouble sleeping.

We will investigate if there is any homogeneous association between all two-factor interaction terms, i.e., XY (Depression × Sleeping). The appropriate model would be the Log-linear Homogeneous Association Model (M_0). The general form of the Loglinear Homogeneous Association Model for this data is given by:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \text{for } i = 1,2,3,4, j = 1,2,3,4 \tag{84}$$

The fitted model can be represented as follows:

$$ln(\mu_{ijk}) = -1.348789 + 0.313508_i^{\text{Depression}} + 0.525302_j^{\text{Sleeping}} - 0.062089_{ij}^{\text{Depression} \times \text{Sleeping}} \tag{85}$$

To perform a test to assess the overall fit of this model, the deviance statistic, $G^2(M_0)$, is calculated by finding the difference between the null deviance, which is the saturated or full model and the residual deviance which is the reduced model.

The associated hypotheses for the deviance statistic are stated as follows:

Null Hypothesis (H_0): Extra model parameters are zero (not significant).

Alternative Hypothesis (H_a): Extra model parameters are non-zero
(at least one is significant).

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 6177.6 - 5242.2 \\ &= 935.35 \end{aligned}$$

The degree of freedom of the chi-square test of the Log-linear Homogeneous Association Model is calculated as follows:

$$\begin{aligned} df &= null_{df} - residual_{df} \\ &= 5161 - 5158 \\ &= 3, \end{aligned}$$

where df is the degree of freedom, $null_{df}$ is the degree of freedom of the null deviance and $residual_{df}$ is the degree of freedom of the residual deviance.

By comparing the test value to the critical value ($\chi^2_{0.05}(3) = 7.815$), we would reject the null hypothesis. Therefore, the data provide sufficient evidence to conclude the homogeneous association model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, sleeping, and feeling bad about ourselves.

We will model a categorical response, "Feeling Bad About Yourself," using any predictor and assume the total number of responses is fixed. The appropriate model would be the Multinomial Logistic Regression model. The Multinomial Logistic Regression Model uses three standard link functions: cumulative logit, adjacent categories logit, and continuation ratio logit.

The Cumulative Logit Model is used in this analysis to split categories in half model lower half versus upper half. The model has different intercepts because it models the cumulative probabilities of observing an outcome falling into a specific category or below. Each intercept represents the log odds of being in or below a specific category relative to a reference category.

The model can be written mathematically as follows:

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \quad , \quad j = 0, 1, 2$$

So,

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 \text{ Depression}_i + \beta_2 \text{ Sleeping}_i$$

More specifically, the model can be expanded as follows:

$$\begin{aligned}
\log\left(\frac{P_0}{P_1 + P_2 + P_3}\right) &= 1.5523 + 0.4206 \text{ Depression}_i + 0.7741 \text{ Sleeping}_i, \\
\log\left(\frac{P_0 + P_1}{P_2 + P_3}\right) &= 2.9644 + 0.4206 \text{ Depression}_i + 0.7741 \text{ Sleeping}_i, \text{ and} \\
\log\left(\frac{P_0 + P_1 + P_2}{P_3}\right) &= 3.8131 + 0.4206 \text{ Depression}_i + 0.7741 \text{ Sleeping}_i
\end{aligned}$$

To perform a test to assess the overall fit of this model, the deviance statistics, $G^2(M_0)$ is calculated and compared to the chi-square critical value. The associated hypotheses of the deviance statistic are stated as follows:

Null Hypothesis (H_0): Model fit is good.

Alternative Hypothesis (H_a): Model fit is not good.

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 8414.769 - 7699.33 \\ &= 715.44 \end{aligned}$$

The degree of freedom of the chi-square test of the Cumulative Logit Model is calculated as follows:

$$\begin{aligned} df &= K - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where df is the degree of freedom, K is the number of levels of "Feeling Bad About Yourself".

By comparing the test value to the critical value ($\chi^2_{0.05}(3) = 7.815$), we would reject the null hypothesis. Therefore, the data provide sufficient evidence to conclude the cumulative logit model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, sleeping, and feeling bad about ourselves.

We will model the response variable, "Feeling Bad About Yourself," with generalized linear models (GLMs). Since the data's response variable is a count variable, we will use the Poisson regression and the negative binomial regression.

Using a log-linear model, Poisson regression models the relationship between predictor variables and the expected value of a count variable (response variable). The general form of the model is given by:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{86}$$

where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Depression) and X_{2i} (Trouble Sleeping) respectively.

So, the model will be represented as follows:

$$\ln(\hat{\mu}) = -1.27599 + 0.23081(\text{ Depression}_i) + 0.41123(\text{ Sleeping}_i). \tag{87}$$

Now, we are interpreting the coefficients in two ways, general and specific interpretations.

The general interpretation is as follows:

"There is an expected increase in the mean count of people feeling bad about themselves".

The specific interpretations are given as follows:

"The number of times people feel bad about themselves with depression is expected to be $e^{0.23081} \approx 1.259$ times, while trouble sleeping is constant" and "the number of times people feel bad about themselves with trouble sleeping is expected to be $e^{0.41123} \approx 1.509$ times, while depression is constant".

The assumptions and fit of the Poisson model are independent of response, there is no collinearity among predictors, Poisson sampling is good, and variance is equal to the mean. These assumptions are checked to determine whether the model is a good fit. The model is a Poisson sampling in which time is fixed, and the total is unknown. The data was collected over two weeks.

The histogram of the data is right-skewed which is shown in Figure 3 below:

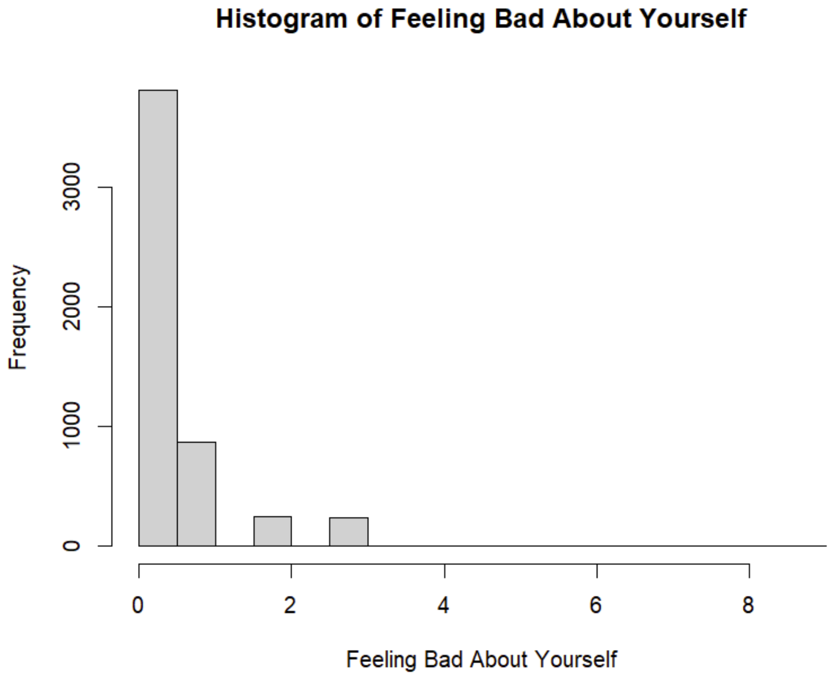


Figure 3: The histogram of feeling bad about yourself

The overdispersion test checks if the variance is equal to the mean. The associated hypotheses of overdispersion test is stated as:

Null Hypothesis (H_0): variance = mean

Alternative Hypothesis (H_1): variance > mean

The dispersion estimate ($\hat{\phi}$) = 1.437269, which is greater than 1. The p-value for the over-dispersion test is 2.2×10^{-16}, which is less than $\alpha = 0.05$. The variance is equal to 0.6681336, greater than the mean, equal to 0.4074002. Therefore, the null hypothesis is rejected, and we conclude that the variance is greater than the mean.

A scatter plot is commonly used for visualizing the relationship between two continuous variables. The plot in Figure 4 depicts the scatter plot of the Poisson Regression Model. The x-axis represents feeling bad about yourself, while the y-axis displays deviance residuals from the Poisson model. These residuals quantify the deviation between observed and model-predicted values, with values closer to zero indicating better model fit. The plot's spread showcases how "Feeling Bad About Yourself" categories correlate with deviance residuals, with discernible vertical alignments potentially indicating discrete or categorical data. Clusters of points at various "Feeling Bad About Yourself" levels demonstrate how residuals vary across this variable, with patterns or trends suggesting model influential outliers. Outlying points, located farther from the central horizontal line (y=0), denote larger discrepancies between observed and predicted values, warranting further investigation.

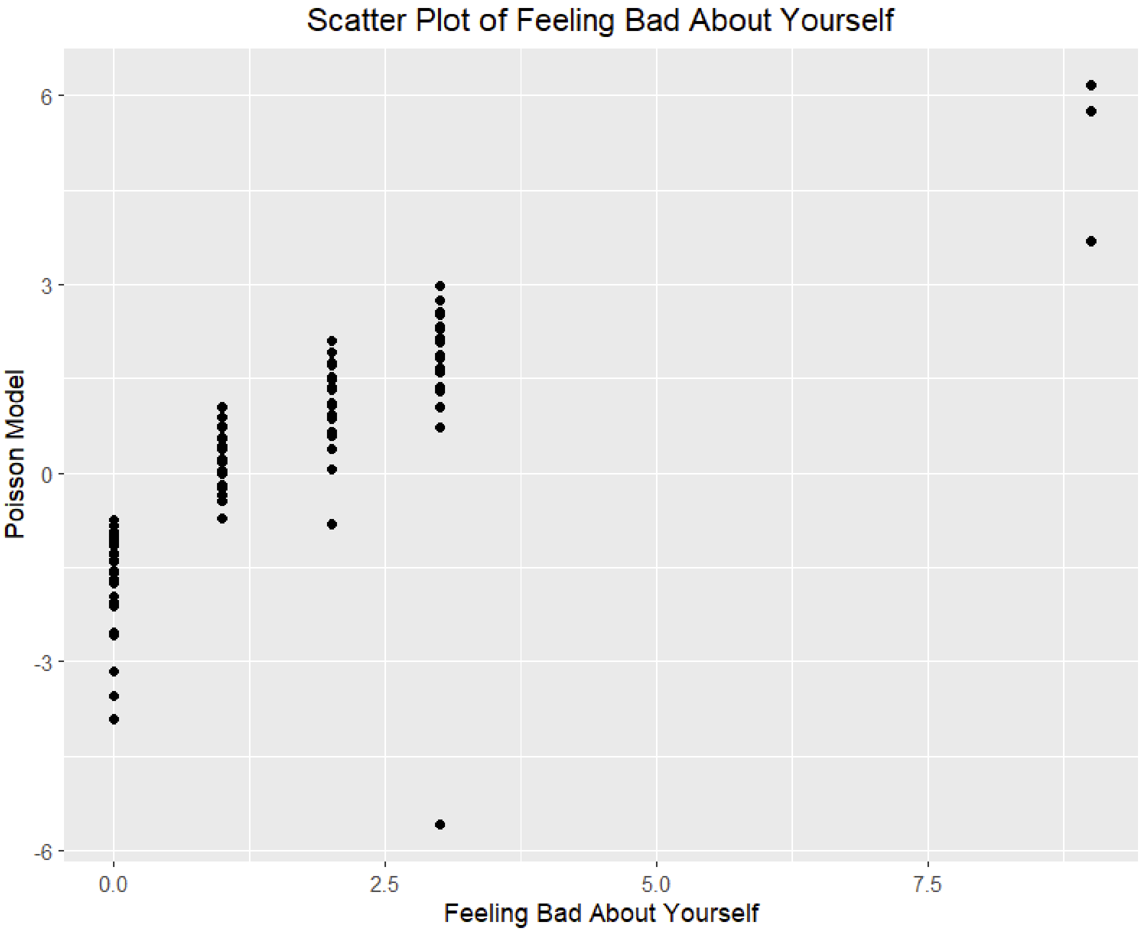


Figure 4: Scatter Plot of Poisson Model of Feeling Bad About Yourself

The over-dispersion test shows the Poisson model is a poor fit; therefore, we changed the model to Negative Binomial Regression.

The general form of the Negative Binomial Regression model is given as follows:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{88}$$

where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Depression) and X_{2i} (Trouble Sleeping) respectively.

Therefore, the fitted model would be:

$$\ln(\hat{\mu}) = -1.37535 + 0.31645 \text{ Depression}_i + 0.47472 \text{ Sleeping}_i \tag{89}$$

The Negative Binomial Regression model is presented in a scatter plot shown in Figure 5. Ideally, the residuals should be randomly dispersed around the horizontal line representing zero, indicating that the model fits well across all levels of the "Feeling" variable. In the scatter plot, the residuals seem to be symmetrically distributed around zero. Still, the clustering at fixed intervals on the x-axis indicates that "Feeling Bad About Yourself" is not a continuous variable. The spread of residuals does not show a clear pattern indicating systematic model error for "Feeling Bad About Yourself", but the presence of outliers, particularly for higher values of "Feeling Bad About Yourself," could warrant further investigation.

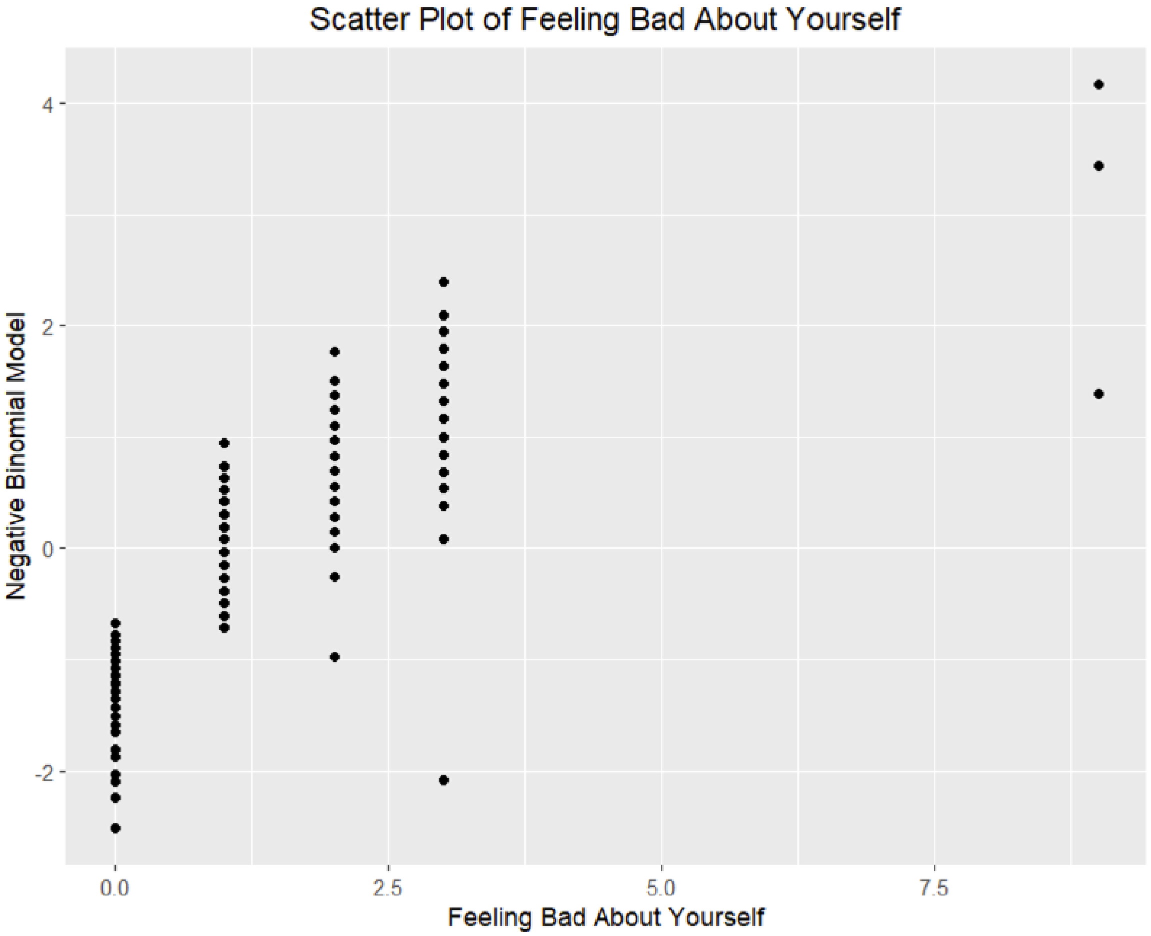


Figure 5: Scatter Plot of Negative Binomial Regression Model of Feeling Bad About Yourself

The statistical measures obtained from the analysis of "Feeling Bad About Yourself" are shown in Table 14 below. These measures are visualized graphically below, and colored vertical lines represent the best model.

Table 14: Feeling Bad About Yourself

	Statistical Measures			
Models	AIC	BIC	G^2	MSE
Log-Linear	8347.642	8373.839	935.4	0.615654
Multinomial	7711.33	7750.624	715.439	0.07756587
GLM(Poisson)	8400.514	8420.161	880.5	1.384698
GLM (NB)	8113.765	8139.961	655.3	1.069236

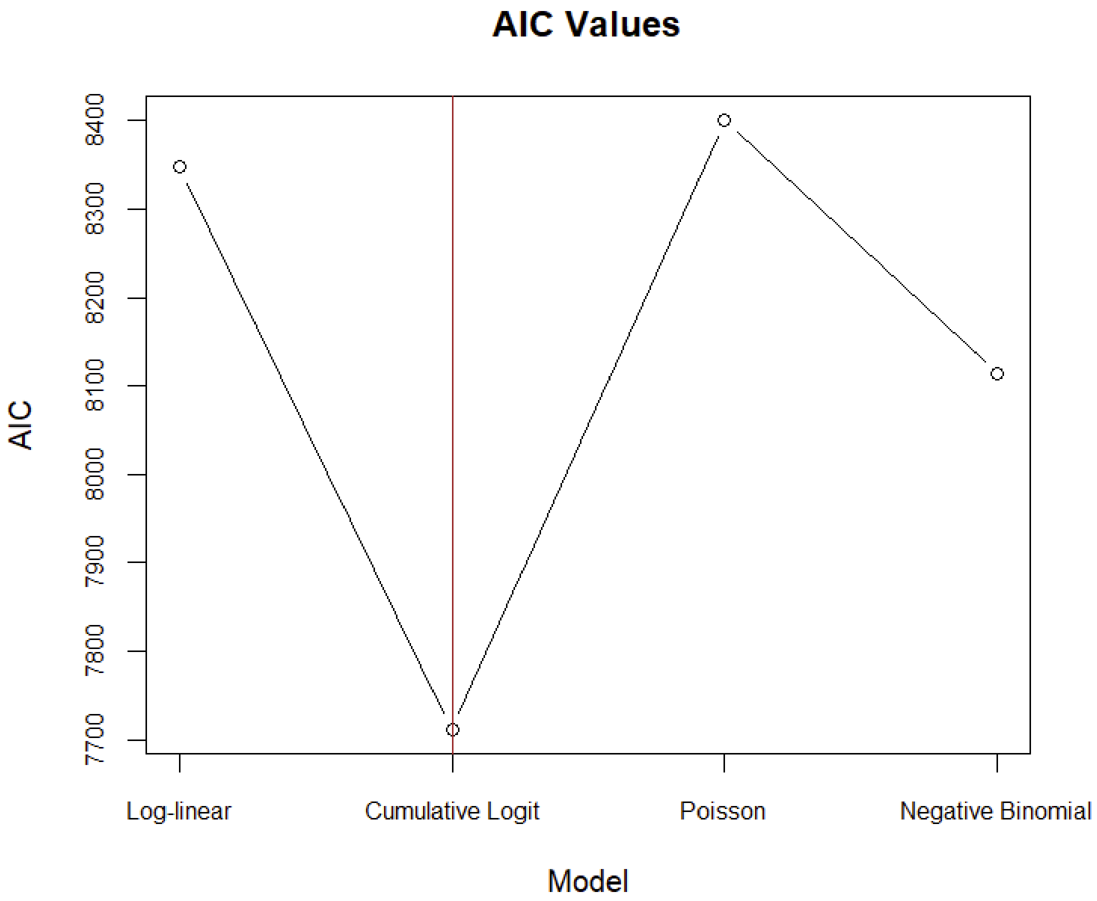


Figure 6: Graph of AIC values for "Feeling Bad About Yourself".

Figure 6 above shows the graph of AIC values for "Feeling Bad About Yourself," the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

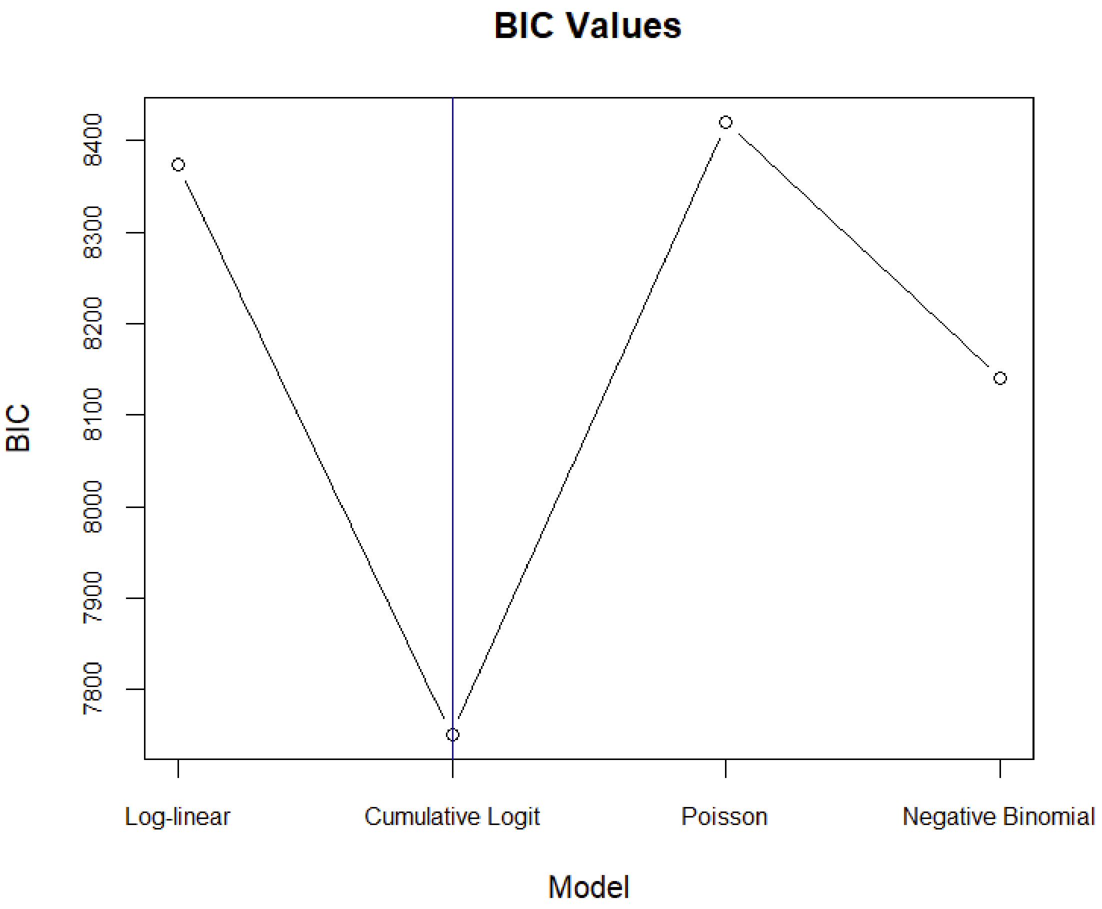


Figure 7: Graph of BIC values for “Feeling Bad About Yourself”.

Figure 7 above shows the graph of BIC values for “Feeling Bad About Yourself,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

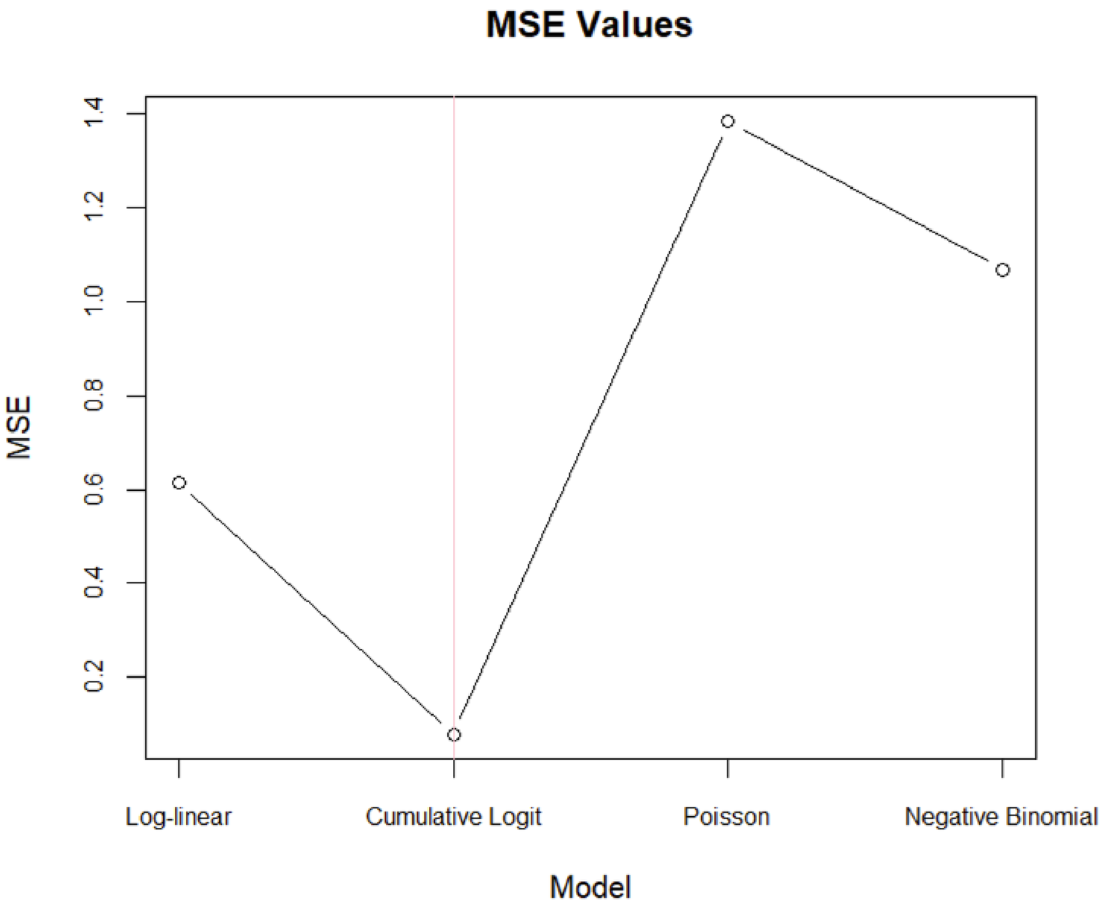


Figure 8: Graph of MSE values for “Feeling Bad About Yourself”.

Figure 8 above shows the graph of MSE values for “Feeling Bad About Yourself,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

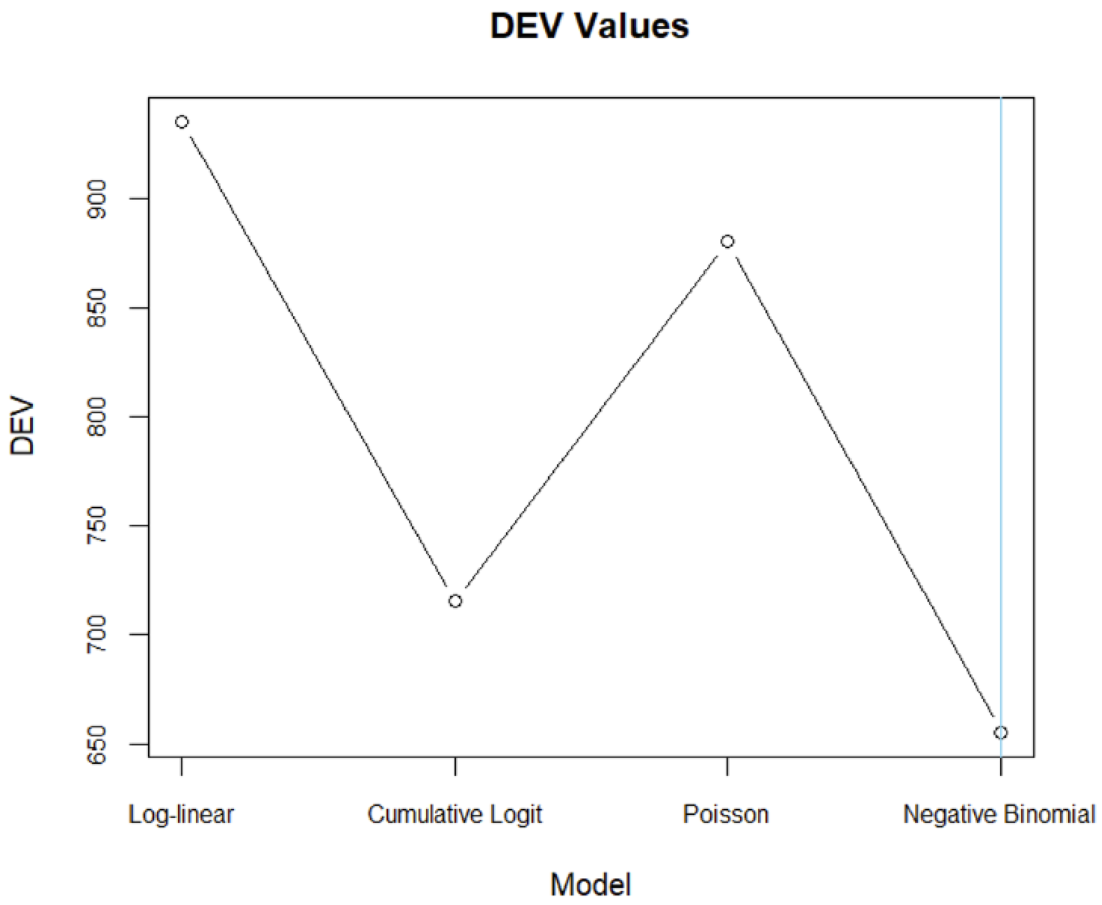


Figure 9: Graph of Deviance Statistic for “Feeling Bad About Yourself”.

Figure 9 above shows the graph of deviance statistic values for “Feeling Bad About Yourself,” the colored vertical line depicts that the Negative Binomial Regression Model is the best model for the analysis.

The graphs of AIC, BIC, and MSE values depict that the cumulative logit model consistently has the lowest value among other models. This finding suggests that the cumulative logit model is the best model to describe the relationship between feeling bad about yourself, depression, and trouble sleeping. On the other hand, the graph of Deviance Statistic value shows that the negative binomial model has the lowest value among other models. This finding indicates that the negative binomial model is the best model to describe the relationship between the variables.

4.4 Analysis of Depression

The analysis of depression is figured out through two variables, which are feeling bad about yourself and having trouble sleeping.

The mosaic plot in Figure 10 is drawn for this data, and we would like to highlight if too many trouble-sleeping people feel bad about themselves. The plot depicts relationships between "Feeling Bad About Yourself", "Trouble Sleeping," and "Depression." "Trouble Sleeping" levels are shown on the vertical axis, and "Feeling Bad About Yourself" levels are shown on the horizontal axis, ranging from 0 to 3. On the right side of the plot, there is a vertical bar labeled "Pearson residuals" with a scale from $-$ 12 to 69 which provide a visual presentation of the deviations of observed cell frequencies from the frequencies expected under the assumption of independence between "Feeling Bad About Yourself" and "Trouble Sleeping". Large positive residuals greater than 4 can be found for "Not At All" levels of "Trouble Sleeping" and "Feeling Bad About Yourself" and are colored in green. These positive residuals (4 to 69) indicate that the observed frequency is higher than expected under the independence assumption. On the other hand, there are large negative residuals, which are less than 4 for "Several Days of "Feeling Bad About Yourself" or "Several Days of Trouble Sleeping," colored in orange. These negative residuals indicate that the observed frequency is lower than expected under the independence assumption.

Residuals between -4 and 4, shaded in gray, indicate that the observed frequency is close to the expected frequency under the independence assumption. The p-value less than $2.22e^{-}16$ represents the probability of observing deviations from independence computed from a chi-square distribution with a degree of freedom 9. The degree of freedom of the chi-square (χ^2) test of independence is calculated as follows:

$$\begin{aligned} df &= (I-1)(J-1) \\ &= (4-1)(4-1)) \\ &= 9, \end{aligned}$$

where df is the degree of freedom, I is the number of levels of "Feeling Bad About Yourself" and J is the number of levels of "Trouble Sleeping".

The association between "Feeling Bad About Yourself" and "Trouble Sleeping" will be evaluated statistically with the p-value from the chi-square distribution in the mosaic plot. The associated hypotheses for the chi-square distribution are:

Null Hypothesis (H_0): There is no association between feeling bad about yourself and trouble sleeping.

Alternative Hypothesis (H_a): There is an association between feeling bad about yourself and trouble sleeping.

The p-value is small enough to reject the null hypothesis of independence and conclude that there is an association between feeling bad about yourself and trouble sleeping.

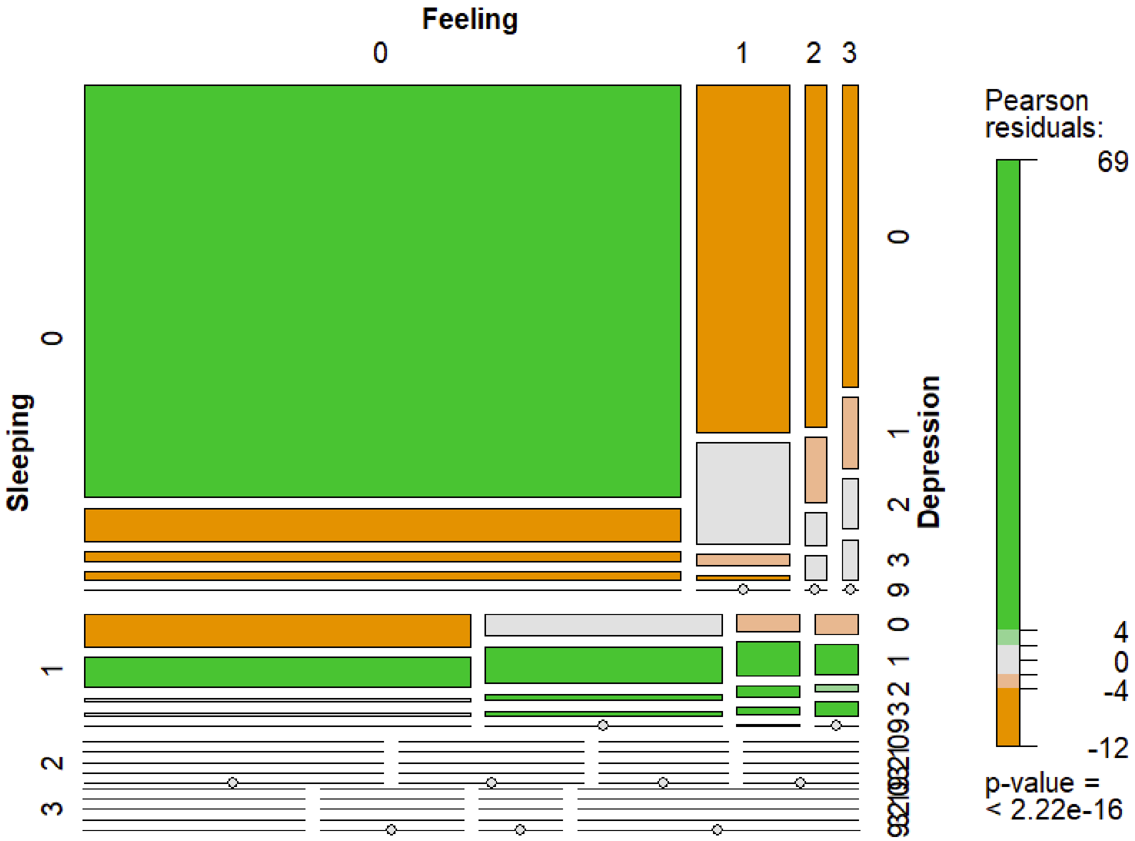


Figure 10: The mosaic plot of depression

The conditional contingency tables and marginal contingency tables constructed for each level of feeling bad about yourself and trouble sleeping are shown below. Tables 15, 16, 17, and 18 display the conditional contingency table between ”Feeling Bad About Yourself” and ”Trouble Sleeping” for each level of ”Depression.”

Table 15 displays the conditional contingency table between “Feeling Bad About Yourself” and “Trouble Sleeping” for the “Not At All” level of “Depression.”

Table 15: Conditional contingency table for “Not At All” level of “Feeling Depressed”.

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	2851	374	83	59
Several	216	90	19	15
More than half	42	6	9	6
Nearly	19	3	1	9

Table 16 displays the conditional contingency table between ”Feeling Bad About Yourself” and ”Trouble Sleeping” for” Several Days” level of “Depression.”

Table 16: Conditional contingency table for “Several Days” level of “Feeling Depressed”.

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	233	109	16	14
Several	198	146	36	22
More than half	14	25	10	9
Nearly	9	4	1	5

Table 17 displays the conditional contingency table between “Feeling Bad About Yourself” and “Trouble Sleeping” for the “More Than Half The Days” level of “Depression.”

Table 17: Conditional contingency table for "More Than Half The Days" level of "Feeling Depressed".

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	65	12	8	10
Several	26	28	11	5
More than half	22	14	14	9
Nearly	8	8	8	7

Table 18 displays the conditional contingency table between "Feeling Bad About Yourself" and "Trouble Sleeping" for "Nearly Everyday" level of "Depression".

Table 18: Contingency conditional contingency table for "Nearly Every Days" level of "Feeling Depressed"

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	57	5	6	8
Several	20	18	8	10
More than half	8	8	4	9
Nearly	11	16	8	40

Since each category has four levels, we can calculate several odds ratios. For simplicity, one level of the "Depression" variable, which is more meaningful, is picked to compute the odds ratio. The calculation of the odds ratio between "Several Days" and "More Than Half The Days" levels of "Feeling Bad About Yourself" and "Trouble Sleeping" for "More Than Half The Days" level of "Feeling Down, Depressed and Hopeless" is done with the circled values in the conditional contingency table 19.

Table 19: Conditional contingency table for “More Than Half The Days” level of “Feeling Depressed” for odds ratio calculation.

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	65	12	8	10
Several	26	㉘	⑪	5
More than half	22	⑭	⑭	9
Nearly	8	8	8	7

The conditional probability $\theta_{XY|Z}$ can be calculated as follows:

$$\begin{aligned}
\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\
&= \theta_{XY|4} = \frac{n_{222}n_{322}}{n_{232}n_{332}} \\
&= \frac{28 \times 14}{14 \times 11} = \frac{28}{11} \\
&= 2.545
\end{aligned}$$

It means that for those who have trouble sleeping for several days, the odds of “Several Days of Feeling Bad About Yourself” will be 2.545 times more than those who have trouble sleeping for more than half the days due to more than half the days of feeling down, depressed, and hopeless.

Also, the calculation of the odds ratio between “Several Days” and “Nearly Everyday” levels of “Trouble Sleeping” and “Feeling Bad About Yourself” for “More Than Half The Days” level of “Feeling Down, Depressed, and Hopeless” is done with the circled values in the conditional contingency table 20.

Table 20: Conditional contingency table for “More Than Half The Days” level of “Feeling Depressed” for second odds ratio calculation.

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	65	12	8	10
Several	26	㉘	11	⑤
More than half	22	14	14	9
Nearly	8	⑧	8	⑦

The conditional probability $\theta_{XY|Z}$ can be calculated as follows:

$$\begin{aligned}\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\ &= \theta_{XY|4} = \frac{n_{222}n_{442}}{n_{242}n_{422}} \\ &= \frac{28 \times 7}{8 \times 5} = \frac{196}{40} \\ &= 4.90\end{aligned}$$

It means that for those who have trouble sleeping for several days, the odds of several Days of feeling bad about yourself” will be 4.90 times more than those who have trouble sleeping nearly every day due to more than half the days of feeling down, depressed, and hopeless. If the rest of the odds ratio is calculated, it is found that all odds ratios are greater than one. In other words, $\theta_{XY|4} = c > 1, \forall X, Y$, where c is a constant.

We conclude that trouble sleeping has a greater effect on mental health than feeling bad about yourself, which are independent of each other condition on “more than half the days feeling down, depressed, and hopeless. ”.

We will determine whether there is an increasing or decreasing trend between the levels of depression and trouble sleeping. The associated hypotheses for the linear trend are stated as follows:

Null Hypothesis (H_0): There is no trend between depression and trouble sleeping.

Alternative Hypothesis (H_a): A positive linear trend exists between depression and trouble sleeping.

The test statistic of the linear trend is calculated as follows:

$$\begin{aligned} Q &= \sqrt{((n_{++} - 1)r^2)} \\ &= \sqrt{((255 - 1)(1)^2)} \\ &\approx 15.937, \end{aligned}$$

where n_{++} is the total number of counts in the conditional contingency table for "More Than The Days" level of "Feeling Down, Depressed, and Hopeless" and r is the weighted correlation between "Feeling Bad About Yourself" and "Trouble Sleeping."

The linear trend test statistic value is compared with the critical value of chi-square with degree of freedom 3 ($\chi^2(3) = 7.815$). The degree of freedom of chi-square is calculated as follows:

$$\begin{aligned} df &= k - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where df is the degree of freedom and k is the number of levels of "Feeling Down, Depressed, and Hopeless.".

The comparison of the test statistic value to the critical value $\chi^2(3)$ suggests the null hypothesis (H_0) should be rejected since the test statistic value is greater than the critical value. The data provide sufficient evidence to conclude there is a positive linear association between feeling bad about yourself and trouble sleeping due to more than half the days feeling down, depressed, and hopeless. This test must be appropriate only if researchers suspect a positive linear association before seeing data.

Then, the marginal contingency table that presents the total counts across all levels of "Depression" is shown below in Table 21.

Table 21: Marginal contingency table between "Feeling Bad About Yourself" and "Trouble Sleeping".

	Feeling Bad About Yourself			
Trouble Sleeping	Not at all	Several	More than half	Nearly
Not at all	3209	500	113	91
Several	461	282	75	52
More than half	86	53	37	33
Nearly	48	31	18	61

The odds ratio between "Several Days" and "More Than Half The Days" levels of "Trouble Sleeping" and "Feeling Bad About Yourself" would be calculated.

The joint probability (θ_{XY}) between "Feeling Bad About Yourself" (X) and "Trouble Sleeping" (Y) can be calculated as follows:

$$\begin{aligned}\theta_{XY} &= \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}} \\ &= \frac{282 \times 37}{53 \times 75} \\ &= \frac{10434}{3975}\end{aligned}$$

$$= 2.625$$

It means that for those who have trouble sleeping for several days, the odds of several days of feeling bad about yourself will be 2.63 times more than those who have trouble sleeping for more than half the days of feeling down, depressed, and hopeless. If we calculate the odds for the marginal table, we will find that all are greater than 1. Therefore, we would conclude that the odds of success are greater for those with trouble sleeping.

Finally, a test to evaluate the conditional independence (conditional on "Depression" levels) is performed. An appropriate test would be the Cochran-Mantel-Haenszel (C M H)Test of conditional independence [28].

The associated hypotheses of the Cochran-Mantel-Haenszel Test are stated as follows:

Null Hypothesis (H_0): Feeling bad about yourself and Trouble sleeping are conditionally independent.

Alternative Hypothesis (H_a): Feeling bad about yourself and Trouble sleeping are not conditionally independent.

The Cochran-Mantel-Haenszel (C M H) χ^2 statistic is approximately 280.35, which is greater than any $\chi^2(9)$ critical value of the degree of freedom 9 ($\chi^2(9) = 16.919$). The degree of freedom of the chi-square test of conditional independence is calculated as follows:

$$
\begin{aligned}
df &= (I-1)(J-1) \\
&= (4-1)(4-1)) \\
&= 9,
\end{aligned}
$$

where *df* is the degree of freedom, I is the number of levels of "Feeling Bad About Yourself" and J is the number of levels of "Trouble Sleeping."

We would reject the null hypothesis (H_0) since the test statistic value is greater than the critical value and conclude that "Feeling Bad About Yourself" and "Trouble Sleeping" are not conditionally independent of "Depression" levels. For at least one "Depression" level, there is a significant association between trouble sleeping and feeling bad about yourself.

We will investigate if there is any homogeneous association between all two-factor interaction terms, i.e., XY (Sleeping × Feeling). The appropriate model would be the Log-linear Homogeneous Association Model (M_0). The general form of the Log-linear Homogeneous Association Model for this data is given by:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \forall i = 1,2,3,4, j = 1,2,3,4. \tag{90}$$

The fitted model can be represented as follows:

$$\ln(\mu_{ijk}) = -1.42978 + 0.63156_i^{\text{Sleeping}} + 0.36246_j^{\text{Feeling}} - 0.07570_{ij}^{\text{Sleeping} \times \text{Feeling}} \tag{91}$$

To perform a test to assess the overall fit of this model, the deviance statistic, $G^2(M_0)$, is calculated by finding the difference between the null deviance, which is the saturated or full model, and the residual deviance, which is the reduced model.

The associated hypotheses of the deviance statistics are stated as follows:

Null Hypothesis (H_0): Extra model parameters are zero (not significant).

Alternative Hypothesis (H_a): Extra model parameters are non-zero
(at least one is significant).

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 6442.9 - 5062.3 \\ &= 1380.6 \end{aligned}$$

The degree of freedom of the chi-square test of the Log-linear Homogeneous Association Model is calculated as follows:

$$\begin{aligned} df &= null_{df} - residual_{df} \\ &= 5161 - 5158 \\ &= 3, \end{aligned}$$

where df is the degree of freedom, $null_{df}$ is the degree of freedom of the null deviance and $residual_{df}$ is the degree of freedom of the residual deviance.

By comparing the test value to the critical value ($\chi^2_{0.05}(3) = 7.815$), since the test value is greater than the critical value, we would reject the null hypothesis. Therefore, the data provide sufficient evidence to conclude the homogeneous association model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, sleeping, and feeling bad about ourselves.

For the Multinomial Logistic Regression Model, the model can be written mathematically as follows:

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \quad ; \quad j = 0, 1, 2$$

So,

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 \text{ Sleeping}_i + \beta_2 \text{ Feeling}_i$$

More specifically, the model can be expanded as follows:

$$\begin{aligned}
\log\left(\frac{P_0}{P_1 + P_2 + P_3}\right) &= 1.7994 + 1.264 \text{ Sleeping}_i + 0.495 \text{ Feeling}_i, \\
\log\left(\frac{P_0 + P_1}{P_2 + P_3}\right) &= 3.3841 + 1.264 \text{ Sleeping}_i + 0.495 \text{ Feeling}_i, and \\
\log\left(\frac{P_0 + P_1 + P_2}{P_3}\right) &= 4.3852 + 1.264 \text{ Sleeping}_i + 0.495 \text{ Feeling}_i.
\end{aligned}$$

To perform a test to assess the overall fit of this model, the deviance statistics, $G^2(M_0)$ is calculated and compared to the chi-square critical value. The associated hypotheses of the deviance statistic are stated as follows:

Null Hypothesis (H_0): Model fit is good.

Alternative Hypothesis (H_a): Model fit is not good.

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned}
G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\
&= 8557.117 - 7214.218 \\
&= 1342.899
\end{aligned}$$

The degree of freedom of the chi-square test of the Cumulative Logit Model is calculated as follows:

$$\begin{aligned} df &= k - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where *df* is the degree of freedom and k is the number of levels of “Feeling Down, Depressed, and Hopeless”.

By comparing the test value to the critical value ($\chi^2_{0.05}(3) = 7.815$), we would reject the null hypothesis (H_0) since the test value is greater than the critical value. Therefore, the data provide sufficient evidence to conclude the cumulative logit model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, sleeping, and feeling bad about ourselves.

We will model the response variable, ”Depression,” with generalized linear models (GLMs). Since the data’s response variable is a count variable, we will use the Poisson regression and the negative binomial regression.

The general form of the Poisson model is given by:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{92}$$

where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Trouble Sleeping) and X_{2i} (Feeling Bad About Yourself) respectively.

So, the fitted model will be represented as follows:

$$\ln(\hat{\mu}) = -1.36237 + 0.54893 \text{ Sleeping}_i + 0.25197 \text{ Feeling}_i \tag{93}$$

The interpretation of the coefficients will be made in two ways, general and specific interpretations.

The general interpretation is as follows:
"There is an expected increase in the mean count of people feeling down, depressed, and hopeless.".

The specific interpretations are given as follows:
"The number of times people feel depressed with trouble sleeping is expected to be $e^{0.54893} \approx 1.73$ times, feeling bad about yourself is constant", and "the number of times people feel depressed with feeling bad about themselves is expected to be $e^{0.25197} \approx 1.29$ times, depression is constant".

The histogram of the data is right-skewed, shown in Figure 11 below:

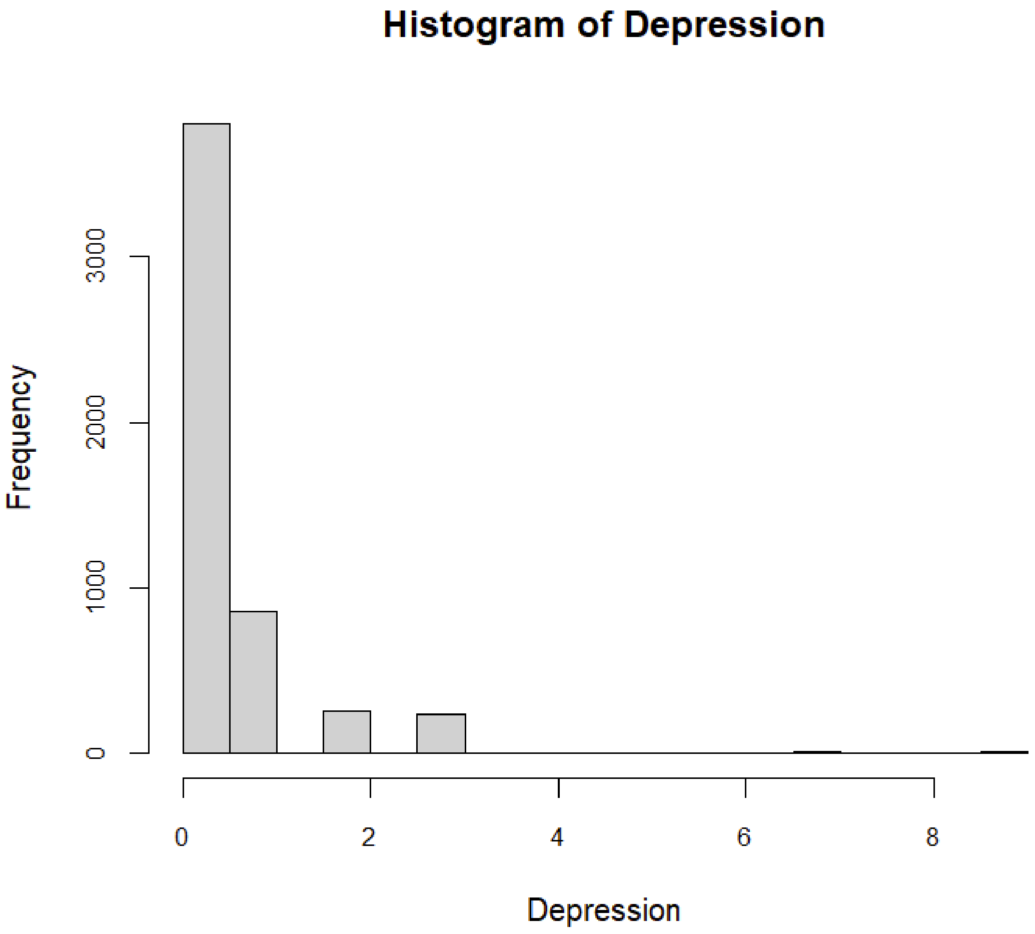


Figure 11: The histogram of depression

The over-dispersion test is used to check if the variance is equal to the mean. The associated hypotheses of the over-dispersion test are stated as:

Null Hypothesis (H_0): variance = mean

Alternative Hypothesis (H_a): variance > mean

The dispersion estimate ($\hat{\phi}$) = 1.624339, greater than 1. The p-value for the over-dispersion test is 8.788×10^{-9}, which is less than $\alpha = 0.05$. The Variance, equal to 0.7530648, is greater than the mean, equal to 0.4213483. Therefore, the null hypothesis is rejected, and we conclude that the variance is greater than the mean.

The plot in Figure 12 depicts the scatter plot of the Poisson Model of “Depression”.If the Poisson model fits well, we expect to see the residuals dispersed randomly around the zero line on the y-axis without any clear pattern. The presence of clusters of points vertically aligned at specific values of “Depression” implies several unique levels of this variable. Residuals that are far from zero, especially if they form a pattern, may indicate potential issues with the model’s fit at certain levels of depression. There are some potential outliers, especially at higher levels of depression, which have higher deviance residuals. Suppose these outliers are not random but correspond to specific levels of the depression variable. In that case, it might suggest that the Poisson model does not capture all the nuances of the relationship between depression and the outcome variable.

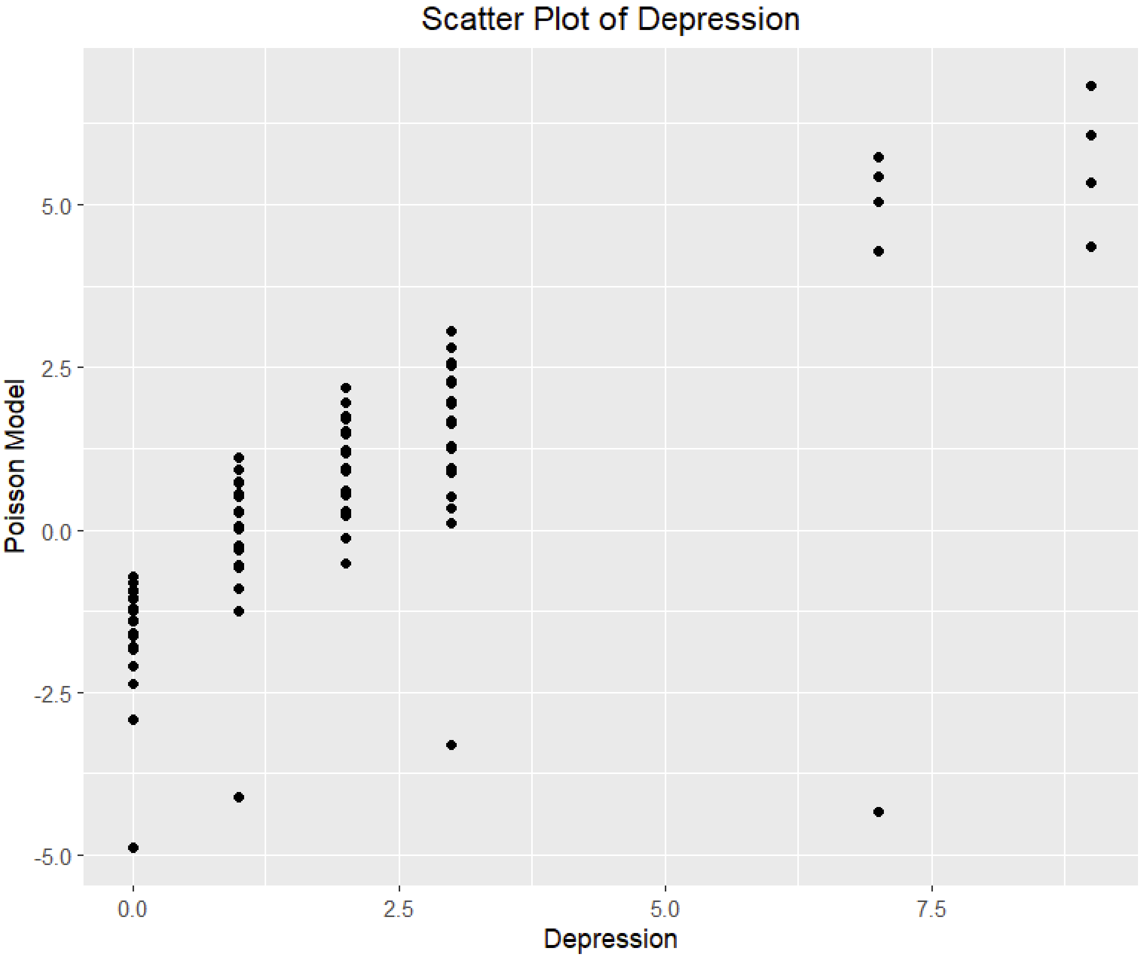


Figure 12: Scatter Plot of Poisson Regression Model of Depression

The overdispersion test shows the model is a poor fit. Therefore, we changed the model to Negative Binomial Regression.

The general form of the Negative Binomial Regression model is:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{94}$$

where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Trouble Sleeping) and X_{2i} (Feeling Bad

About Yourself) respectively.

$$\ln(\hat{\mu}) = -1.47999 + 0.69791 \text{ Sleeping}_i + 0.27767 \text{ Feeling}_i \tag{95}$$

From the scatter plot of the negative binomial model in Figure 13, the residuals do not show a clear pattern, which might initially indicate that the model does not have systematic bias. However, the presence of outliers, particularly at higher depression levels, indicates instances where the model's predictions deviate significantly from the observed data.

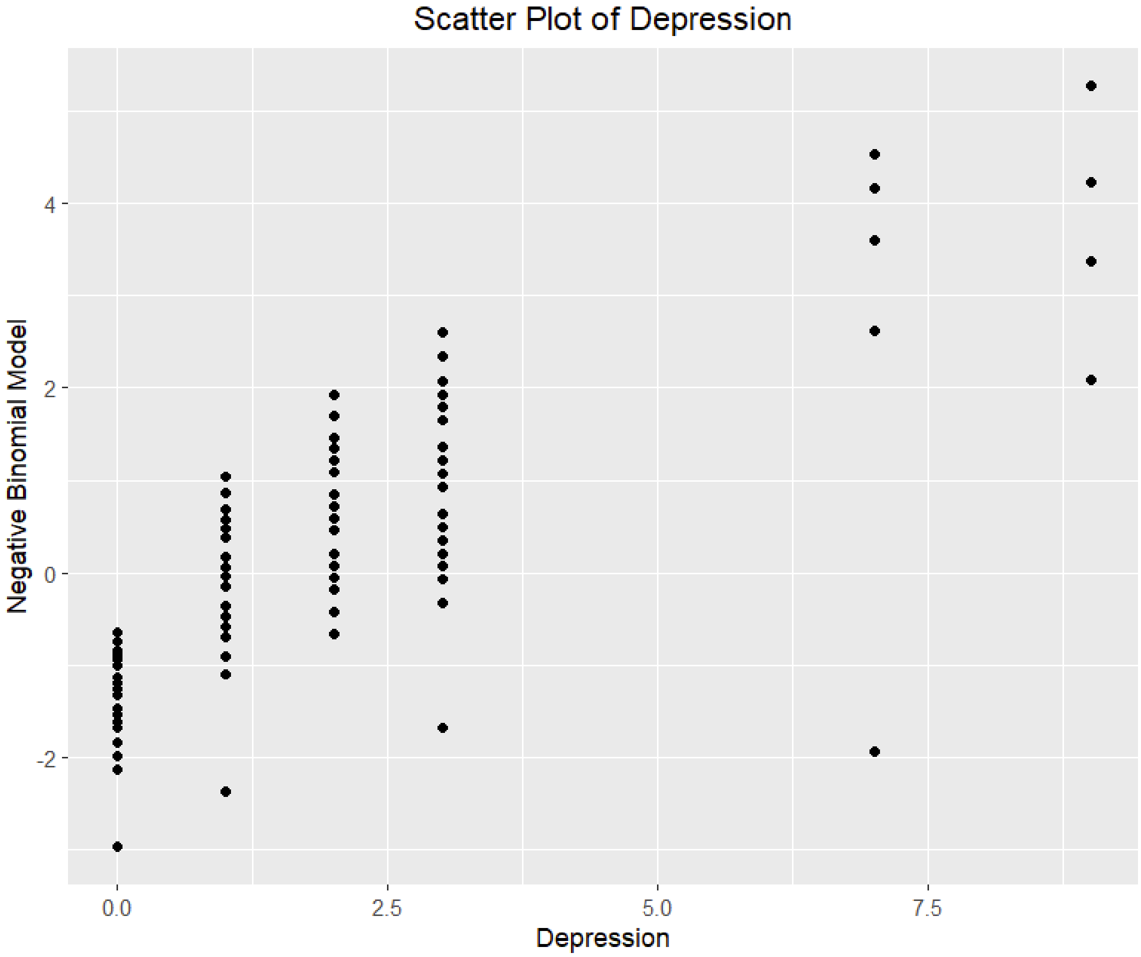


Figure 13: Scatter Plot of Negative Binomial Regression Model of Depression

The statistical measures obtained from the analysis of “Depression” are shown in Table 22. These measures are visualized graphically below, and colored vertical lines represent the best model.

Table 22: Statistical measures of “Feeling Down, Depressed, and Hopeless”.

	Statistical Measures			
Models	AIC	BIC	G^2	MSE
Log-Linear	8152.288	8178.484	1431.3	0.7365462
Multinomial	7228.218	7274.062	1342.899	0.05878926
GLM(Poisson)	8201.003	8220.65	1380.6	1.542902
GLM (NB)	7947.697	7973.894	1064.6	1.324138

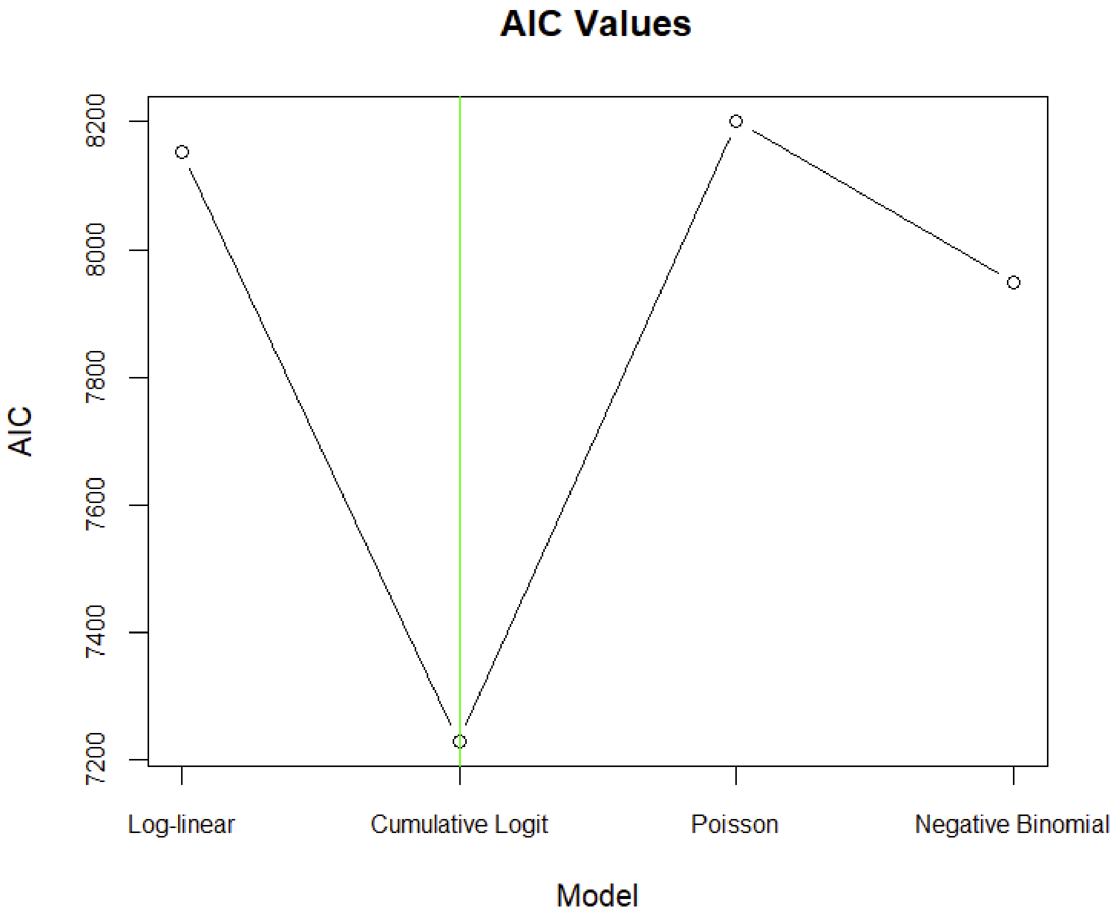


Figure 14: Graph of AIC values for “Depression”.

Figure 14 above shows the graph of AIC values for “Depression,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

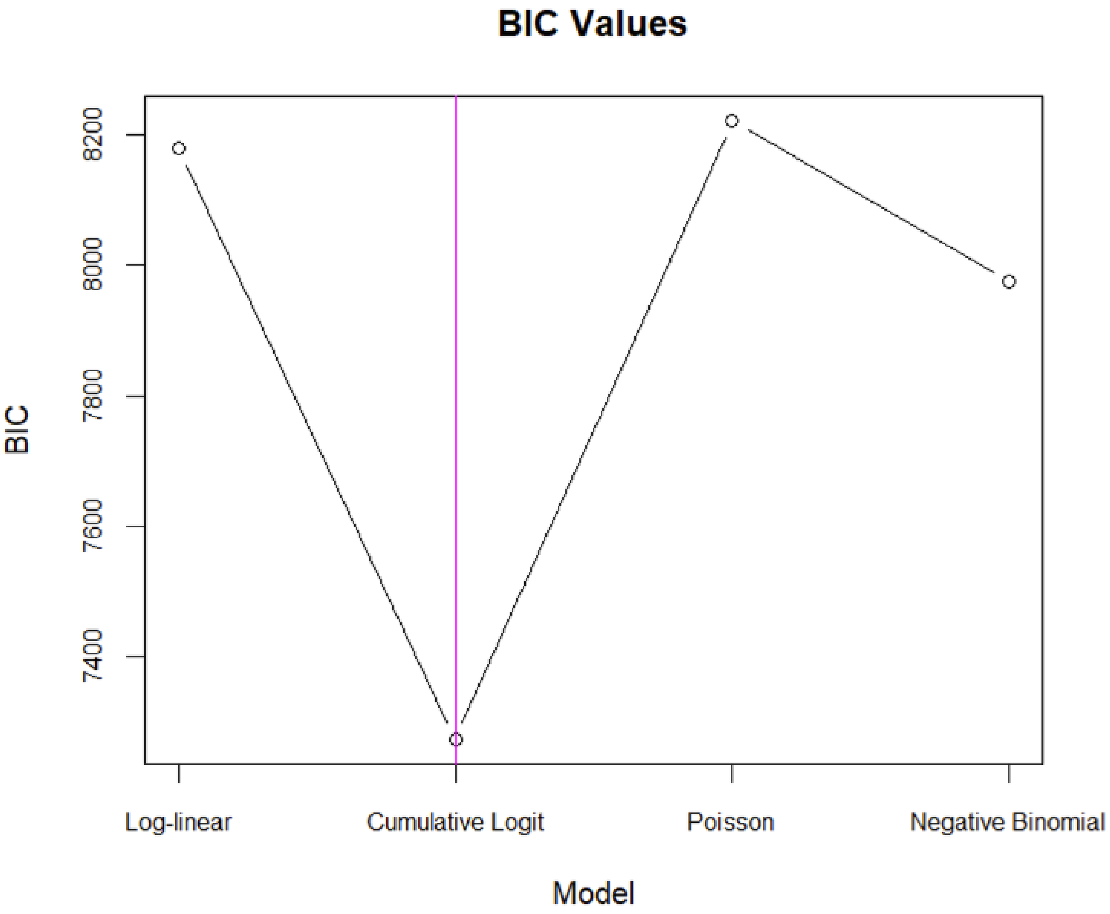


Figure 15: Graph of BIC values for "Depression".

Figure 15 above shows the graph of BIC values for "Depression," the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

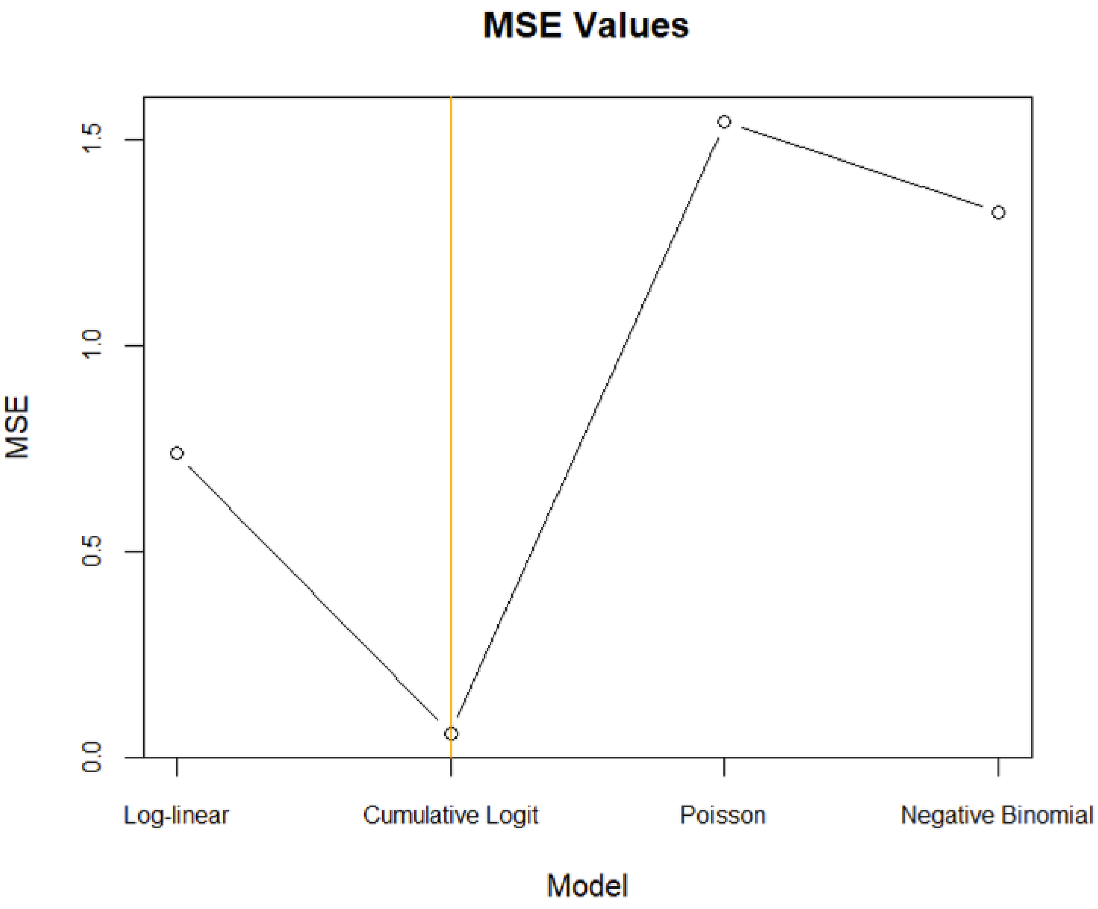


Figure 16: Graph of MSE values for “Depression”.

Figure 16 above shows the graph of MSE values for “Depression,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

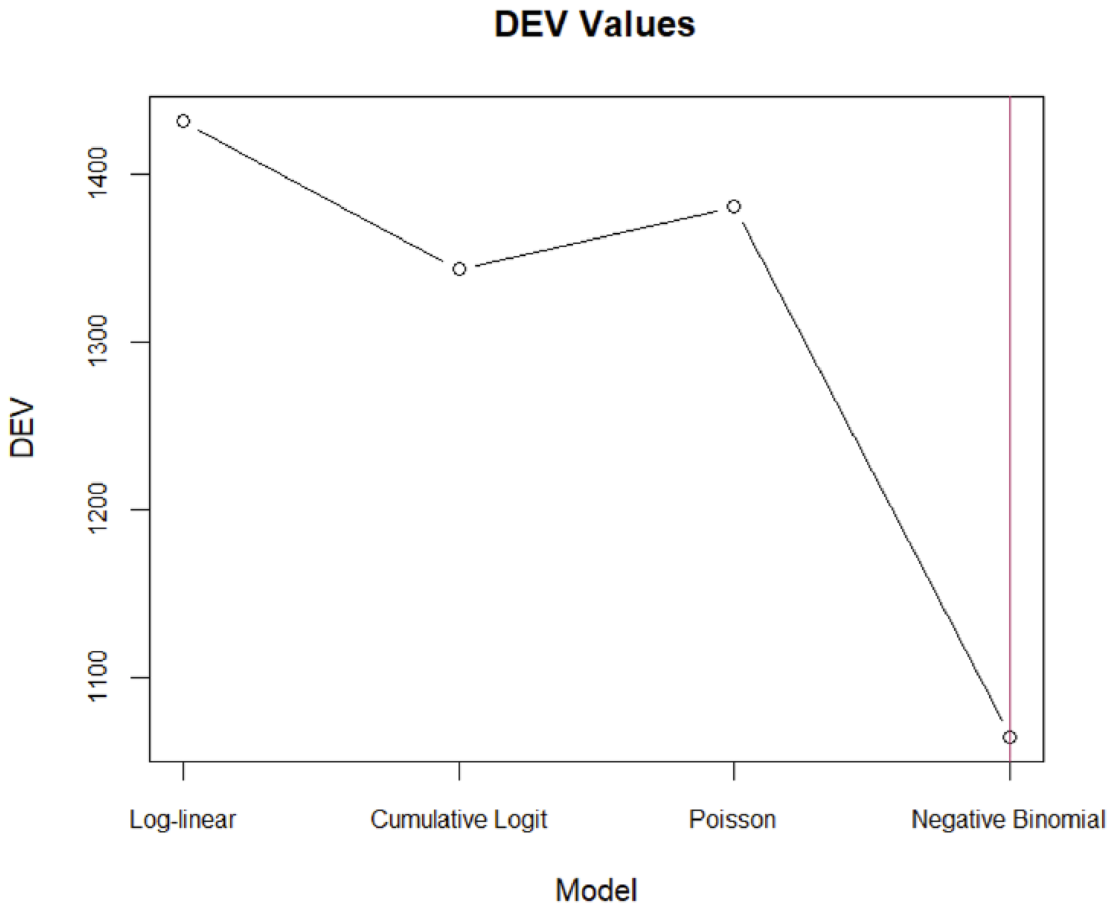


Figure 17: Graph of Deviance Statistics for “Depression”.

Figure 17 above shows the graph of Deviance Statistics values for “Depression,” and the colored vertical line depicts that the Negative Binomial Regression Model is the best model for the analysis.

The graphs of AIC, BIC, and MSE values depict that the cumulative logit model has the lowest value among other models, which shows that it is the best model for describing the relationship between the three variables. The graph of the Deviation Statistics value, on the other hand, depicts that the negative binomial model has the lowest value among other models, which shows that it is the best model to describe the relationship between the variables.

4.5 Analysis of Trouble Sleeping

The analysis of trouble sleeping is figured out through two variables: feeling bad about yourself and depression.

The mosaic plot in Figure 18 is drawn for this data, and we would like to highlight if there are too many depressed people who feel bad about themselves. The plot depicts relationships between “Feeling Bad About Yourself”, “Depression,” and “Trouble Sleeping.” “Depression” levels are shown on the vertical axis, and “Feeling Bad About Yourself” levels are shown on the horizontal axis, ranging from 0 to 3. On the right side of the plot, there is a vertical bar labeled “Pearson residuals” with a scale from − 12 to 69 which provide a visual presentation of the deviations of observed cell frequencies from the frequencies expected under the assumption of independence between “Feeling Bad About Yourself” and “Depression.” Large positive residuals greater than 4 can be found for “Not At All” levels of “Feeling Bad About Yourself” and “Depression” and are colored in green. These positive residuals (4 to 69) indicate that the observed frequency is higher than expected under the independence assumption. On the other hand, there are large negative residuals, which are less than 4 for “Several Days of “Feeling Bad About Yourself” or “Several Days of Depression,” colored in orange. These negative residuals indicate that the observed frequency is lower than expected under the independence assumption. Residuals between − 4 and 4, shaded in gray, indicate that the observed frequency is close to the expected frequency under the independence assumption. The p-value, less than $2.22e^{-}16$, represents the probability of observing deviations from independence computed from a chi-square distribution with a degree of freedom 9. The degree of freedom of the chi-square (χ^2) test of independence is calculated as follows:

$$\begin{aligned} df &= (I-1)(J-1) \\ &= (4-1)(4-1)) \\ &= 9, \end{aligned}$$

where df is the degree of freedom, I is the number of levels of “Feeling Bad About Yourself” and J is the number of levels of “Depression”.

The association between “Feeling Bad About Yourself” and “Depression” will be evaluated statistically with the p-value from the chi-square distribution in the mosaic plot. The associated hypotheses of a chi-square distribution are stated:

Null Hypothesis (H_0): There is no association between feeling bad about yourself and depression.

Alternative Hypothesis (H_a): There is an association between feeling bad about yourself and depression.

The p-value is small enough to reject the null hypothesis of independence and conclude that there is an association between feeling bad about yourself and depression.

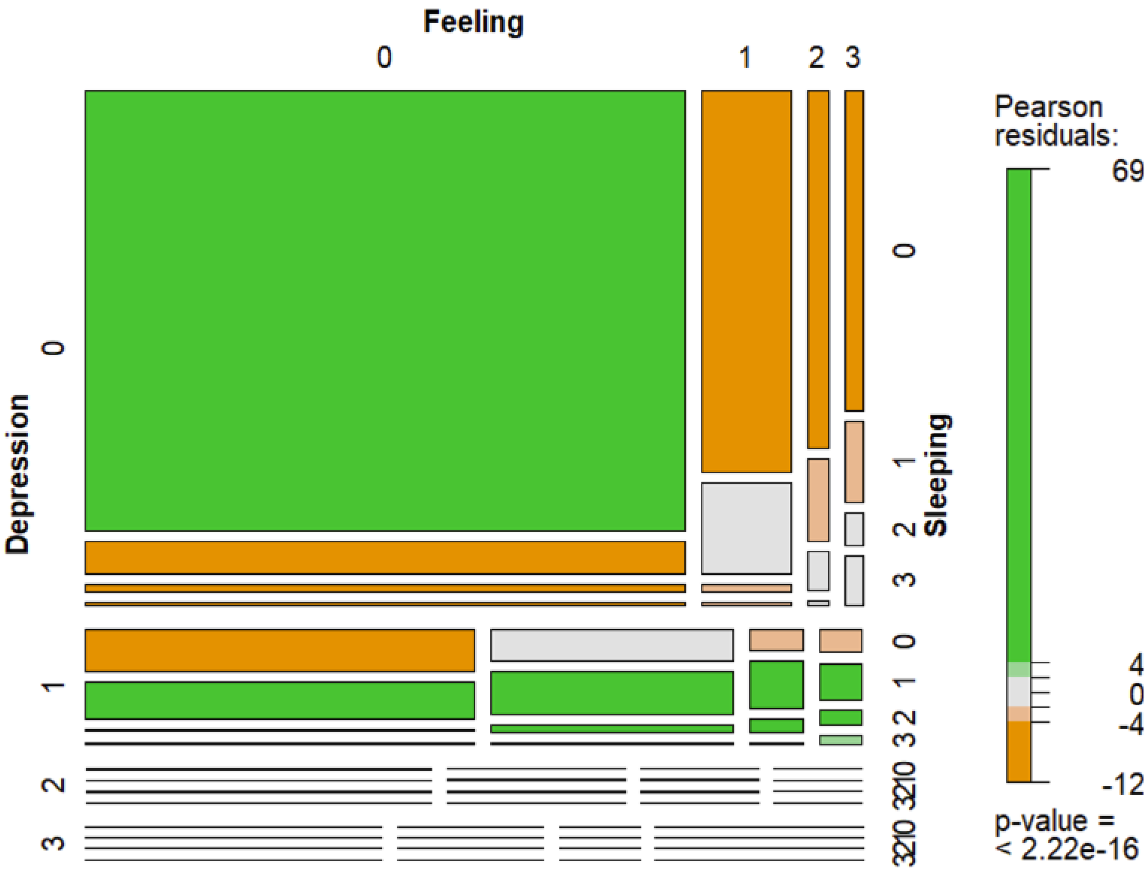


Figure 18: The mosaic plot of trouble sleeping

The Conditional contingency tables and marginal contingency table constructed for each level of feeling bad about yourself and depression are as follows. Tables 23,24,25, and 26 display the conditional contingency table between "Feeling Bad About Yourself" and "Depression" for each level of "Trouble Sleeping."

Table 23 displays the conditional contingency table between "Feeling Bad About Yourself" and "Depression" for the "Not At All" level of "Trouble Sleeping."

Table 23: Conditional contingency table for “Not At All” level of “Trouble Sleeping”.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	2851	374	83	59
Several	233	109	16	14
More than half	65	12	8	10
Nearly	57	5	6	8

Table 24 displays the conditional contingency table between “Feeling Bad About Yourself” and “Depression” for the “Several Days” level of “Trouble Sleeping.”

Table 24: Conditional contingency table for “Several Days” level of “Trouble Sleeping”.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	216	90	19	15
Several	198	146	36	22
More than half	26	28	11	5
Nearly	20	18	8	10

Table 25 displays the conditional contingency table between “Feeling Bad About Yourself” and “Depression” for the “More Than Half The Days” level of “Trouble Sleeping.”

Table 25: Conditional contingency table for “More Than Half Days” level of “Trouble Sleeping”.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	42	6	9	6
Several	14	25	10	9
More than half	22	14	14	9
Nearly	8	8	4	9

Table 26 displays the conditional contingency table between “Feeling Bad About Yourself” and “Depression” for the “Nearly Everyday” level of “Trouble Sleeping.”

Table 26: Conditional contingency table for “Nearly Every Day” level of “Trouble Sleeping”.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	19	3	1	9
Several	9	4	1	5
More than half	8	8	8	7
Nearly	11	16	8	40

Since each category has four levels, we can calculate several odds ratios. For simplicity, one level of the “Sleeping” variable, which is more meaningful, is picked to compute the odds ratio. The calculation of the odds ratio between “Several Days” and “More Than Half The Days” levels of “Depression” and “Feeling Bad About Yourself” for “More Than Half The Days” of “Trouble Sleeping” is done with the circled values in the conditional contingency table 27.

Table 27: Conditional contingency table for "More Than Half The Days" level of "Trouble Sleeping" for odds ratio calculation.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	42	6	9	6
Several	14	25	10	9
More than half	22	14	14	9
Nearly	8	8	4	9

The conditional probability $\theta_{XY|Z}$ between the "Several Days" and the "More Than Half The Days" levels of "Feeling Bad About Yourself" and "Depression" can be calculated as follows:

$$\begin{aligned}
\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\
= \theta_{XY|4} &= \frac{n_{222}n_{322}}{n_{232}n_{332}} \\
&= \frac{25 \times 14}{14 \times 10} = \frac{25}{10} \\
&= 2.50
\end{aligned}$$

It means that for those who have had depression for several days, the odds of several days of feeling bad about yourself will be 2.50 times more than those who have had depression for more than half the days due to more than half the days of trouble sleeping.

Also, the calculation of the odds ratio between "Several Days" and "Nearly Everyday" levels of "Depression" and "Feeling Bad About Yourself" for "More Than Half

Days" level of "Trouble Sleeping" is done with the circled values in the conditional contingency table 28.

Table 28: Conditional contingency table for "More Than Half The Days" level of "Trouble Sleeping" for second odds ratio calculation.

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	42	6	9	6
Several	14	㉕	10	⑨
More than half	22	14	14	9
Nearly	8	⑧	4	⑨

The Conditional Probability $\theta_{XY|Z}$ between "Several Days" and "Nearly Everyday" levels of "Feeling Bad About Yourself" and "Trouble Sleeping" is calculated as follows:

$$\begin{aligned}\theta_{XY|Z} = \theta_{XY|k} &= \frac{n_{ijk}n_{i'j'k}}{n_{ij'k}n_{i'jk}} \\ &= \theta_{XY|4} = \frac{n_{222}n_{442}}{n_{242}n_{422}} \\ &= \frac{25 \times 9}{8 \times 9} = \frac{25}{8} \\ &= 3.125\end{aligned}$$

It means that for those who have depression for several days, the odds of several days of feeling bad about themselves will be 3.125 times more than those who have depression nearly every day due to more than half the days of trouble sleeping. If the rest of the odds ratio is calculated, it is found that all odd ratios are greater than one. In other words, $\theta_{XY|4} = c > 1$, $\forall X, Y$, where c is a constant.

We conclude that depression has a greater effect on mental health than feeling bad about yourself, which are independent of each other condition on "more than half the day's trouble sleeping."

We will determine whether there is an increasing or decreasing trend between the levels of depression and trouble sleeping. The associated hypotheses of the linear trend are stated as follows:

Null Hypothesis (H_0): There is no trend between feeling bad about yourself and depression.

Alternative Hypothesis (H_a): There is a positive linear trend between feeling bad about yourself and depression.

The test statistic of the linear trend is calculated as follows:

$$
\begin{aligned}
\mathrm{Q} &= \sqrt{((n_{++} - 1)r^2)} \\
&= \sqrt{((209 - 1)(1)^2)} \\
&\approx 14.422,
\end{aligned}
$$

where n_{++} is the total number of counts in the conditional contingency table for "More Than The Days" level of "Trouble Sleeping" and r is the weighted correlation between "Feeling Bad About Yourself" and "Depression".

The linear trend test statistic value is compared with the critical value of chi-square with degree of freedom 3 ($\chi^2(3) = 7.815$). The degree of freedom of chi-square is calculated as follows:

$$\begin{aligned} \text{df} &= k - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where *df* is the degree of freedom, and k is the number of levels of "Trouble Sleeping".

The linear trend test is a crucial part of our analysis. The comparison of the test statistic value to the critical value $\chi^2(3)$ suggests that the null hypothesis (H_0) should be rejected since the test statistic value is greater than the critical value. This means that the data provide sufficient evidence to conclude there is a positive linear association between depression and feeling bad about yourself due to more than half the days of feeling down, depressed, and hopeless. It's important to note that this test is appropriate only if a positive linear association is suspected by researchers before seeing data, further validating our findings.

The marginal contingency table, which presents the total counts across all levels of "Trouble Sleeping," is shown in Table 29.

Table 29: Marginal contingency table for "More Than Half the Days" of "Trouble Sleeping".

	Feeling Bad About Yourself			
Depression	Not at all	Several	More than half	Nearly
Not at all	3128	473	112	89
Several	454	284	63	50
More than half	121	62	41	31
Nearly	96	47	26	67

The odds ratio between the ”Several Days” and ”More Than Half The Days” levels of ”Feeling Bad About Yourself” and ”Depression” would be calculated.

The joint probability (θ_{XY}) between “Feeling Bad About Yourself” (X) and “Depression” (Y) can be calculated as follows:

$$\begin{aligned}\theta_{XY} &= \frac{n_{11+} \cdot n_{22+}}{n_{12+} \cdot n_{21+}}\\ &= \frac{284 \times 41}{62 \times 63}\\ &= \frac{11644}{3906}\\ &= 2.98\end{aligned}$$

It means that for those who have depression for several days, the odds of several days of feeling bad about yourself will be 2.98 times more than those who have depression for more than half the days of trouble sleeping. If we calculate the odds for the marginal table, we will find that all are greater than 1. Therefore, we would conclude that the odds of success are greater for those with trouble sleeping.

Finally, a test is performed to evaluate the conditional independence (conditional on Feeling level). An appropriate test would be the Cochran-Mantel-Haenszel(C M H) Test of conditional independence [28].

The associated hypotheses for the Cochran-Mantel-Haenszel Test are stated as follows:

Null Hypothesis (H_0): Feeling bad about yourself and depression are conditionally independent.

Alternative Hypothesis (H_a): Feeling bad about yourself and depression are not conditionally independent.

The Cochran-Mantel-Haenszel(C M H) χ^2 statistic is approximately 222.61, which is greater than the chi-square critical value of the degree of freedom 9 ($\chi^2(9) = 16.919$). The degree of freedom of the chi-square test of conditional independence is calculated as follows:

$$\begin{aligned} \text{df} &= k - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where *df* is the degree of freedom and k is the number of levels of "Trouble Sleeping".

We would reject the null hypothesis(H_0) since the test statistic value is greater than the critical value and conclude depression and feeling bad about yourself are not conditionally independent of trouble sleeping levels. For at least one trouble sleeping level, there is a significant association between feeling bad about yourself and depression.

We will investigate if there is any homogeneous association between all two-factor interaction terms, i.e., XY (Depression × Feeling).

The appropriate model would be the Log-linear Homogeneous Association Model (M_0). The general form of the Log-linear Homogeneous Association Model for this data is given by:

$$\ln(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad \forall i = 1,2,3,4, j = 1,2,3,4. \tag{96}$$

So, the fitted model can be represented as follows:

$$\ln(\mu_{ij}) = -1.621580 + 0.434453_i^{\text{Depression}} + 0.428898_j^{\text{Feeling}} - 0.049198_{ij}^{Depression\times Feeling} \tag{97}$$

To perform a test to assess the overall fit of this model, the deviance statistic, $G^2(M_0)$, is calculated by finding the difference between the null deviance, which is the saturated or full model, and the residual deviance, which is the reduced model.

The associated hypotheses for the deviant statistics are stated as follows:

Null Hypothesis (H_0): Extra model parameters are zero (not significant).

Alternative Hypothesis (H_a): Extra model parameters are non-zero
(at least one is significant).

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 5503.7 - 4291.8 \\ &= 1211.9 \end{aligned}$$

The degree of freedom of the chi-square test of the Log-linear Homogeneous Association Model is calculated as follows:

$$\begin{aligned} \text{df} &= (I-1)(J-1)(K-1) \\ &= (4-1)(4-1)(2-1) \\ &= 9, \end{aligned}$$

where *df* is the degree of freedom, I is the number of levels of "Feeling Bad About Yourself", J is the number of levels of "Depression" and K is the number of levels of parameters (Feeling Bad Yourself and Depression).

By comparing the test value to the critical value $\chi^2_{0.05}(9)$, we would reject the null hypothesis (H_0). Therefore, the data provide sufficient evidence to conclude the homogeneous association model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, sleeping, and feeling bad about ourselves.

The Multinomial Logistic Regression model can be written mathematically as:

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} \quad ; \quad j = 0, 1, 2$$

So,

$$\text{logit}\left(\frac{P \leq j}{P < j}\right) = \beta_0 + \beta_1 \text{Depression}_i + \beta_2 \text{Feeling}_i$$

More specifically, the model can be expanded as follows:

$$\begin{aligned}
\log\left(\frac{P_0}{P_1 + P_2 + P_3}\right) &= 1.9811 + 1.0219\text{Depression}_i + 0.6357\text{Feeling}_i, \\
\log\left(\frac{P_0 + P_1}{P_2 + P_3}\right) &= 3.8039 + 1.0219\text{Depression}_i + 0.6357\text{Feeling}_i, and \\
\log\left(\frac{P_0 + P_1 + P_2}{P_3}\right) &= 4.9173 + 1.0219\text{Depression}_i + 0.6357\text{Feeling}_i
\end{aligned}$$

To perform a test to assess the overall fit of the Cumulative Logit Model, the deviance statistic $G^2(M_0)$ is calculated and compared to the chi-square critical value. The associated hypotheses of the deviance statistic are stated as follows:

Null Hypothesis(H_0): Model fit is good.

Alternative Hypothesis (H_a): Model fit is not good.

The test statistic of the deviance statistic is calculated as follows:

$$\begin{aligned} G^2(M_0|M_1) &= G^2(M_0) - G^2(M_1) \\ &= 7758.285 - 6502.324 \\ &= 1255.961 \end{aligned}$$

The degree of freedom of the chi-square test of the Cumulative Logit Model is calculated as follows:

$$\begin{aligned} \text{df} &= K - 1 \\ &= 4 - 1 \\ &= 3, \end{aligned}$$

where *df* is the degree of freedom, K is the number of levels of "Trouble Sleeping".

By comparing the test value to the chi-square critical value ($\chi^2_{0.05}(3) = 7.815$), we would reject the null hypothesis. Therefore, the data provide sufficient evidence to conclude the cumulative logit model fits well. It means that there is an association between any pair of variables. Additionally, it implies that we have no three-way interaction between depression, trouble sleeping, and feeling bad about ourselves.

We will model the response variable, "Trouble Sleeping," with generalized linear models (GLMs). Since the data's response variable is a count variable, we will use the Poisson regression and negative binomial regression models.

The general form of the Poisson Regression Model is given by:

$$\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}, \tag{98}$$

where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Depression) and X_{2i} (Feeling Bad About Yourself) respectively.

So, the Poisson Regression Model will be represented as follows:

$$\ln(\hat{\mu}) = -1.56708 + 0.40148 \text{ Depression}_i + 0.32368 \text{ Feeling}_i \tag{99}$$

The coefficients will be interpreted in two ways, general and specific interpretations.

The general interpretation is as follows:

"There is an expected increase in the mean count of people having trouble sleeping".

The specific interpretations are given as follows:

"The number of times people have trouble sleeping with depression is expected to be $e^{0.40148} \approx 1.49$ times, feeling bad about yourself is constant, and "The number of times people have trouble sleeping with feeling bad about themselves is expected to be $e^{0.32368} \approx 1.38$ times, depression is constant".

The histogram of the data is right-skewed which is shown in the Figure 19 below:

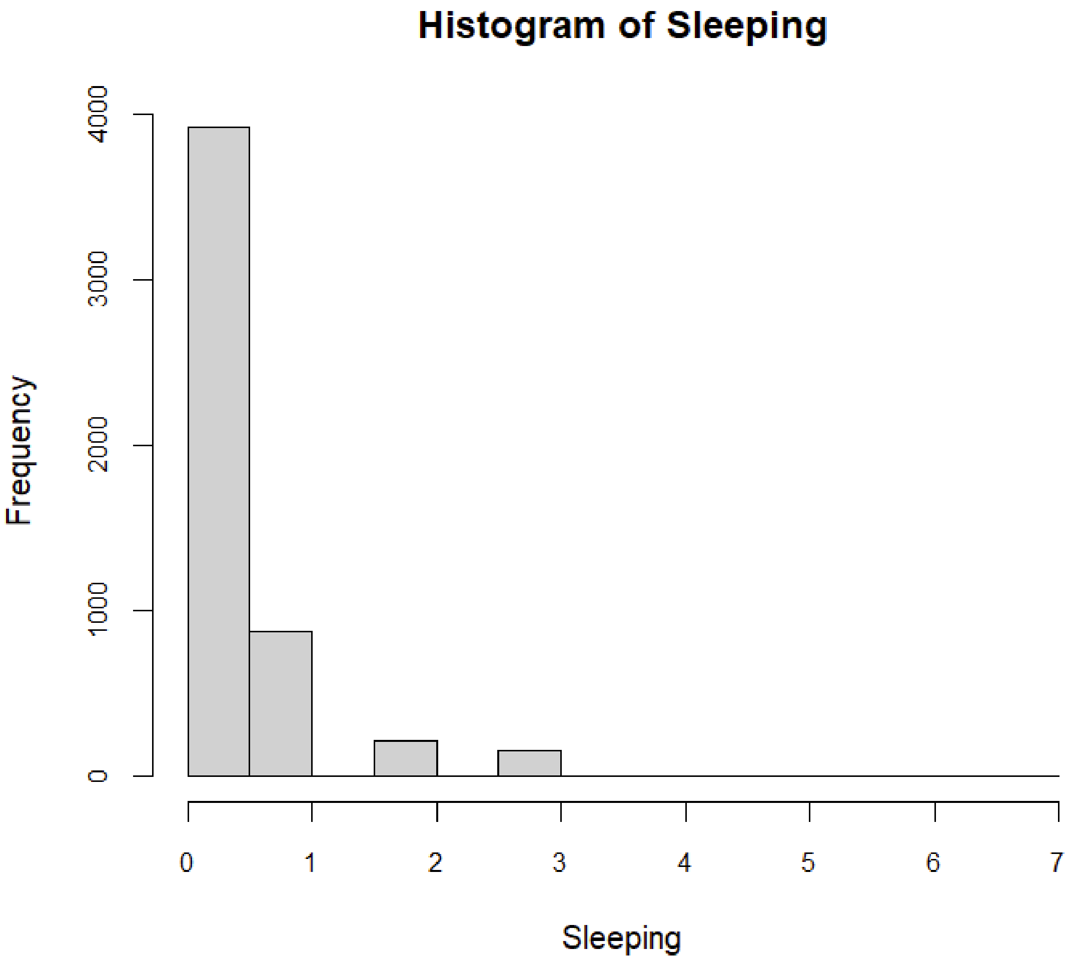


Figure 19: The histogram of trouble sleeping

The over-dispersion test checks if the variance is equal to the mean. The associated hypotheses to the over-dispersion test are stated as:

Null Hypothesis (H_0): variance = mean

Alternative Hypothesis (H_a): variance > mean

The dispersion estimate ($\hat{\phi}$) = 1.230287, greater than 1. The p-value for the over-dispersion test is 2.764×10^{-6}, which is less than $\alpha = 0.05$. The variance is equal to

0.5172042, greater than the mean, equal to 0.3467648. Therefore, the null hypothesis is rejected, and we conclude that the variance is greater than the mean.

The plot in Figure 20 depicts the scatter plot of the Poisson Model of “Trouble Sleeping”. Clusters of points vertically aligned at specific x-values suggest that the “Sleeping” variable might be categorical or has been binned into distinct levels. The spread of the residuals (the distance of the points from the horizontal line at zero on the y-axis) is indicative of the negative binomial model’s performance at different levels of the “Sleeping” variable. Residuals far from zero can indicate poor model fit or potential outliers. A good model fit would be indicated by a random scatter of points around the horizontal line at zero without any discernible pattern. Patterns, trends, or outliers in the plot could indicate issues with the model’s specifications or potential influential observations.

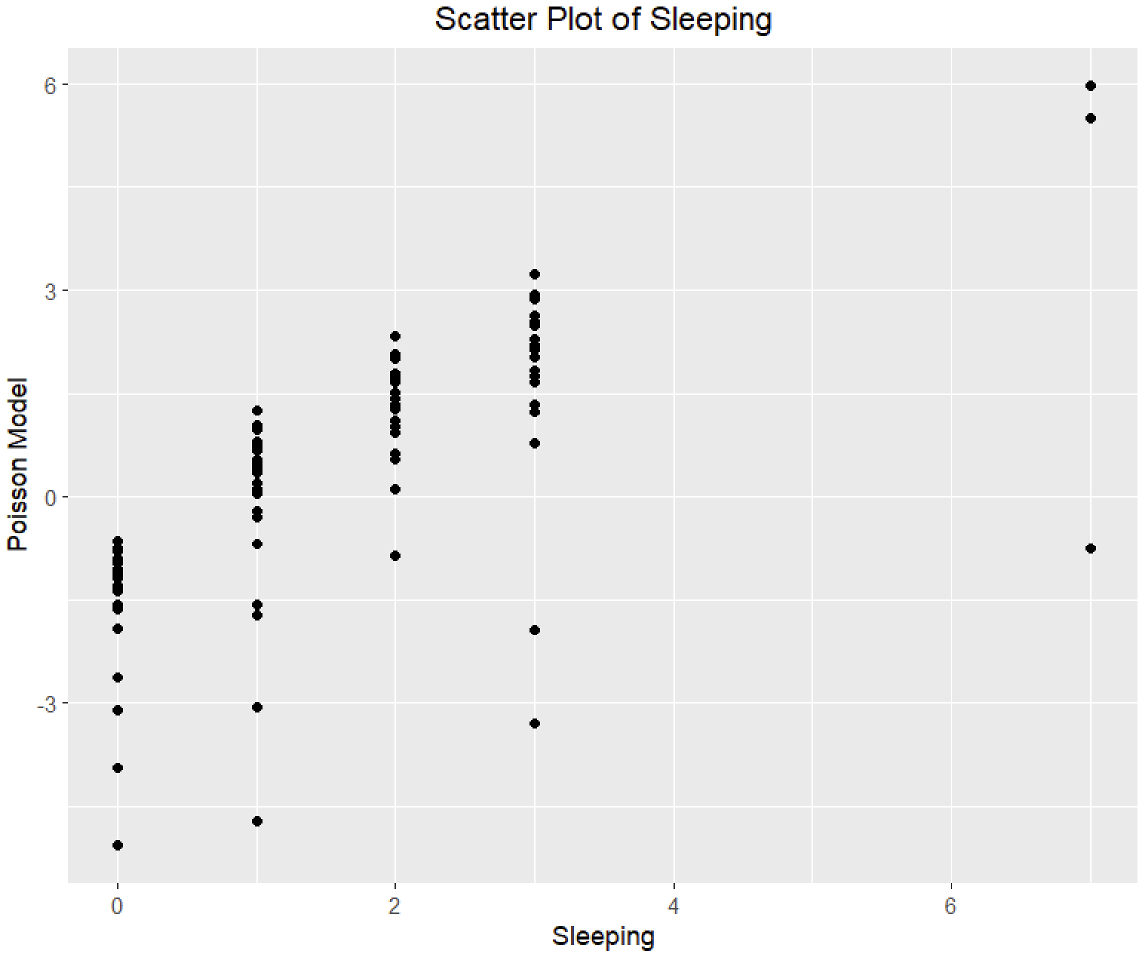


Figure 20: Scatter Plot of Poisson Regression Model for Trouble Sleeping

The over-dispersion test shows the model is a poor fit, therefore, we changed the model to Negative Binomial Regression.

The general form of the Negative Binomial Regression Model is given by: $\ln(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i}$, where μ_i is the mean count, β_0 is the intercept term of the model, β_1 and β_2 are the coefficients of the predictor variables X_{1i} (Depression) and X_{2i} (Feeling Bad About Yourself) respectively.

Therefore, the fitted model would be:

$$\ln(\hat{\mu}) = -1.74508 + 0.57225 \text{ Depression}_i + 0.27767 \text{ Feeling}_i \tag{100}$$

From the scatter plot of the Negative Binomial Regression model in Figure 21, there are distinct columns of data points, which likely correspond to the different categories of 'Sleeping.' Points that deviate significantly from zero may indicate potential outliers or leverage points that could influence the model fit. The absence of a clear pattern across different levels of 'Sleeping' suggests that the model's fit does not systematically vary with this predictor. However, the presence of points with high positive or negative residuals might be areas to investigate further for model improvement.

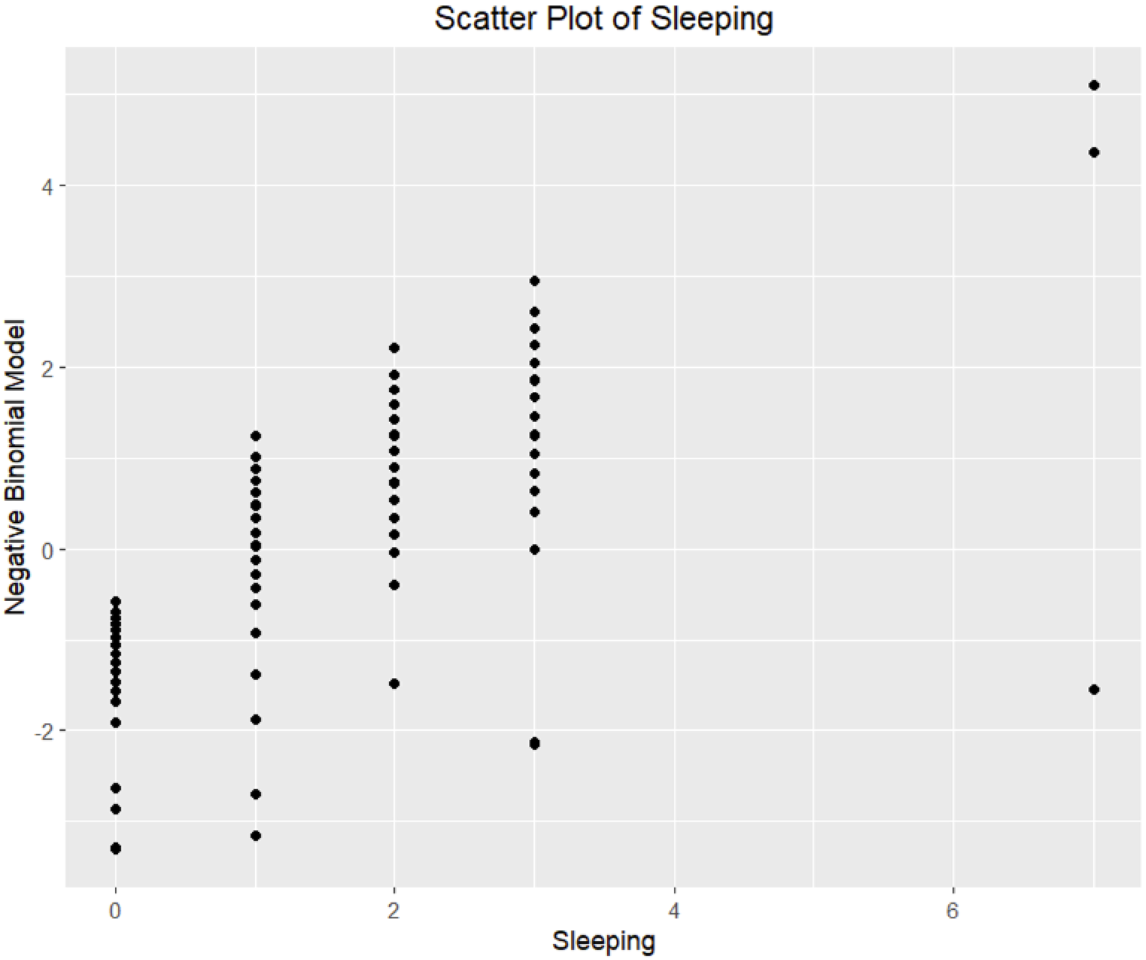


Figure 21: Scatter Plot of Negative Binomial Regression Model for Trouble Sleeping

The statistical measures obtained from the analysis of “Trouble Sleeping” are shown in Table 30. These measures are visualized graphically below, and colored vertical lines represent the best models.

Table 30: Trouble Sleeping

	Statistical Measures			
Models	AIC	BIC	G^2	MSE
Log-Linear	7053.431	7079.627	1238.3	0.5409994
Multinomial	6514.324	6553.619	1255.961	0.06607545
GLM(Poisson)	7077.584	7097.405	1211.9	1.109951
GLM (NB)	6907.558	6933.754	1120.3	0.9982505

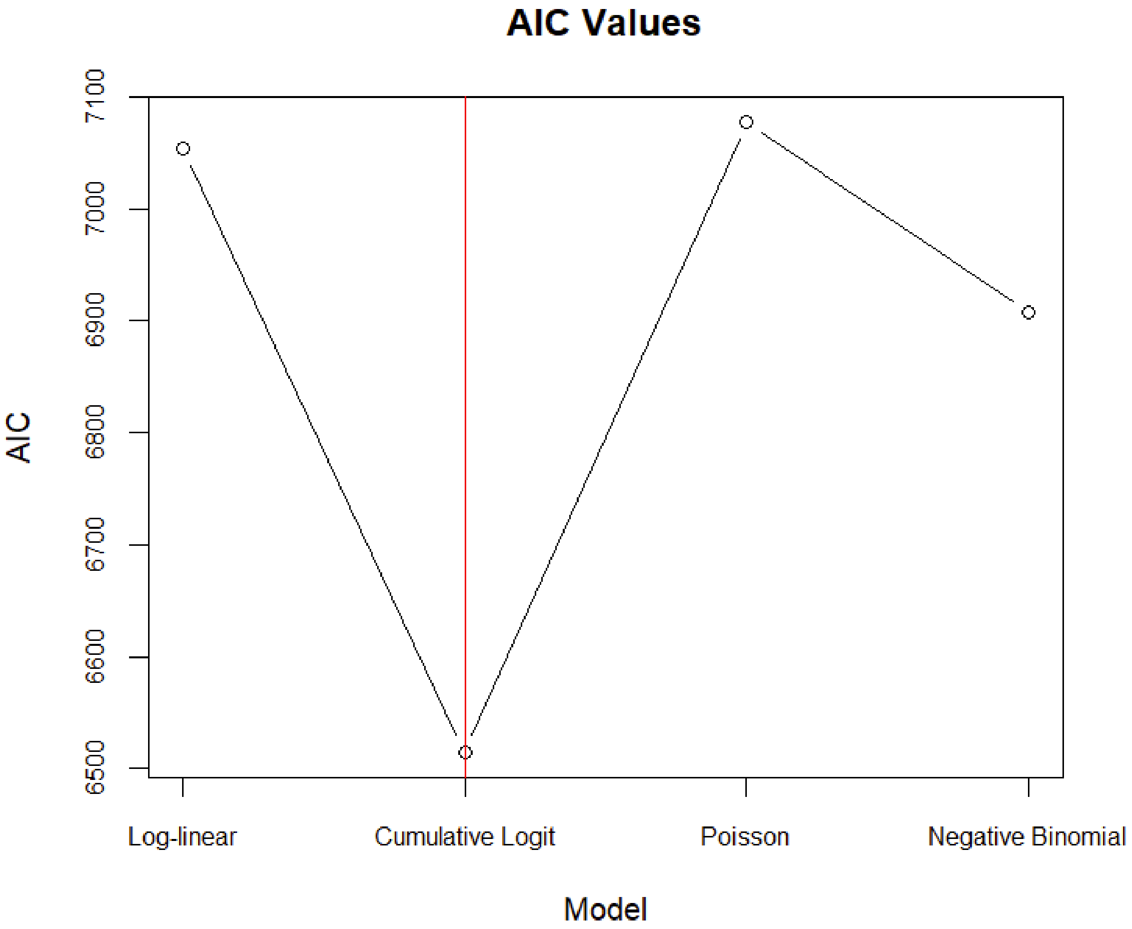


Figure 22: Graph of AIC values for “Trouble Sleeping”

Figure 22 above shows the graph of AIC values for “Trouble Sleeping,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

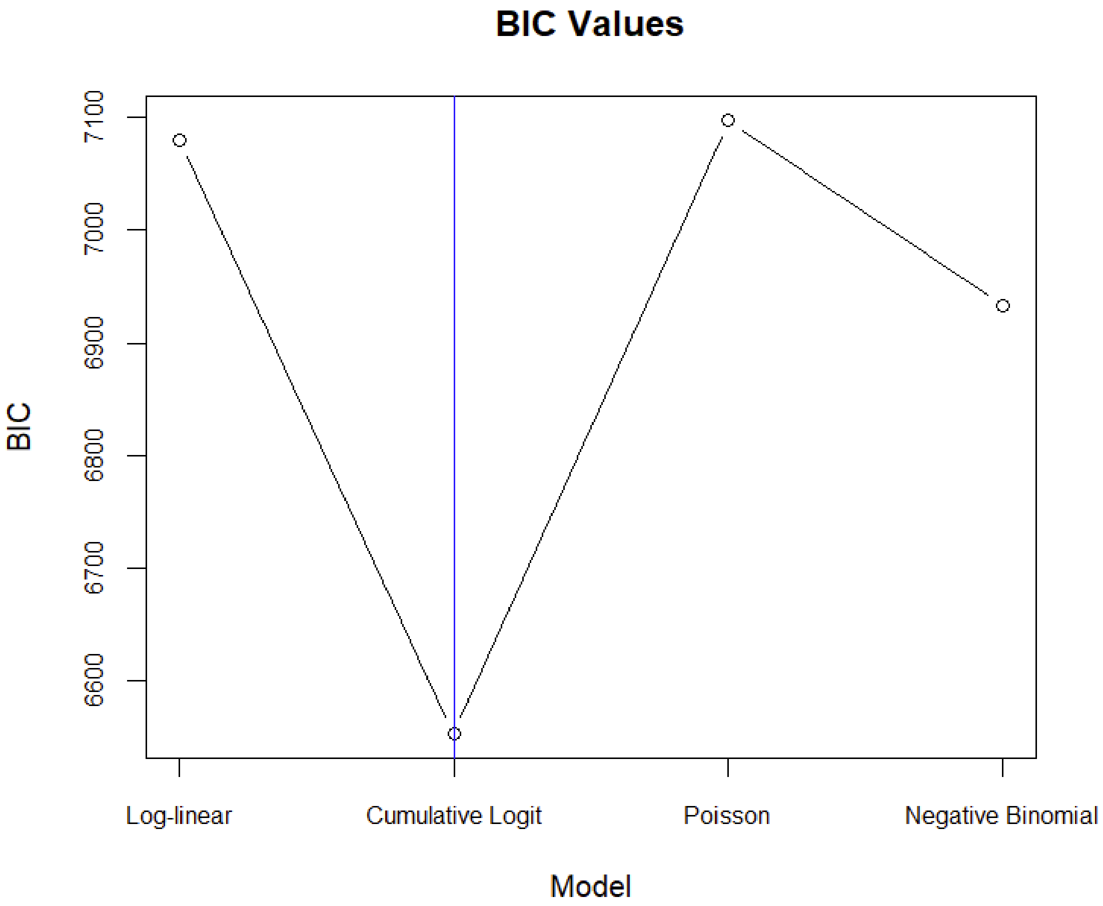


Figure 23: Graph of BIC values for “Trouble Sleeping”

Figure 23 above shows the graph of BIC values for “Trouble Sleeping,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

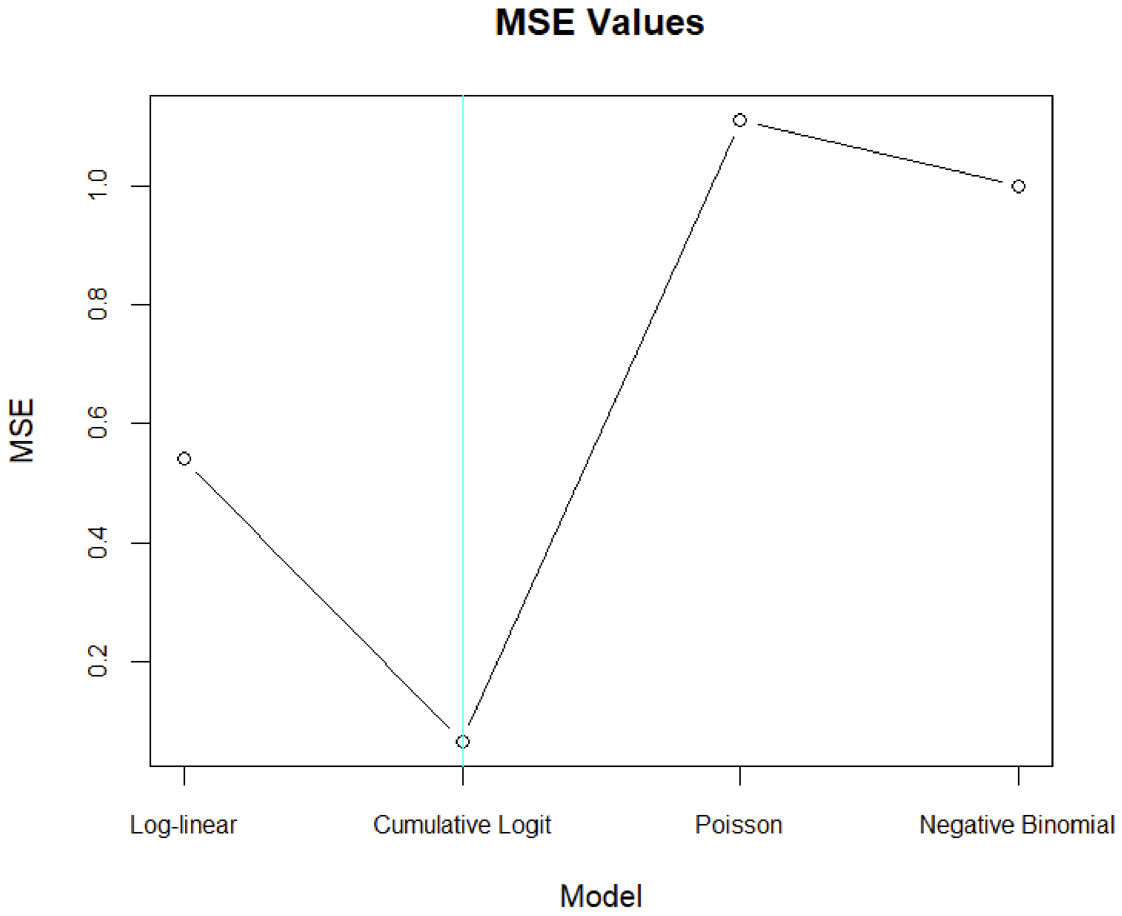


Figure 24: Graph of MSE values for “Trouble Sleeping”

Figure 24 above shows the graph of MSE values for “Trouble Sleeping,” the colored vertical line depicts that the Cumulative Logit Model is the best model for the analysis.

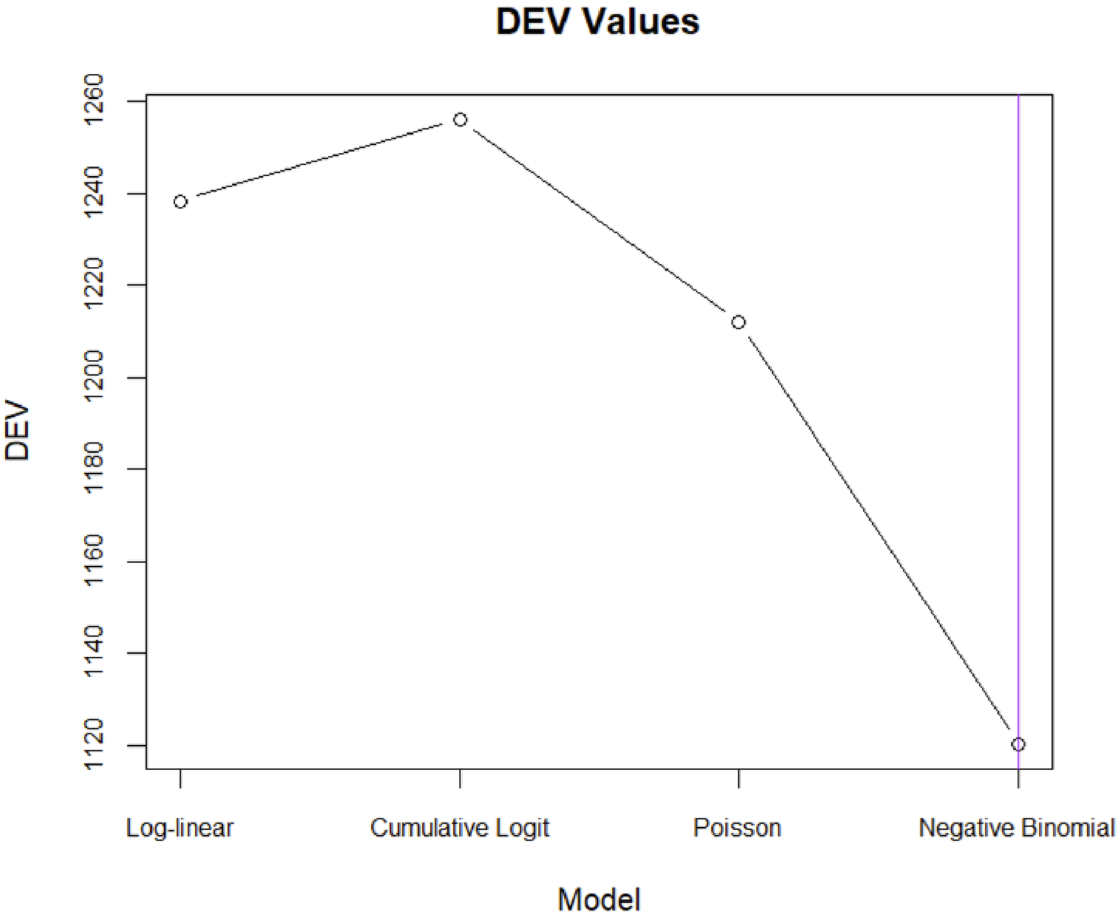


Figure 25: Graph of Deviance Statistics for “Trouble Sleeping”

Figure 25 above shows the graph of Deviance statistic values for “Trouble Sleeping,” and the colored vertical line depicts that the Negative Binomial Regression Model is the best model for the analysis.

The AIC, BIC, and MSE values graphs depict that the Cumulative Logit Model has the lowest value among other models, showing the best model to describe the relationship between the three variables. The Deviance Statistic value, on the other hand, depicts that the Negative Binomial Model has the lowest value among other models, showing the best model to describe the relationship between the variables.

We observed from the tables and graphs of values that the best model to analyze the relationship between Feeling bad about yourself, Feeling down, depressed, and hopeless, and Trouble Sleeping is the Multinomial Logistic Regression Model.

5 DISCUSSION AND FUTURE WORK

5.1 CONCLUSION

In this research, we wanted to investigate if there is any association between depression, trouble sleeping, and feeling bad about yourself. To do this, we employed three models. The first model was the log-linear model which showed We concluded that there are relationships between any pair of variables but no relationship between three variables simultaneously. In other words, people with depression commonly experience disturbed sleep patterns, people with trouble sleeping will feel bad about themselves, and those who have depression will feel bad about themselves. But there is no relationship between the three variables at the same time.

5.2 FUTURE WORK

In future studies exploring the dynamics of the relationship between depression, trouble sleeping, and feeling bad about yourself, non-parametric models such as those that do not assume any assumptions could be employed.

Additionally, supervised machine learning techniques, such as random forest, could assist researchers in seeing if there is any improvement in the relationship of the variables. Furthermore, since there was no association between depression, trouble sleeping, and feeling bad about yourself, adding more variables as predictors could potentially reveal more information regarding mental health.

BIBLIOGRAPHY

[1] Kaneita, Y., Ohida, T., Uchiyama, M., Takemura, S., Kawahara, K., Yokoyama, E., and Fujita, T. (2006). The relationship between depression and sleep disturbances: a Japanese nationwide general population survey. The Journal of Clinical Psychiatry.

[2] National Institute of Mental Health. (2024). Depression. https://www.nimh.nih.gov/health/topics/depression

[3] Freepik. (2020). 8 common causes of depression [Infographic]. https://image.freepik.com/free-vector/8-common-causes-depression-infographics_115990-222.jpg

[4] Dreamstime.(2020). Depression signs symptom infographic [Image of people with mental health problems: sad woman, despair, stress, loneliness].

[5] Bel Marra Health. (2019). Effects of Depression on the Human Body [Infographic].

[6] Passarella, S., Duong, M. T. (2008). Diagnosis and treatment of insomnia. American Journal of Health-System Pharmacy, 65(10), 927-934.

[7] Cleveland Clinic. (2017). Hypersomnia: Symptoms, Causes and Treatment. https://my.clevelandclinic.org/health/diseases/21591-hypersomnia

[8] Healthline. (2018). Insomnia vs. Hypersomnia: Understanding the Differences. https://www.healthline.com/health/insomnia/insomnia-vs-hypersomnia

[9] Healthline. (2022). Low Self-Esteem: Signs, Causes, and How to Improve. https://www.healthline.com/health/low-self-esteem

[10] Tsuno, N., Besset, A., and Ritchie, K. (2005). Sleep and depression. The Journal of Clinical Psychiatry.

[11] Alvaro, P. K., Roberts, R. M., and Harris, J. K. (2013). A systematic review assessing bidirectionality between sleep disturbances, anxiety, and depression. Sleep, 36(7), 1059-1068.

[12] Bhati, S., Richards, K. (2015). A systematic review of the relationship between postpartum sleep disturbance and postpartum depression. Journal of Obstetric, Gynecologic and Neonatal Nursing, 44(3), 350-357.

[13] Dinis, J., Bragança, M. (2018). Quality of sleep and depression in college students: a systematic review. Sleep Science, 11(04), 290-301.

[14] Oh, C. M., Kim, H. Y., Na, H. K., Cho, K. H., Chu, M. K. (2019). The effect of anxiety and depression on sleep quality of individuals with high risk for insomnia: a population-based study. Front Neurol, 10, 849.

[15] Wang, X., Cheng, S., Xu, H. (2019). Systematic review and meta-analysis of the relationship between sleep disorders and suicidal behavior in patients with depression. BMC Psychiatry, 19, 1-13.

[16] Morssinkhof, M. W. L., Van Wylick, D. W., Priester-Vink, S., van der Werf, Y. D., den Heijer, M., van den Heuvel, O. A., Broekman, B. F. P. (2020). Asso-

ciations between sex hormones, sleep problems, and depression: A systematic review. Neuroscience Biobehavioral Reviews, 118, 669-680.

[17] Yin, R., Li, L., Xu, L., Sui, W., Niu, M. E., Xu, R., Srirat, C. (2022). Association between depression and sleep quality in patients with systemic lupus erythematosus: a systematic review and meta-analysis. Sleep and Breathing, 26(1), 429-441.

[18] Jiang, Y., Jiang, T., Xu, L. T., Ding, L. (2022). Relationship of depression and sleep quality, diseases and general characteristics. World journal of psychiatry, 12(5), 722.

[19] Kroenke, K., Spitzer, R. L. (2002). The PHQ-9: a new depression diagnostic and severity measure. Psychiatric Annals, 32(9), 509-515.

[20] Walck, C. (1996). Handbook on statistical distributions for experimentalists (pp. 54-56). Stockholms universitet.

[21] Krishnamoorthy, K. (2006). Handbook of statistical distributions with applications. Chapman and Hall/CRC.

[22] NHANES 2015-2016: Mental Health - Depression Screener Data Documentation, Codebook, and Frequencies. `https://www.cdc.gov/nchs/nhanes/nhanes2015-2016/mh_depression.htm`

[23] Centers for Disease Control and Prevention. (2015). NHANES Questionnaire.

[24] Lauritzen, S. L. (1979). Lectures on contingency tables. University of Copenhagen.

[25] Fabozzi, F. J., Focardi, S. M., Rachev, S. T., Arshanapalli, B. G., Hoechstoetter, M. (2014). Appendix E: model selection criterion: AIC and BIC. The basics of financial econometrics, 41(1979), 399-403.

[26] Sullivan, J. (2024). Mean Squared Error (MSE). Statistics. `https://statisticsbyjim.com/regression/mean-squared-error`

[27] Hornik, K., Zeileis, A., and Meyer, D. (2006). The strucplot framework: visualizing multi-way contingency tables with vcd. Journal of Statistical Software, 17(3), 1-48.

[28] Agresti, A. (1996). An Introduction to Categorical Data Analysis (Vol. 135). New York: Wiley.

APPENDIX: R Code

```
install.packages("haven")

library(Hmisc)

library(haven)

install.packages("remotes")

library(remotes)

# Reading of Data

file.choose()

mental<-read_xpt("C:\\Users\\musli\\OneDrive\\Desktop\\DPQ_I.XPT")

colnames(mental)=c("LitteleInterest","Depression","Sleeping",

                    "LittleEnergy","Appetite","Feeling")

View(mental)

head(mental)

dim(mental)

summary(mental)

nrow(mental)

#Analysis of Feeling

# Contingency Tables

t= table(mental$Depression,mental$Sleeping,mental$Feeling);t

ftable(t,row.vars =3)

ftable(mental$Depression,mental$Sleeping,mental$Feeling,exclude=c(NA),

       row.vars = NULL,col.vars = NULL)

# Feeling Contingency Table
```

```
tab1 <- xtabs(~Depression+Sleeping+Feeling,data=mental)

tab1

summary(xtabs(~Depression+Sleeping+Feeling,data=mental))

tab_1 = tab1[-c(5:6), -c(5:6), -5]

tab_1

# Marginal Contingency Table

addmargins(tab_1)

require(vcd)

# Feeling Mosaic Plot

MP_1=mosaicplot(~Depression+Sleeping+Feeling,data=mental,

                color = TRUE,las=1)

mosaic(tab_1,split_horizontal = c(TRUE, TRUE, FALSE))

mosaic(tab_1,gp = shading_hcl,

       gp_args = list(h = c(130, 43), c = 100, l = c(90, 70)))

mosaic(tab_1,shade=T)

assoc(tab_1)

fill_colors <- matrix(c("dark cyan","gray","gray","dark magenta"),

                            ncol = 2)

mosaic(tab_1, gp = gpar(fill = fill_colors, col = 0))


# Feeling Linear Trend

library(wCorr)
```

```
Mental= data.frame(cbind(Depression=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                    TroubleSleeping=c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4),
                    count=c(83,19,9,1,16,36,10,1,8,11,14,8,6,8,4,8),
                    u=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                     v=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4)))
# CALCULATE CORRELATION DIRECTLY #
weightedCorr(Mental$u, Mental$v, weights=Mental$count, method='Pearson')
# Feeling CMH test
install.packages("xtable")
library(xtable)
design.table <- xtable(mental,auto=TRUE)
print(design.table)
my_ftable1 <- (tab_1)
my_df1 <- as.data.frame(my_ftable1)
my_xtable1 <- xtable(my_df1)
print(my_xtable1)
mantelhaen.test(tab_1)
# Cochran-Mantel-Haenszel test
#data:  tab_1 (Feeling)
#Cochran-Mantel-Haenszel M^2 = 1178, df = 9, p-value < 2.2e-16
# LOGLINEAR MODEL Feeling
attach(mental)
```

```
loglin_model_1 <- glm(Feeling ~ (Depression + Sleeping)^2,
                      data = mental, family = poisson())
# Check the summary of the model
summary(loglin_model_1)
# AIC/ BIC
AIC(loglin_model_1)
# AIC = 8347.642
BIC(loglin_model_1)
# BIC = 8373.839
# Chi Square Test
null_model_1 <- glm(Feeling ~ 1, family = poisson(), data = mental)
chi_square_LM1 <- anova(null_model_1, loglin_model_1, test = "Chisq")
print(chi_square_LM1)
# Pr(Chi)= < 2.2e-16
# MSE
# Fit the log-linear model
loglin_model_1 <- glm(Feeling ~ (Depression + Sleeping)^2,
                      data = mental, family = poisson())
y_hat_1 <- predict(loglin_model_1, type = "response")
y_1 <- mental$Feeling
# Remove missing values
y_1 <- y_1[!is.na(y_1)]
y_hat_1 <- y_hat_1[!is.na(y_hat_1)]
```

```
sse_1 <- sum((y_1 - y_hat_1)^2)

print(sse_1)

# SSE = 3178.006

n_1 <- length(y_1)

mse_1 <- sse_1 / n_1

print(mse_1)

# MSE = 0.615654

# Deviance Stat.

Dev_LM1 <- 6177.6 - 5242.2

Dev_LM1

# Dev = 935.4

#Multi- nominal Logistic regression (Cumulative Logit Model)

library(MASS)

mental$Feeling <- factor(mental$Feeling)

cumulative_model_1 <- polr(Feeling ~ Depression + Sleeping,

                           data = mental, Hess = TRUE)

summary(cumulative_model_1)

# AIC/BIC

AIC(cumulative_model_1 )

# AIC = 7711.33

BIC(cumulative_model_1)

# BIC = 7750.624

# Null Deviance
```

```
null_model_1 <- polr(Feeling ~ 1, data = mental)

null_deviance_1 <- deviance(null_model_1)

print(null_deviance_1)

# Null Deviance =  8414.769

# MSE

cumulative_model_1 <- polr(Feeling ~ Depression + Sleeping,

                           data = mental, Hess = TRUE)

predicted_probs_1 <- predict(cumulative_model_1, type = "probs")

observed_responses_1 <- model.matrix(~ Feeling - 1, data = mental)

squared_residuals_1 <- (observed_responses_1 - predicted_probs_1)^2

mse_CM1 <- mean(squared_residuals_1)

print(mse_CM1)

# MSE =  0.07756587

# Dev. Stat.

Dev_CM1 <- 8414.769 - 7699.33

Dev_CM1

# Dev = 715.439

# GLM Poisson

library(MASS)

mental <- mental[!is.na(mental$Feeling), ]

mental$Feeling <- as.numeric(as.character(mental$Feeling))

if (any(mental$Feeling < 0)) {

  mental <- mental[mental$Feeling >= 0, ]
}

Pmodel_1 <- glm(Feeling ~ Depression + Sleeping,

                data = mental, family = poisson())

summary(Pmodel_1)

#AIC/BIC

AIC(Pmodel_1)

# AIC = 8400.514

BIC(Pmodel_1)

# BIC = 8420.161

#MSE

Pmodel_1 <- glm(Feeling ~ Depression + Sleeping,

                data = mental, family = poisson())

residuals_pb1 <- residuals(Pmodel_1, type = "pearson")

mse_pb1 <- mean(residuals_pb1^2)

print(mse_pb1)

# MSE = 1.384698

# Deviance Stat.

Dev_Pb1 <- 6177.6 - 5297.1

Dev_Pb1

# Dev = 880.5

#Over-Dispersion Test

attach(mental)

library(AER)
library(ggplot2)
dispersiontest(Pmodel_1)
mean(mental$Feeling)
var(mental$Feeling)
hist(mental$Feeling, main = "Histogram of Feeling Bad About Yourself",
     xlab = "Feeling Bad About Yourself")
library(ggplot2)
qplot(mental$Feeling, summary(Pmodel_1)$deviance.resid,
      xlab = "Feeling Bad About Yourself",
      ylab = "Poisson Model") +
  ggtitle("Scatter Plot of Feeling Bad About Yourself") +
  theme(plot.title = element_text(hjust = 0.5))
# Dispersion = 1.437269
# Mean = 0.4074002
# Variance = 0.6681336
# GLM (Negative Binomial)
nbmodel_1 <- glm.nb(Feeling ~ Depression + Sleeping, data = mental)
summary(nbmodel_1)
qplot(mental$Feeling, summary(nbmodel_1)$deviance.resid,
      xlab = "Feeling Bad About Yourself",
      ylab = "Negative Binomial Model") +
  ggtitle("Scatter Plot of Feeling Bad About Yourself") +
  theme(plot.title = element_text(hjust = 0.5))
```

```
# AIC/BIC

AIC(nbmodel_1)

# AIC = 8113.765

BIC(nbmodel_1)

# BIC = 8139.961

# Mean Square Error

residuals_nb1 <- residuals(nbmodel_1, type = "pearson")

mse_nb1 <- mean(residuals_nb1^2)

print(mse_nb1)

# MSE = 1.069236

# Deviance Stat.

Dev_nb1 <- 4428.2 - 3772.9

Dev_nb1

# Dev = 655.3

# Graph of values for Feeling

AIC <- c(AIC(loglin_model_1), AIC(cumulative_model_1),

         AIC(Pmodel_1), AIC(nbmodel_1))

BIC <- c(BIC(loglin_model_1), BIC(cumulative_model_1),

         BIC(Pmodel_1), BIC(nbmodel_1))

MSE <- c(mse_1, mse_CM1, mse_pb1, mse_nb1)

DEV <- c(Dev_LM1,Dev_CM1, Dev_Pb1, Dev_nb1)

best_model_index <- which.min(AIC)

best_model_index_1 <- which.min(BIC)
```

```
best_model_index_2 <- which.min(MSE)

best_model_index_3 <- which.min(DEV)

models <- c("Log-linear", "Cumulative Logit",

            "Poisson","Negative Binomial")

if(length(AIC) != length(BIC) || length(BIC) != length(MSE) ||

   length(MSE) != length(DEV)|| length(DEV) != length(models)){

  stop("Lengths of vectors don’t match.")}

AIC <- na.omit(AIC)

BIC <- na.omit(BIC)

MSE <- na.omit(MSE)

DEV <- na.omit(DEV)

# Set the plot size to accommodate longer labels

par(mar = c(5, 6, 4, 2) + 0.1, cex.axis = 0.8)

# Plot each vector separately

plot(AIC, type="b", xlab="Model", ylab="AIC",

main="AIC Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index, col="brown")

plot(BIC, type="b", xlab="Model", ylab="BIC",

main="BIC Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_1, col="navyblue")

plot(MSE, type="b", xlab="Model", ylab="MSE",
main="MSE Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_2, col="pink")

plot(DEV, type="b", xlab="Model", ylab="DEV",

main="DEV Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_3, col="skyblue")


# Analysis of Depression

# Depression Contingency Table

tab2 <- xtabs(~Sleeping+Feeling+Depression,data=mental)

tab2

summary(xtabs(~Sleeping+Feeling+Depression,data=mental))

tab_2 = tab2[-c(5:6), -c(5:6), -5]

tab_2

#Depression Marginal Contingency table

addmargins(tab_2)

# Depression Mosaic plot

require(vcd)

# Create mosaic plot with title

MP_2 <- mosaicplot(~Sleeping + Feeling + Depression,

                   data = mental, color = TRUE, las = 1,

                   main = "Mosaic Plot of Sleeping
```

```
                and Feeling vs. Depression")
mosaic(tab_2,gp = shading_hcl, gp_args = list(h = c(130, 43),
                                    c = 100, l = c(90, 70)))
mosaic(tab_2,split_horizontal = c(TRUE, TRUE, FALSE))
mosaic(tab_2,shade=T)
assoc(tab_2)
fill_colors <- matrix(c("yellow","green","green","purple")
                                        , ncol = 2)
mosaic(tab_2, gp = gpar(fill = fill_colors, col = 0))
# Depression CMH Test
my_ftable2 <- (tab_2)
my_df2 <- as.data.frame(my_ftable2)
my_xtable2 <- xtable(my_df2)
print(my_xtable2)
mantelhaen.test(tab_2)
# Cochran-Mantel-Haenszel test
# data: Depression
#Cochran-Mantel-Haenszel M^2 = 280.35, df = 9,
# p-value < 2.2e-16
# Depression Linear Trend
library(wCorr)
Mental= data.frame(cbind(Depression=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                TroubleSleeping=c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4),
                count=c(65,12,8,10,26,28,11,5,22,14,14,9,8,8,8,7),
                u=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                v=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4)))
# CALCULATE CORRELATION DIRECTLY #
weightedCorr(Mental$u, Mental$v, weights=Mental$count,
                method=’Pearson’)
# LOGLINEAR MODEL Depression
loglin_model_2 <- glm(Depression ~ (Sleeping + Feeling)^2,
                      data = mental, family = poisson())
summary(loglin_model_2)
# AIC/ BIC
AIC(loglin_model_2)
# AIC = 8152.288
BIC(loglin_model_2)
# BIC = 8178.484
# Chi Square Test
null_model_2 <- glm(Depression ~ 1, family = poisson(), data = mental)
chi_square_test_2 <- anova(null_model_2, loglin_model_2, test = "Chisq")
print(chi_square_test_2)
# Chi-Square = < 2.2e-16
# MSE
# Fit the log-linear model
loglin_model_2 <- glm(Depression ~ (Sleeping + Feeling)^2,
                      data = mental, family = poisson())
y_hat_2 <- predict(loglin_model_2, type = "response")
y_2 <- mental$Depression
# Remove missing values
y_2 <- y_2[!is.na(y_2)]
y_hat_2 <- y_hat_2[!is.na(y_hat_2)]
sse_2 <- sum((y_2 - y_hat_2)ˆ2)
print(sse_2)
n_2 <- length(y_2)
mse_2 <- sse_2 / n_2
print(mse_2)
# MSE =  0.7365462
# Deviance Stat.
Dev_LM2 <- 6442.9 - 5011.6
Dev_LM2
# Dev = 1431.3
#Multi- nominal Logistic regression (Cumulative Logit Model)
library(MASS)
mental$Depression <- factor(mental$Depression)
cumulative_model_2 <- polr(Depression ˜ Sleeping + Feeling,
                           data = mental, Hess = TRUE)
summary(cumulative_model_2)
# AIC/BIC
```

```
AIC(cumulative_model_2 )
# AIC = 7228.218
BIC(cumulative_model_2)
# BIC = 7274.062
# Null Deviance
null_model_2 <- polr(Depression ~ 1, data = mental)
null_deviance_2 <- deviance(null_model_2)
print(null_deviance_2)
# Null Deviance = 8557.117
# 1342.899
# Deviance Stat.
Dev_CM2 <- 8557.117 - 7214.218
Dev_CM2
# Dev = 1342.899
# MSE
mental$Depression <- factor(mental$Depression)
cumulative_model_2 <- polr(Depression ~ Sleeping + Feeling,
                           data = mental, Hess = TRUE)
predicted_probs_2 <- predict(cumulative_model_2, type = "probs")
observed_responses_2 <- model.matrix(~ Depression - 1, data = mental)
squared_residuals_2 <- (observed_responses_2 - predicted_probs_2)^2
mse_CM2 <- mean(squared_residuals_2)
print(mse_CM2)
```

```
# MSE =  0.05878926

# GLM Poisson

mental <- mental[!is.na(mental$Depression), ]

mental$Depression <- as.numeric(as.character(mental$Depression))

if (any(mental$Depression< 0)){

  mental <- mental[mental$Depression>= 0, ]}

Pmodel_2 <- glm(Depression ~ Sleeping + Feeling,

                data = mental, family = poisson())

summary(Pmodel_2)

#AIC/BIC

AIC(Pmodel_2)

# AIC = 8201.003

BIC(Pmodel_2)

# BIC = 8220.65

#MSE

Pmodel_2 <- glm(Depression ~ Sleeping + Feeling,

                data = mental, family = poisson())

residuals_pb2 <- residuals(Pmodel_2, type = "pearson")

mse_pb2 <- mean(residuals_pb2^2)

print(mse_pb2)

# MSE = 1.542902

# Deviance Stat.

Dev_Pb2 <- 6442.9 - 5062.3
```

```
Dev_Pb2

# Dev = 1380.6

#Over-Dispersion Test

library(AER)

library(ggplot2)

dispersiontest(Pmodel_2)

mean(mental$Depression)

var(mental$Depression)

hist(mental$Depression, main = "Histogram of Depression",

        xlab = "Depression")

library(ggplot2)

qplot(mental$Depression, summary(Pmodel_2)$deviance.resid,

      xlab = "Depression",

      ylab = "Poisson Model") +

  ggtitle("Scatter Plot of Depression") +

  theme(plot.title = element_text(hjust = 0.5))

# Dispersion = 1.624339

# Mean = 0.4213483

# Variance = 0.7530648

# GLM (Negative Binomial)

nbmodel_2 <- glm.nb(Depression ~ Sleeping + Feeling,

                data = mental)

summary(nbmodel_2)
```

```
qplot(mental$Depression, summary(nbmodel_2)$deviance.resid,
      xlab = "Depression",
      ylab = "Negative Binomial Model") +
  ggtitle("Scatter Plot of Depression") +
  theme(plot.title = element_text(hjust = 0.5))
# 1064.6
# AIC/BIC
AIC(nbmodel_2)
# AIC = 7947.697
BIC(nbmodel_2)
# BIC = 7973.894
# Mean Square Error
residuals_nb2 <- residuals(nbmodel_2, type = "pearson")
mse_nb2 <- mean(residuals_nb2^2)
print(mse_nb2)
# MSE = 1.324138
# Deviance Stat.
Dev_nb2 <- 4819.6 - 3755.0
Dev_nb2
# Dev = 1064.6
# Graph of values for Depression
AIC <- c(AIC(loglin_model_2), AIC(cumulative_model_2),
         AIC(Pmodel_2), AIC(nbmodel_2))
```

```
BIC <- c(BIC(loglin_model_2), BIC(cumulative_model_2),
         BIC(Pmodel_2), BIC(nbmodel_2))
MSE <- c(mse_2, mse_CM2, mse_pb2, mse_nb2)
DEV <- c(Dev_LM2,Dev_CM2,Dev_Pb2,Dev_nb2)
best_model_index <- which.min(AIC)
best_model_index_1 <- which.min(BIC)
best_model_index_2 <- which.min(MSE)
best_model_index_3 <- which.min(DEV)
models <- c("Log-linear", "Cumulative Logit",
                "Poisson","Negative Binomial")
if(length(AIC) != length(BIC) || length(BIC) != length(MSE) ||
   length(MSE) != length(DEV)|| length(DEV) != length(models)){
  stop("Lengths of vectors don’t match.")}
AIC <- na.omit(AIC)
BIC <- na.omit(BIC)
MSE <- na.omit(MSE)
DEV <- na.omit(DEV)
# Set the plot size to accommodate longer labels
par(mar = c(5, 6, 4, 2) + 0.1, cex.axis = 0.8)
# Plot each vector separately
plot(AIC, type="b", xlab="Model", ylab="AIC",
        main="AIC Values", xaxt="n")
axis(1, at=1:length(models), labels=models)
```

```
abline(v=best_model_index, col="green")
plot(BIC, type="b", xlab="Model", ylab="BIC",
        main="BIC Values", xaxt="n")
axis(1, at=1:length(models), labels=models)
abline(v=best_model_index_1, col="magenta")
plot(MSE, type="b", xlab="Model", ylab="MSE",
        main="MSE Values", xaxt="n")
axis(1, at=1:length(models), labels=models)
abline(v=best_model_index_2, col="orange")
plot(DEV, type="b", xlab="Model", ylab="DEV",
        main="DEV Values", xaxt="n")
axis(1, at=1:length(models), labels=models)
abline(v=best_model_index_3, col="maroon")

# Analysis for Sleeping
# Sleeping Contingency Table
tab3 <- xtabs(~Depression+Feeling+Sleeping,data=mental)
tab3
summary(xtabs(~Depression+Feeling+Sleeping,data=mental))
tab_3 = tab3[-c(5:6), -c(5:6), -5]
tab_3
# Sleeping Marginal Contingency Table
addmargins(tab_3)
```

```
#  Sleeping Mosaic Plot

MP_3 =mosaicplot(~Depression+Feeling+Sleeping,

            data=mental,color = TRUE,las=1)

mosaic(tab_3,split_vertical = c(TRUE, TRUE, FALSE))

mosaic(tab_3,gp = shading_hcl,

       gp_args = list(h = c(130, 43), c = 100,

            l = c(90, 70)))

mosaic(tab_3,shade=T)

assoc(tab_3)

fill_colors <- matrix(c("red","blue","blue","orange"),

                ncol = 2)

mosaic(tab_3, gp = gpar(fill = fill_colors, col = 0))

# Sleeping CMH Test

my_ftable3 <- (tab_3)

my_df3 <- as.data.frame(my_ftable3)

my_xtable3 <- xtable(my_df3)

print(my_xtable3)

mantelhaen.test(tab_3)

# Cochran-Mantel-Haenszel test

# data: Sleeping

# Cochran-Mantel-Haenszel M^2 = 222.61, df = 9, p-value < 2.2e-16

# Sleeping Linear Trend

library(wCorr)
```

```
Mental= data.frame(cbind(Depression=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                TroubleSleeping=c(1,2,3,4,1,2,3,4,1,2,3,4,1,2,3,4),
                count=c(42,6,9,6,14,25,10,9,22,14,14,9,8,8,4,9),
                u=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4),
                v=c(1,1,1,1,2,2,2,2,3,3,3,3,4,4,4,4)))
# CALCULATE CORRELATION DIRECTLY #
weightedCorr(Mental$u, Mental$v, weights=Mental$count,
                method='Pearson')
# LOGLINEAR MODEL
loglin_model_3 <- glm(Sleeping ~ (Depression + Feeling)^2,
                      data = mental, family = poisson())
summary(loglin_model_3)
# AIC/ BIC
AIC(loglin_model_3)
# AIC =  7053.431
BIC(loglin_model_3)
# BIC = 7079.627
# MSE
# Fit the log-linear model
loglin_model_3 <- glm(Sleeping ~ (Depression + Feeling)^2,
                      data = mental, family = poisson())
y_hat_3 <- predict(loglin_model_3, type = "response")
y_3 <- mental$Sleeping
```

```
# Remove missing values

y_3 <- y_3[!is.na(y_3)]

y_hat_3 <- y_hat_3[!is.na(y_hat_3)]

sse_3 <- sum((y_3 - y_hat_3)^2)

print(sse_3)

n_3 <- length(y_3)

mse_3 <- sse_3 / n_3

print(mse_3)

# MSE = 0.5409994

Dev_LM3 <- 5503.7 -  4265.4

Dev_LM3

# Dev Stat. = 1238.3

#Multi- nominal Logistic regression (Cumulative Logit Model)

library(MASS)

mental$Sleeping <- factor(mental$Sleeping)

cumulative_model_3 <- polr(Sleeping ~ Depression + Feeling,

                           data = mental, Hess = TRUE)

summary(cumulative_model_3)

# AIC/BIC

AIC(cumulative_model_3)

# AIC = 6514.324

BIC(cumulative_model_3)

# BIC = 6553.619
# Null Deviance
null_model_3 <- polr(Sleeping ~ 1, data = mental)
null_deviance_3 <- deviance(null_model_3)
print(null_deviance_3)
# Null Deviance = 7758.285
#Deviance Stat.
Dev_CM3 <- 7758.285 - 6502.324
Dev_CM3
# Dev = 1255.961
# MSE
mental$Sleeping <- factor(mental$Sleeping)
cumulative_model_3 <- polr(Sleeping ~ Depression + Feeling,
                           data = mental, Hess = TRUE)
predicted_probs_3 <- predict(cumulative_model_3, type = "probs")
observed_responses_3 <- model.matrix(~ Sleeping - 1, data = mental)
squared_residuals_3 <- (observed_responses_3 - predicted_probs_3)^2
mse_CM3 <- mean(squared_residuals_3)
print(mse_CM3)
# MSE =  0.06607545
# GLM Poisson
mental <- mental[!is.na(mental$Sleeping), ]
mental$Sleeping <- as.numeric(as.character(mental$Sleeping))
if (any(mental$Sleeping< 0)){
  mental <- mental[mental$Sleeping>= 0, ]}

Pmodel_3 <- glm(Sleeping ~ Depression + Feeling,

                data = mental, family = poisson())

summary(Pmodel_3)

# AIC

AIC(Pmodel_3)

# AIC = 7077.758

#BIC

BIC(Pmodel_3)

# BIC = 7097.405

#MSE

Pmodel_3 <- glm(Sleeping ~ Depression + Feeling,

                data = mental, family = poisson())

residuals_pb3 <- residuals(Pmodel_3, type = "pearson")

mse_pb3 <- mean(residuals_pb3^2)

print(mse_pb3)

# MSE = 1.109951

#Dev. Stat

Dev_Pb3 <- 5503.7 - 4291.8

Dev_Pb3

# Dev. Stat = 1211.9

#Over-Dispersion Test

library(AER)
library(ggplot2)

dispersiontest(Pmodel_3)

mean(mental$Sleeping)

var(mental$Sleeping)

hist(mental$Sleeping, main = "Histogram of Sleeping",

            xlab = "Sleeping")

library(ggplot2)

qplot(mental$Sleeping, summary(Pmodel_3)$deviance.resid,

      xlab = "Sleeping",

      ylab = "Poisson Model") +

  ggtitle("Scatter Plot of Sleeping") +

  theme(plot.title = element_text(hjust = 0.5))

#Dispersion = 1.230287

#Mean = 0.3467648

#Variance= 0.5172042

# GLM (Negative Binomial)

nbmodel_3 <- glm.nb(Sleeping ~ Depression + Feeling,

                    data = mental, maxit = 1000)

summary(nbmodel_3)

# 1120.3

library(ggplot2)

qplot(mental$Sleeping, summary(nbmodel_3)$deviance.resid,

      xlab = "Sleeping",
      ylab = "Negative Binomial Model") +
  ggtitle("Scatter Plot of Sleeping") +
  theme(plot.title = element_text(hjust = 0.5))
# AIC/BIC
AIC(nbmodel_3)
# AIC = 6907.558
BIC(nbmodel_3)
# BIC = 6933.754
# Mean Square Error
residuals_3 <- residuals(nbmodel_3, type = "pearson")
mse_nb3 <- mean(residuals_3^2)
print(mse_nb3)
# MSE = 0.9982505
# Dev Stat.
Dev_nb3 <- 4536.9 - 3416.6
Dev_nb3
# Dev Stat. = 1120.3
# Graph of values for Sleeping
AIC <- c(AIC(loglin_model_3), AIC(cumulative_model_3),
         AIC(Pmodel_3), AIC(nbmodel_3))
BIC <- c(BIC(loglin_model_3), BIC(cumulative_model_3),
         BIC(Pmodel_3), BIC(nbmodel_3))
MSE <- c(mse_3, mse_CM3, mse_pb3, mse_nb3)
```

```
DEV <- c(Dev_LM3,Dev_CM3,Dev_Pb3,Dev_nb3)
best_model_index <- which.min(AIC)
best_model_index_1 <- which.min(BIC)
best_model_index_2 <- which.min(MSE)
best_model_index_3 <- which.min(DEV)
models <- c("Log-linear", "Cumulative Logit",
        "Poisson","Negative Binomial")
if(length(AIC) != length(BIC) || length(BIC) != length(MSE) ||
   length(MSE) != length(DEV)|| length(DEV) != length(models)){
  stop("Lengths of vectors don’t match.")}
AIC <- na.omit(AIC)
BIC <- na.omit(BIC)
MSE <- na.omit(MSE)
DEV <- na.omit(DEV)
# Set the plot size to accommodate longer labels
par(mar = c(5, 6, 4, 2) + 0.1, cex.axis = 0.8)
# Plot each vector separately
plot(AIC, type="b", xlab="Model", ylab="AIC",
            main="AIC Values", xaxt="n")
axis(1, at=1:length(models), labels=models)
abline(v=best_model_index, col="red")
plot(BIC, type="b", xlab="Model", ylab="BIC",
        main="BIC Values", xaxt="n")
```

```
axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_1, col="blue")

plot(MSE, type="b", xlab="Model", ylab="MSE",

        main="MSE Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_2, col="cyan")

plot(DEV, type="b", xlab="Model", ylab="DEV",

        main="DEV Values", xaxt="n")

axis(1, at=1:length(models), labels=models)

abline(v=best_model_index_3, col="purple")
```

VITA

MUSLIHAT ADEJOKE GAFFARI

Education: B.S. Mathematics, Osun State University
Osogbo, Osun State, Nigeria, October 2015
M.S. Mathematics(Applied), University of Lagos
Akoka, Lagos State, Nigeria, January 2022
M.S. Mathematical Sciences, East Tennessee State
University, Johnson City, Tennessee, August 2024

Professional Experience: Graduate Assistant, East Tennessee State University
Johnson City, Tennessee, 2022–2024