STATISTICAL METHODS FOR CLINICAL DATA: SURVIVAL ANALYSIS, LONGITUDINAL REGRESSIONS, AND BAYESIAN MONITORING

A THESIS

Presented to the Department of Mathematics and Statistics

California State University, Long Beach

In Partial Fulfillment

of the Requirements for the Degree

Master of Science in Applied Statistics

Committee Members:

Olga Korosteleva, Ph.D. (Chair)
Tianni Zhou, Ph.D
Xiyue Liao, Ph.D.

College Designee:

William Murray, Ph.D.

By Kevin E. Nguyen

B.S., 2019, California State University, Long Beach

May 2021

ABSTRACT

Implementation of clinical trials is a necessary step in increasing medical knowledge, such as providing information about the efficacy of an innovative medical device, procedure, or a medication. To establish the efficacy, human participants are carefully selected based on their characteristics suitable for the study. They are randomly assigned to either a treatment or control group and are monitored and measured over time to detect any physical changes. Clinical data obtained this way is vital in determining the efficacy of the tested product. In this thesis, we give an overview of a broad range of statistical methodology used in analysis of clinical data. We present techniques from survival analysis, longitudinal regression modeling, and Bayesian monitoring of clinical trials. For each method, we discuss theoretical framework and illustrate with an application to a suitable data set.

ACKNOWLEDGMENTS

I would like to thank all my parents and all my friends for their support in my educational journey. Most importantly, I want to thank my thesis advisor, Dr. Olga Korosteleva, for all of her support for writing my thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview of Clinical Trials

A clinical trial is a research study, or investigation, performed in human subjects with the purpose of evaluating the efficacy and/or safety of either an innovative medical device, medical procedure, or medication that is administered to the treatment group patients as compared to the control group patients who are administered a placebo or the standard therapy. The innovative treatment is tested before it gets marketed to the general consumer population.

1.1.1. Carrying Out a Clinical Trial

To briefly summarize the process, for a clinical trial procedure to be carried out, researchers must first select qualified candidates from a pool of individuals for their study. To do so, they must find candidates who have certain characteristics suitable for their study. Once qualified candidates are chosen, they are notified of their rights, benefits, and risks of participating in these trials. This process is known as the informed consent process. During this process, candidates (which we will now refer to as "patients") sign a consent form, confirming their intent to participate in the trial.

At enrollment for the clinical trial, patients receive an initial treatment. From there, patients are expected by the researchers to make follow-up visits, meaning that patients must check into the research lab in timely intervals specified by the researchers, to monitor any physical changes resulting from the treatment. After the last scheduled follow-up visit, each patient has the choice of continuing in the study or dropping out. Those who want to continue must sign another informed consent form, accepting the willingness of continuing in the study. Those who wish not to continue can drop out of the study. There are many reasons why a patient

would want to drop out of the study. These reasons include either adverse life events (such as mental health problems, contracting a certain disease, etc.), or voluntary discontinuation, meaning maybe the trial was not effective for the patient.

A clinical trial may be stopped earlier when a predetermined number of subjects have been accrued, or if the collected data strongly prove the efficacy of the tested treatment or show that the treatment is harmful. We will illustrate the concept later in Chapter 4.

1.1.2 Phases of a Clinical Trial

According to the National Institute of Aging (2020), clinical trials typically progress through four phases to test a treatment, find the appropriate dosage, and look for side effects. Now if, for example, after the first three phases, the researchers find a drug or medical device to be safe and effective, then the FDA approves it for clinical use and continues to monitor its effects.

Clinical trials are divided and described by their phase. These four phases are summarized as follows.

Phase I: A new product or treatment is tested on a small group of healthy subjects (typically 20-80 individuals) to determine its safety.

Phase II: This is where the initial clinical investigation begins. A phase II trial tests the product or treatment on a larger group of patients (typically 100-300 individuals) to determine, from preliminary data, the effectiveness of the drug on patients who have a certain disease or condition. In this phase, the drug's safety and side effects are closely monitored.

Phase III: More information about safety and effectiveness of either the drug or treatment is gathered. The test is usually carried out on a very large group of patients (typically 500-3000) with diverse backgrounds. In this stage, the new product is typically compared to either a

placebo or a standard treatment, where the side effects and efficacy are closely observed in the comparison. After this phase is completed, if the FDA agrees that the trial results are positive, it will approve the experimental drug or device. From there, the drug is then marketed to the consumer.

Phase IV: This phase occurs after the FDA approves the experimental drugs or devices for marketing. Known as the post-marketing surveillance phase, the device or drug's effectiveness and safety is then monitored in the general population after the product is marketed to collect additional information on the product's safety and efficacy over an extended period. Sometimes, the side effects may not become clear until more people have taken it over a longer period.

1.2. Literature Review

The use of statistical methods in clinical data has become an increasingly important topic through advances in healthcare and technology. In this thesis we present statistical techniques from survival analysis and regression analysis for longitudinal data.

Survival analysis focuses on modeling the distribution of the time until occurrence of an event (for example, death or remission) in the presence of censored observations. Observations are called censored if the person drops out of the study before occurrence of the event. Survival function for times to event (the probability of exceeding certain time) is modeled by a product-limit estimator (known commonly as the Kaplan-Meier estimator). Proposed by Kaplan and Meier (1958), this estimator is a step-function with discontinuities at the time of event. The Kaplan-Meier estimator, however cannot accommodate the presence of predictor variables.

To do so, Cox (1972) extended the work of Kaplan and Meier by proposing a regression model to fit life table data. His work, widely known in medical statistics as the Cox proportional

hazards model, assumes measurements on each individual are collected, and the relation between the distribution of time-to-event is modeled through what is known as a hazard function. Within the same year of Cox's publication, a statistician named Nelson (1972) developed the theory of hazard plotting. In his paper, the author presents an application to data plotting by modeling censored data through the use of a hazard function involving a cumulative distribution function.

Transitioning away from the field of survival analysis, longitudinal data analysis is another technique used in medical data for modeling repeated measures on specific individuals, over periods of times, ranging from a few days to even a few years. In their paper, Caruana et al. (2015) state that there are generally two types of designs incorporating longitudinal data: the standard longitudinal design and the cross-sectional design. The longitudinal design collects data repeatedly over time on the same individuals, whereas the cross-sectional design measures multiple variables at a single time point without regard to the influence of time on these variables. Both the standard longitudinal and cross-sectional designs have distinct advantages. An advantage of a standard longitudinal design is that it is useful for evaluating the relationship between risk factors causing a disease and the outcomes of a clinical trial treatment observed over a period of time. On the other hand, an advantage of using a cross-sectional study is that it is quicker to set up which may be useful in performing quicker evaluations.

It is often noteworthy to know that clinical trials are often very expensive to carry out and require lots of optimization. Gupta (2012) noted that there are two statistical methods used in evaluation of efficacy of clinical trials: frequentist and Bayesian. Frequentist methods model the prior information through the design of the clinical trial, but not the analysis of data. Bayesian, on the other hand, provides a mathematical method of calculating the likelihood of future events,

given knowledge of past events. In this thesis, we discuss both frequentist and Bayesian analysis techniques.

1.3. A Brief History of Clinical Trials

Clinical research has been ever evolving throughout the centuries. The first documented clinical trial (non-medical) happened during circa 500 BCE in Babylon. During that time, King Nebuchadnezzar (Bhatt 2010) ordered his people to eat meat and drink wine, believing the diet would keep them in good shape; however, several citizens objected to his rule by eating vegetables instead. Therefore, Nebuchadnezzar experimented by allowing 10 days for the objectors to follow a diet of legumes. After 10 days, the king found out that the vegetarians were more nourished than the meat-eaters. Thus, the king permitted the vegetarians to continue their diet.

As time progresses, more advances have been made to clinical trials. A noteworthy example was the development of a controlled trial. Discovered by James Lind (Bartholomew, 2002) in the mid-18th century, a controlled trial is a study design that randomly places participants into an experimental group or control group. The difference between the outcome variables in the two groups is then investigated.

After the basic approaches to clinical trials were introduced in the 18th century, the efforts were made to refine both the design as well as the statistical aspects. It took another century of research before the emergence of a significant milestone in clinical trials, known as the placebo. First introduced in the early 1800s, the placebo was referred to as “an epithet given to any medicine more to please than benefit the patient” (Shapiro 1964, p. 52). Putting this quote into layman’s terms, the placebo is used in clinical research as a “dummy drug” to instill

confidence within a patient, making them believe that the treatment might work. The effects of placebos are then compared to an active treatment to establish validity.

As scientific advances continued to occur in clinical trials, new ethical and regulatory challenges emerged. The ethical framework for human experimentation dates back to the ancient Hippocratic Oath, which mentions that the primary duty of a physician is to avoid harming the patient; this oath, however, it was violated many times in past human experimentations (such as Nazi human experimentations during World War II). To solve this issue, several laws and regulations were created in the mid-20th century. The first was the Nuremberg Code of 1947 which stressed voluntary consent in clinical trials. Another notable one was the Helsinki Declaration created in 1964. Widely regarded as the cornerstone document on human research ethics, the Helsinki Declaration established regulatory guidelines outlining the general principals in clinical trials as well as the rights, risks, and privacy of using humans in medical research (Bhatt, 2010).

Today clinical trials have been heavily regulated by the government as a response to ethical guidelines. Founded in 1862, the FDA has evolved as one of the world's foremost institutional authorities for conducting and evaluating controlled clinical drug trials (Davies and Kermani 2008). For a new drug to be marketed, the FDA requires that at least two adequate and well-controlled clinical trials be conducted to provide substantial evidence regarding the efficacy of the drug product under investigation (Davies and Kermani 2008).

1.4. Aims and Objectives for Thesis

In this thesis, we explore properties of clinical data using various statistical methods. All datasets and scenarios are simulated except for the publicly available Primary Biliary Cirrhosis dataset which we will analyze in Chapter 2. Understanding these techniques is not only

necessary for clinical trials, but other related fields such as epidemiology. Below we summarize to topics that this thesis explores.

Chapter 2 is devoted to survival analysis methodology. In the field of survival analysis, we estimate the distribution of time until an adverse event and compare the distributions in two or more distinct groups using the Kaplan-Meier estimator of the survival function and log-rank test. Furthermore, we dive deeper into investigating how certain factors can influence the rate of a particular event happening (such as infection or death) through the use of the Cox proportional hazards model.

In medical data, most measurements are collected through a longitudinal study, meaning repeated observations of the same variables in the same individuals are collected through periods of time. Chapter 3 presents regression models for longitudinal data from different settings where the response variables have normal, gamma, binary, and Poisson distributions.

In Chapter 4 we discuss interim data monitoring in clinical trials. With a standard approach, clinical trials continue until the pre-determined number of patients has been accrued and followed for a certain period of time. The required number of patients is determined based on an acceptable power of the statistical test of superiority of the product under investigation. However, if researchers have a strong belief in superiority of the tested product, they might conduct a sequential testing that allows stopping a trial earlier if the data show enough evidence. Chapter 4 explores the concept of interim data monitoring where both the classical group sequential testing procedure and Bayesian sequential procedure are discussed and illustrated with several clinical trial examples.

CHAPTER 2

SURVIVAL ANALYSIS

2.1. Theoretical Framework

2.1.1 The Survival, Hazard, and Cumulative Hazard Functions

Let *T* denote the survival time of an individual. Assume that T is a random variable with the probability density function $f(x)$. The density function $f(x)$ along with the cumulative distribution function $F(x) = \int_0^x f(u)du$ does not provide much information about the individual's chance of survival past a fixed time t. Therefore, the survival, hazard, and cumulative hazard functions are used instead.

The survival function is given as $S(t) = 1 - F(t) = P(T > t)$. Therefore $S(t)$ is the probability of survival past time t. The relation between $f(t)$ and $S(t)$ is derived as $f(t) = F'(t) = \big(1 - S(t)\big)' = -S'(t)$.

The hazard function is defined as the instantaneous rate of failure, given that the individual has survived past time t. The expressions that connect the hazard function with $f(t), F(t)$, and $S(t)$ are obtained as follows:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t<T\leq t+\delta t \mid T>t)}{\delta t} = \lim_{\delta t \to 0} \frac{P(t<T\leq t+\delta t)}{\delta t P(T>t)} = \frac{1}{S(t)} \lim_{\delta t \to 0} \frac{F(t+\delta t)-F(t)}{\delta t} = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -(\ln S(t))'.$$

In addition to the hazard function, it is sometimes convenient to operate with the cumulative hazard function defined as $H(t) = \int_0^t h(u)du = -\int_0^t \frac{d \ln S(u)}{du} du = -\int_0^t d \ln S(u) = -\ln S(t) + \ln S(0) = -\ln S(t)$ since $S(0) = 1$ and $\ln(1) = 0$.

2.1.2. The Kaplan-Meier Estimator

The Kaplan-Meier estimator, also known as the product-limit estimator, is a nonparametric method widely used to estimate the survival function from lifetime data. The specificity of the Kaplan-Meier estimator is that it can accommodate censored observations. If for an individual the lifetime data are not observed until an event but rather until the individual drops out of the study, the lifetime observation is referred to as right-censored. It is known that the individual hasn't experienced an event up to certain time, and nothing can be said about survival of the individual past that time. In terms of medical applications, the Kaplan-Meier estimator is used to estimate the survival curve of every patient that is followed until death or censoring.

The estimator of the survival function $S(t)$, the probability that life is longer than t, is given by $\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$ where t_i is the time of the ith event occurring, d_i is the number of individuals experiencing the event at time t_i, and n_i is the number of individuals who have survived up until just before time t_i. These individuals are termed "at risk" at time t_i. This expression is nothing more than the maximum likelihood estimator derived as follows:

Let $0 = t_0 < t_1 < t_2 < \cdots < t_k$ denote the k distinct event times. Let $\pi_i = P(T > t_i \mid T > t_{i-1}), i = 1, \ldots, k,$ be the conditional probabilities that an individual survives past time t_i , given that the individual has survived past time t_{i-1}. The survival function $S(t)$ at time t_j is the product of π_i's; that is, $S(t_j) = \prod_{i=1}^{j} \pi_i$. The probabilities π_i can be estimated by the method of maximum likelihood. Each of the d_i individuals who experience an event at time t_i contributes a $1 - \pi_i$ term to the likelihood function, whereas each of the $n_i - d_i$ individuals who survive past time t_i contributes a π_i term to the likelihood function. Consequently, the

likelihood function has the form $L = \prod_{i=1}^{k}(1-\pi_i)^{d_i}\pi_i^{n_i-d_i}$. The log-likelihood function becomes $\ln L = \sum_{i=1}^{k} d_i \ln(1-\pi_i) + \sum_{i=1}^{k}(n_i - d_i)\ln \pi_i$. Differentiating with respect to π_i and setting the derivative to zero, we see that the estimators $\hat{\pi}_i$'s solve $\frac{d \ln L}{d\pi_i}|_{\pi_i=\hat{\pi}_i} = 0 = -\frac{d_i}{1-\hat{\pi}_i} - \frac{n_i-d_i}{\hat{\pi}_i}$. From here, $\hat{\pi}_i = 1 - \frac{d_i}{n_i}$ and $\hat{S}(t_j) = \prod_{i=1}^{j}\left(1-\frac{d_i}{n_i}\right)$. Now take any t. Since $t_j \leq t \leq t_{j+1}$ for some $j = 0, \dots, k,$ and there are no events in the open interval (t_j, t_{j+1}), we have that $S(t) = S(t_j)$ and the result follows.

To find the variance of the Kaplan-Meier estimator, we note that $d_i's$ can be assumed to follow a binomial distribution with parameters n_i and probability of event π_i where we estimate $\hat{\pi}_i = \frac{d_i}{n_i}$. We can approximate $\widehat{Var}(\hat{\pi}_i) \approx \frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}$. Further, consider $\ln[\hat{S}(t_j)] = \sum_{i=1}^{j}\ln(1-\hat{\pi}_i)$. Now we will apply the delta method that states that if $\{X_n, n \geq 1\}$ is a sequence of random variables such that $\sqrt{n}(X_n - \theta) \xrightarrow[n\to\infty]{} N(0, \sigma^2),$ in distribution, and there exists a function $g(x)$ that is differentiable at θ where $g'(\theta) \neq 0$, then $\sqrt{n}(g(X_n) - g(\theta)) \xrightarrow[n\to\infty]{} N\left(0, \sigma^2\left(g'(\theta)\right)^2\right)$. By the delta method and the approximate independence of the $\hat{\pi}_i's$,

$$\widehat{Var}(\ln[\hat{S}(t_j)]) = \sum_{i=1}^{j}\widehat{Var}[\ln(1-\hat{\pi}_i)] = \sum_{i=1}^{j}\left(\frac{1}{1-\hat{\pi}_i}\right)^2\widehat{Var}(\hat{\pi}_i) = \sum_{i=1}^{j}\left(\frac{1}{1-\hat{\pi}_i}\right)^2\frac{\hat{\pi}_i(1-\hat{\pi}_i)}{n_i}$$

$$= \sum_{i=1}^{j}\frac{\hat{\pi}_i}{(1-\hat{\pi}_i)n_i} = \sum_{i=1}^{j}\frac{\frac{d_i}{n_i}}{\left(1-\frac{d_i}{n_i}\right)n_i} = \sum_{i=1}^{j}\frac{d_i}{(n_i-d_i)n_i}.$$

Finally, recalling that for $t, t_j \le t \le t_{j+1,}\ S(t) = S(t_j)$, and writing $\hat{S}(t) = e^{\ln(\hat{S}(t))} = e^{\ln(\hat{S}(t_j))}$, we use the delta method again to obtain the Greenwood's formula for variance $\widehat{Var}\left(\hat{S}(t)\right) = \widehat{Var}\left(\hat{S}(t_j)\right) = \left[\hat{S}(t_j)\right]^2 \widehat{Var}[\ln(\hat{S}(t_j))] = \left[\hat{S}(t)\right]^2 \sum_{i:t_i \le t} \frac{d_i}{(n_i - d_i)n_i}$.

2.1.3. Kaplan-Meier Survival Curve

The Kaplan-Meier survival curve is the plot of the Kaplan-Meier estimator of the survival function $\hat{S}(t)$ against time t. $\hat{S}(t)$ is represented as a step function that decreases at the times of events, and remains constant between two observed event times. Traditionally, event times for censored observations are denoted by an "x", and if an observation happens to be censored at an event time, the "x" is placed at the bottom of the step.

2.1.4. The Nelson-Aalen Estimator

The Nelson-Aalen estimator is a nonparametric estimator of the cumulative hazard rate function from censored event data. Let $t_1 < t_2 < \cdots < t_n$ represent the times of events, d_i be the number of observed events at time t_i, and n_i be the number of subjects at risk. The Nelson-Aalen estimator for the cumulative hazard rate is given by $\tilde{H}(t) = \Sigma_{t_i \le t} \frac{d_i}{n_i}$.

To derive this estimator, consider the relation $S(t) = \exp(-H(t))$ and the Kaplan-Meier estimator $\hat{S}(t) = \prod_{i:\, t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$ as an estimator of $S(t)$. If $d_i \ll n_i$, then $\ln\left(1 - \frac{d_i}{n_i}\right) \approx -\frac{d_i}{n_i}$, and therefore, $\tilde{H}(t) = -\ln\left(\hat{S}(t)\right) = -ln \prod_{i:\, t_i \le t} \left(1 - \frac{d_i}{n_i}\right) = -\sum_{i:\, t_i \le t} \ln\left(1 - \frac{d_i}{n_i}\right) \approx \sum_{i:\, t_i \le t} \frac{d_i}{n_i}$. Note that from here, the Nelson-Aalen estimator of the survival function has the form $\tilde{S}(t) = \exp\left(-\sum_{t_i \le t} \frac{d_i}{n_i}\right)$. The variance of $\tilde{H}(t)$ can be approximated by $\widehat{Var}[-\ln(\hat{S}(t))] = \widehat{Var}[\ln(\hat{S}(t_j))] = \sum_{i=1}^{j} \frac{d_i}{(n_i - d_i)n_i}$, which we obtained above on our way to the Greenwood's formula.

2.1.5. The Log-Rank Test

The log-rank test is a test of hypotheses that compares survival functions as functions of time for two categories, for example, survival functions for men vs. women or for intervention group vs. control group. The objective is to test $H_0: S_1(t) = S_2(t)$ for all $t \geq 0$ against $H_1: S_1(t) \neq S_2(t)$ for some $t \geq 0$. Below we derive the expression for the chi-squared test statistic. Let $t_1 < t_2 < \cdots < t_k$ denote the event times, and let d_{1i} and d_{2i} be the number of individuals who experience the event at time t_i in categories 1 and 2, respectively. Also denote by n_{1i} and n_{2i} the number of individuals at risk at time t_i in categories 1 and 2, respectively. We have that $d_{1i} + d_{2i} = d_i$, the total number of individuals who experienced the event at time t_i, and $n_{1i} + n_{2i} = n_i$, the total number of at-risk individuals at time t_i. Under the null hypothesis, d_{1i} has a hypergeometric distribution with mean $E(d_{1i}) = \frac{n_{1i}\, d_i}{n_i}$ and variance $Var(d_{1i}) = \frac{n_{1i} n_{2i} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}$, $i = 1, \ldots, k$. The test statistic is defined as $\chi^2 = \left(\frac{U}{\sqrt{Var(U)}} \right)^2$ where $U = \sum_{i=1}^{k} (d_{1i} - E(d_{1i}))$ and $Var(U) = \sum_{i=1}^{k} \frac{n_{1i} n_{2i} (n_i - d_i) d_i}{n_i^2 (n_i - 1)}$. Under the null hypothesis, the test statistic has a χ^2-distribution with one degree of freedom.

2.1.6. Parametric Estimation of Survival Function

2.1.6.1. Definition

The survival function $S(t)$ is estimated by a parametric method if an explicit algebraic expression is for this function is assumed known and the parameters are estimated from the data. For example, the survival function for a Weibull distribution is widely implemented. A Weibull distribution has the density $f(t) = \alpha \lambda t^{\alpha - 1} e^{-\lambda t^{\alpha}}$ for $t \geq 0$ and $\alpha, \lambda > 0$, which for $\alpha = 1$ reduces to an exponential distribution with the density $f(t) = \lambda e^{-\lambda t}, \lambda > 0, t \geq 0$. The estimator

for the survival function of a Weibull distribution is $\hat{S}(t) = e^{-\lambda t^{\alpha}}, t \geq 0$, which reduces to an exponential survival function $\hat{S}(t) = e^{-\lambda t}, t \geq 0$, when $\alpha = 1$.

The parameters are estimated by the method of maximum likelihood. However, since censored data are present, they have to be taken into consideration when deriving the likelihood function.

2.1.6.2. Random Censoring Model

A random censoring model assumes that times to event and censoring times are independent. Denote by T_i the time to event of the ith subject, and let C_i be the censoring time of the ith subject. We assume that T_i has pdf $f_i(t)$ and cdf $F_i(t)$, and C_i has pdf $g_i(t)$ and cdf $G_i(t)$. The subjects for which an event occurs (termed uncensored), the time to event is smaller than the censoring time, that is, $T_i < C_i$, while for censored observations, $C_i < T_i$.

Thus, the contribution to the likelihood function of an uncensored ith subject with the observed event time t_i is $\lim_{dt\to 0} P(T_i \in (t_i, t_i + dt),\ C_i > t_i)/dt = f_i(t)\big(1 - G_i(t)\big)$, and the contribution of the ith subject censored at time t_i is $\lim_{dt\to 0} P(C_i \in (t_i, t_i + dt),\ T_i > t_i)/dt = g_i(t)\big(1 - F_i(t)\big)$. Let $\delta_i = 1$ if the ith observation is uncensored and 0, otherwise. Therefore, the likelihood function for the survival with random censoring is:

$$L = \prod_{i=1}^{n} [f_i(t_i)(1 - G(t_i))]^{\delta_i} [\big(1 - F_i(t_i)\big) g_i(t_i)]^{1-\delta_i}$$

$$= \prod_{i=1}^{n} (1 - G(t_i))^{\delta_i} (g(t_i))^{1-\delta_i} \prod_{i=1}^{n} f_i(t_i)^{\delta_i} \big(1 - F_i(t_i)\big)^{1-\delta_i}$$

$$\propto \prod_{i=1}^{n} f_i(t_i)^{\delta_i} \big(1 - F_i(t_i)\big)^{1-\delta_i}.$$

The log-likelihood function is proportional to

$$\ln L \ \propto \ \sum_{i=1}^{n} \delta_i \ln f_i(t_i) + \sum_{i=1}^{n} (1 - \delta_i) \ln(1 - F_i(t_i)).$$

By maximizing this function, we find the parameters of the cdf F and estimate the survival function as $\hat{S}(t) = 1 - \hat{F}(t), t \geq 0$.

2.1.6.3. The Weibull Distribution Model

Suppose the time to event T_i has a Weibull distribution with pdf $f(t) = \alpha\lambda t^{\alpha-1}e^{-\lambda t^{\alpha}}, \alpha, \lambda > 0, t \geq 0$, and cdf $F(t) = e^{-\lambda t^{\alpha}}, \alpha, \lambda > 0, t \geq 0$. We estimate the parameters α and λ by the method of maximum likelihood. The log-likelihood function is proportional to

$$\ln L(\alpha, \lambda) \propto \sum_{i=1}^{n} \delta_i \ln f_i(t_i) + \sum_{i=1}^{n} (1 - \delta_i) \ln\big(1 - F_i(t_i)\big)$$

$$= \sum_{i=1}^{n} \delta_i \ln(\alpha\lambda {t_i}^{\alpha-1} e^{-\lambda t_i^{\alpha}}) + \sum_{i=1}^{n} (1 - \delta_i) \ln\big(e^{-\lambda t_i^{\alpha}}\big)$$

$$= \ln\alpha \sum_{i=1}^{n} \delta_i + \ln\lambda \sum_{i=1}^{n} \delta_i + (\alpha - 1) \sum_{i=1}^{n} \delta_i \ln t_i - \lambda \sum_{i=1}^{n} \delta_i t_i^{\alpha} - \lambda \sum_{i=1}^{n} (1 - \delta_i) t_i^{\alpha}$$

$$\propto \ln\alpha \sum_{i=1}^{n} \delta_i + \ln\lambda \sum_{i=1}^{n} \delta_i + \alpha \sum_{i=1}^{n} \delta_i \ln t_i - \lambda \sum_{i=1}^{n} t_i^{\alpha}.$$

Thus, the maximum-likelihood estimators $\hat{\alpha}$ and $\hat{\lambda}$ are numeric solutions to the system of normal equations

$$\begin{cases} \dfrac{\partial \ln L(\hat{\alpha}, \hat{\lambda})}{\partial \hat{\alpha}} = 0 = \dfrac{\sum_{i=1}^{n} \delta_i}{\hat{\alpha}} + \displaystyle\sum_{i=1}^{n} \delta_i \ln t_i - \hat{\lambda} \sum_{i=1}^{n} t_i^{\hat{\alpha}} \ln t_i \ , \\ \dfrac{\partial \ln L(\hat{\alpha}, \hat{\lambda})}{\partial \hat{\lambda}} = 0 = \dfrac{\sum_{i=1}^{n} \delta_i}{\hat{\lambda}} - \displaystyle\sum_{i=1}^{n} t_i^{\hat{\alpha}} \ . \end{cases}$$

The estimated survival function has the form $\hat{S}(t) = e^{-\hat{\lambda} t^{\hat{\alpha}}}, t \geq 0$.

2.1.7. The Weibull Regression Model

The Weibull regression model estimates the survival function as $\hat{S}(t) = e^{-\hat{\lambda}t^{\hat{\alpha}}}$, $t \geq 0$, where $\hat{\lambda} = e^{-(\hat{\beta}_0+\hat{\beta}_1x_1+\cdots+\hat{\beta}_kx_k)/\hat{\sigma}}$ and $\hat{\alpha} = 1/\hat{\sigma}$. The maximum-likelihood estimates $\hat{\sigma}, \hat{\beta}_0, \ldots, \hat{\beta}_k$ are the numeric solutions of the normal equations where the log-likelihood function f has the form:

$$\ln L(\beta_0, \ldots, \beta_k, \sigma) \propto -\ln \sigma \sum_{i=1}^{n} \delta_i - \frac{1}{\sigma}\sum_{i=1}^{n}[\delta_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik})] + \left(\frac{1}{\sigma} - 1\right)\sum_{i=1}^{n} \delta_i \ln t_i$$

$$- \sum_{i=1}^{n} \exp\left[\frac{1}{\sigma}(\ln t_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}))\right].$$

To check goodness of fit of the fitted model, the deviance test is employed. The test statistic, called the deviance, is computed as

$$deviance = -2(\ln L(null\ model) - \ln L(fitted\ model))$$

where the fitted model is the full model with $k + 1$ regression coefficients and the scale parameter σ. The null model is the intercept-only model with β_0 and σ as parameters.

Under H_0, the test statistic follows a χ^2- distribution with the number of degrees of freedom calculated as the difference between the number of parameters in the two models; that is, the number of degrees of freedom is $k + 2 - 2 = k$, the same as the number of predictors in the fitted model.

2.1.8 The Cox Proportional Hazards Model

The Cox proportional hazards model is customarily defined in terms of the hazard function. It is written as $h(t, x_1, \ldots, x_k, \beta_1, \ldots, \beta_k) = h_0(t) \exp(\beta_1 x_1 + \cdots + \beta_k x_k)$ where $h_0(t)$ is called the baseline hazard function. It represents the hazard function when all the predictors are equal to zero, which corresponds to an often-hypothetical individual called a baseline

individual. The quantity exp $(\beta_1 x_1 + \cdots + \beta_k x_k)$ is referred to as the relative risk of an individual with predictors $x_1, \ldots, x_k$. The term "proportional hazards" symbolizes the fact that in this model, the ratio of the hazard functions for two individuals with relative risks exp $(\beta_1 x_{11} + \cdots + \beta_k x_{1k})$ and exp $(\beta_1 x_{21} + \cdots + \beta_k x_{2k})$ is a constant, not depending on time:

$$\frac{h(t,x_{11},\ldots,x_{1k},\beta_1,\ldots,\beta_k)}{h(t,x_{21},\ldots,x_{2k},\beta_1,\ldots,\beta_k)} = \frac{h_0(t)\exp(\beta_1 x_{11}+\cdots+\beta_k x_{1k})}{h_0(t)\exp(\beta_1 x_{21}+\cdots+\beta_k x_{2k})} = \frac{\exp(\beta_1 x_{11}+\cdots+\beta_k x_{1k})}{\exp(\beta_1 x_{21}+\cdots+\beta_k x_{2k})} = constant.$$

Alternatively, the Cox proportional hazards model can be formulated in terms of the survival function. To derive the alternative definition, we note that

$$S(t) = \exp\left(-\int_0^t h(u|x_1, x_2, \ldots, x_k, \beta_1, \ldots, \beta_k)du\right)$$

$$= \exp\left(-\int_0^t h_0(u)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)\,du\right) = [S_0(t)]^r$$

where $r = \exp(\beta_1 x_1 + \cdots + \beta_k x_k)$ is the relative risk, and $S_0(t) = e^{-\int_0^t h_0(u)du}$ is the baseline survival function for the baseline individual.

The regression coefficients $\beta_1, \ldots, \beta_k$ can be estimated by maximizing the partial-likelihood function, which is defined as the portion of the likelihood function in the random censoring model (see Section 2.1.6.2) that does not depend on time t. To derive the expression for the partial-likelihood function, we proceed as follows. We start with the multiplicative factor of the likelihood function that depends only on the distribution of the time-to-event:

$$L = \prod_{i=1}^n f_i(t_i)^{\delta_i}\left(1 - F_i(t_i)\right)^{1-\delta_i}.$$

Next, we let the time-to-event distribution for the ith subject have the survival function $S_i(t)$, and hazard function $h_i(t_i) = h_0(t_i)\exp(\beta_1 x_1 + \cdots + \beta_k x_k)$, $i = 1 \ldots., n$. Using the expression $f_i(t_i) = h_i(t_i)S_i(t_i)$, we obtain

$$L = \prod_{i=1}^{n} f_i(t_i)^{\delta_i} \left(S_i(t_i)\right)^{1-\delta_i} = \prod_{i=1}^{n} (h_i(t_i) S_i(t_i))^{\delta_i} \left(S_i(t_i)\right)^{1-\delta_i}$$

$$= \prod_{i=1}^{n} (h_i(t_i))^{\delta_i} S_i(t_i) = \prod_{i=1}^{n} \left(\frac{h_i(t_i)}{\sum_{j \in R(t_i)} h_j(t_i)} \right)^{\delta_i} \left(\sum_{j \in R(t_i)} h_j(t_i) \right)^{\delta_i} S_i(t_i)$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}}}{\sum_{j \in R(t_i)} \mathrm{e}^{\beta_1 x_{j1} + \cdots + \beta_k x_{jk}}} \right)^{\delta_i} \cdot \prod_{i=1}^{n} \left(\sum_{j \in R(t_i)} h_j(t_i) \right)^{\delta_i} S_i(t_i) \, .$$

Here R_i represents the relative-risk set at time t_i. Now, we discard the portion that depends on times $t_i, i = 1, \dots, n$, and define the partial-likelihood function as

$$L_p(\beta_1, \dots, \beta_n) = \prod_{i=1}^{n} \left(\frac{e^{\beta_1 x_{i1} + \cdots + \beta_k x_{ik}}}{\sum_{j \in R(t_i)} \mathrm{e}^{\beta_1 x_{j1} + \cdots + \beta_k x_{jk}}} \right)^{\delta_i}.$$

Next, the estimators of $\beta_1, \dots, \beta_k$ are obtained by maximizing the log-partial-likelihood function

$$\ln L_p(\beta_1, \dots, \beta_k) = \sum_{i=1}^{n} \delta_i (\beta_1 x_{i1} + \cdots + \beta_k x_{ik}) - \sum_{i=1}^{n} \delta_i \ln \left(\sum_{j \in R(t_i)} \mathrm{e}^{\beta_1 x_{j1} + \cdots + \beta_k x_{jk}} \right).$$

The estimates $\hat{\beta}_1, \dots, \hat{\beta}_k$ are numerical solutions of the partial-likelihood score equations

$$\frac{\partial \ln L_p(\hat{\beta}_1, \dots, \hat{\beta}_k)}{\partial \beta_m} = \sum\nolimits_{i=1}^{n} \delta_i x_{im} - \sum\nolimits_{i=1}^{n} \delta_i \frac{\sum_{j \in R(t_i)} x_{jm} e^{(\hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_k x_{jk})}}{\sum_{j \in R(t_i)} e^{(\hat{\beta}_1 x_{j1} + \cdots + \hat{\beta}_k x_{jk})}} = 0, \qquad m = 1, \dots, k.$$

The estimates of the regression coefficients $\hat{\beta}_1, \dots, \hat{\beta}_k$ for the Cox proportional hazards model yield the following interpretation. For a numeric predictor, say, x_1, the percent change in the estimated hazard function when $x_1$is increased by one unit is equal to

$$\frac{\hat{h}(t, x_1 + 1, x_2, \dots, x_k, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k) - \hat{h}(t, x_1, x_2, \dots, x_k, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)}{\hat{h}(t, x_1, x_2, \dots, x_k, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)} \cdot 100\%$$

$$= \left(\frac{\hat{h}_0(t) \exp(\hat{\beta}_1 (x_1 + 1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)}{\hat{h}_0(t) \exp(\hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)} - 1 \right) \cdot 100\% = \left(e^{\hat{\beta}_1} - 1 \right) \cdot 100\%.$$

If x_1 is a 0-1 predictor, the percent ratio of the estimated hazard functions for $x_1 = 1$ and $x_1 = 0$ is equal to

$$\frac{\hat{h}(t, 1, x_2, \dots, x_k, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)}{\hat{h}(t, 0, x_2, \dots, x_k, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k)} \cdot 100\% = \frac{\hat{h}_0(t) \exp(\hat{\beta}_1 \cdot 1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)}{\hat{h}_0(t) \exp(\hat{\beta}_1 \cdot 0 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k)} \cdot 100\%$$

$$= e^{\hat{\beta}_1} \cdot 100\%.$$

Remark: A noteworthy parametric model, sharing characteristics of the Cox proportional hazards model, is the Weibull regression (see Section 2.1.7) that models the survival function as $S(t) = e^{-\lambda t^{\alpha}}$, $t \geq 0$, where $\lambda = e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)/\sigma}$ and $\alpha = 1/\sigma$. To give interpretation of the estimated regression coefficients, we note that the hazard function of the Weibull distribution is of the form $h(t) = -\frac{S'(t)}{S(t)} = \alpha \lambda t^{\alpha - 1} = h_0(t) \exp\{\beta_1^* x_1 + \cdots + \beta_k^* x_k\}$ where $h_0(t) = \frac{1}{\sigma} e^{-\frac{\beta_0}{\sigma}} t^{\frac{1}{\sigma} - 1}$, and $\beta_i^* = -\frac{\beta_i}{\sigma}, i = 1, \dots, k$. This shows that the Weibull regression is a special case of the Cox proportional hazards model, and $\hat{\beta}_i^* = -\frac{\hat{\beta}_i}{\hat{\sigma}}, i = 1, \dots, k$, are interpreted as in the Cox model, in terms of estimated percent change (or percent ratio) of the hazard function.

2.2 Application of the Survival Analysis

2.2.1. Data Description

The Primary Biliary Cirrhosis dataset presents a clinical trial of a liver disease conducted between the years 1974 and 1984. This dataset was obtained publicly through kaggle.com. The goal of this study was to determine the effectiveness of a placebo drug, known as Penicillamine, on the survival of the patients. The study originally consisted of a total of 424 patients; however, 112 cases did not participate in the clinical trial. Therefore, the patients not participating were omitted from the dataset, reducing the dataset size down to 312 patients. The table below (see Table 1) lists all the variables used in the analysis, along with their attributes.

TABLE 1. Description of the Variables in the Primary Biliary Cirrhosis Dataset

Name	Description	Type	Values
time	The number of days between registration and event	Numeric	Ranges from 41 to 4795 days until event
status	Status of patient	Categorical Numeric	0=Alive and doesn't need liver transplant 1=Alive but needs liver transplant 2=Dead (i.e., censored)
trt	Type of drug that patient received	Binary categorical	1=D-Penicillamine 2=Placebo
age	Age of patients in years	Numeric	Ranges from 26 to 78 years old
sex	Sex of patient	Binary categorical	0=Female, 1=Male
ascites	Presence of ascites, the accumulation of fluid in the peritoneal cavity	Binary categorical	0=No 1=Yes
hepato	Abnormal enlargement of the liver not related to the underlying disease.	Binary categorical	0=No 1=Yes
spiders	Blood vessel malformations in the skin	Binary categorical	0=No 1=Yes
edema	Swelling caused by excess fluid trapped in the body's tissues	Numeric categorical	0=No edema present, therefore no diuretic therapy is needed 0.5=Edema is present without diuretics 1=Edema is present despite diuretic therapy
bilirubin	Amount of Bilirubin, a yellowish pigment made from the breakdown of red blood cells, in milligrams per deciliter of blood (mg/dl)	Numeric	Ranges between 0.3mg/dl to 28mg/dl
cholesterol	Amount of cholesterol, in milligrams per deciliter of blood (mg/dl)	Numeric	Ranges between 120mg/dl to 1,775mg/dl
albumin	Amount of albumin, a protein made by the liver to prevent blood fluids from leaking into other tissues, in milligrams per deciliter of blood (mg/dl)	Numeric	Ranges between 1.96mg/dl to 4.64mg/dl

TABLE 1. Continued

Name	Description	Type	Values
copper	Amount of copper (in micrograms per day, μg/day)	Numeric	Ranges between 4μg/day to 588 μg/day
alk.phos	Alkaline phosphate concentration in U/Liter	Numeric	Ranges between 289U/Liter to 13,862U/Liter
ast	Aspartate aminotransferase (AST) concentration in U/ml	Numeric	Ranges between 26.35U/ml to 457.25U/ml
trig	Triglyceride concentration in mg/dl	Numeric	Ranges between 33mg/dl to 598mg/dl
platelet	Amount of platelets per cubic ml/1000	Numeric	Ranges between 62ml/liter to 563ml/liter
protime	Prothrombin time, the amount of time it takes for blood to clot	Numeric	Ranges between 9 seconds to 17.1 seconds
stage	Histologic stage of disease which describes how much damage has been done to the liver. Stage 1=Inflammation and damage to the walls of medium-sized bile ducts Stage 2=Blockage of small bile ducts Stage 3=Beginning of scarring Stage 4=Permanent Cirrhosis has developed, resulting in severe damage to the liver	Categorical Numeric	Stages 1,2,3, and 4
copper	Amount of copper (in micrograms per day, μg/day)	Numeric	Ranges between 4μg/day to 588 μg/day
alk.phos	Alkaline phosphate concentration in U/Liter	Numeric	Ranges between 289U/Liter to 13,862U/Liter
ast	Aspartate aminotransferase (AST) concentration in U/ml	Numeric	Ranges between 26.35U/ml to 457.25U/ml
trig	Triglyceride concentration in mg/dl	Numeric	Ranges between 33mg/dl to 598mg/dl
platelet	Amount of platelets per cubic ml/1000	Numeric	Ranges between 62ml/liter to 563ml/liter

2.2.2. Kaplan-Meier Estimator

The Kaplan-Meier survival curve was fitted. The plot of the survival curve along with the confidence band is given in Figure 1 below.

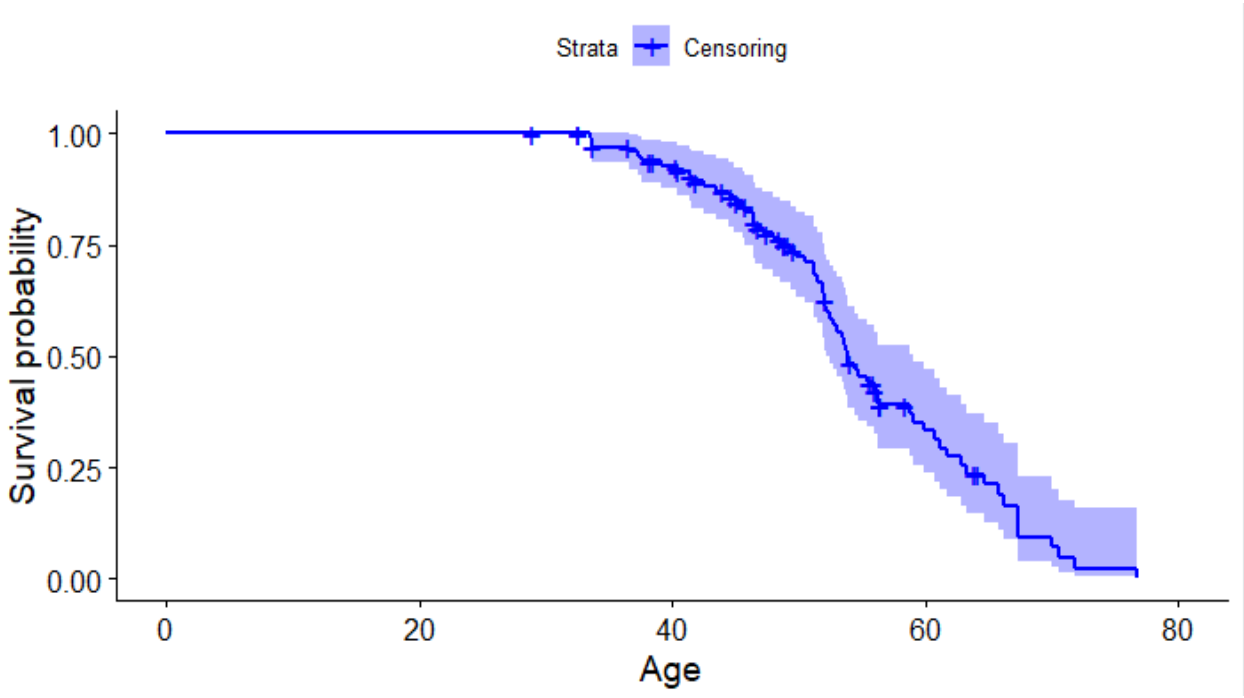

FIGURE 1. The Kaplan-Meier survival curve.

From the table and the graph, there is a 100% survival up to age 25, after which patients start dying. Only about 60% of cohort survive past age 56, and around 20% are still alive at about age 69. The survival curve exhibits roughly linear downward trend, with a slight curvature downward and then upward. The change of curvature occurs around age 55.

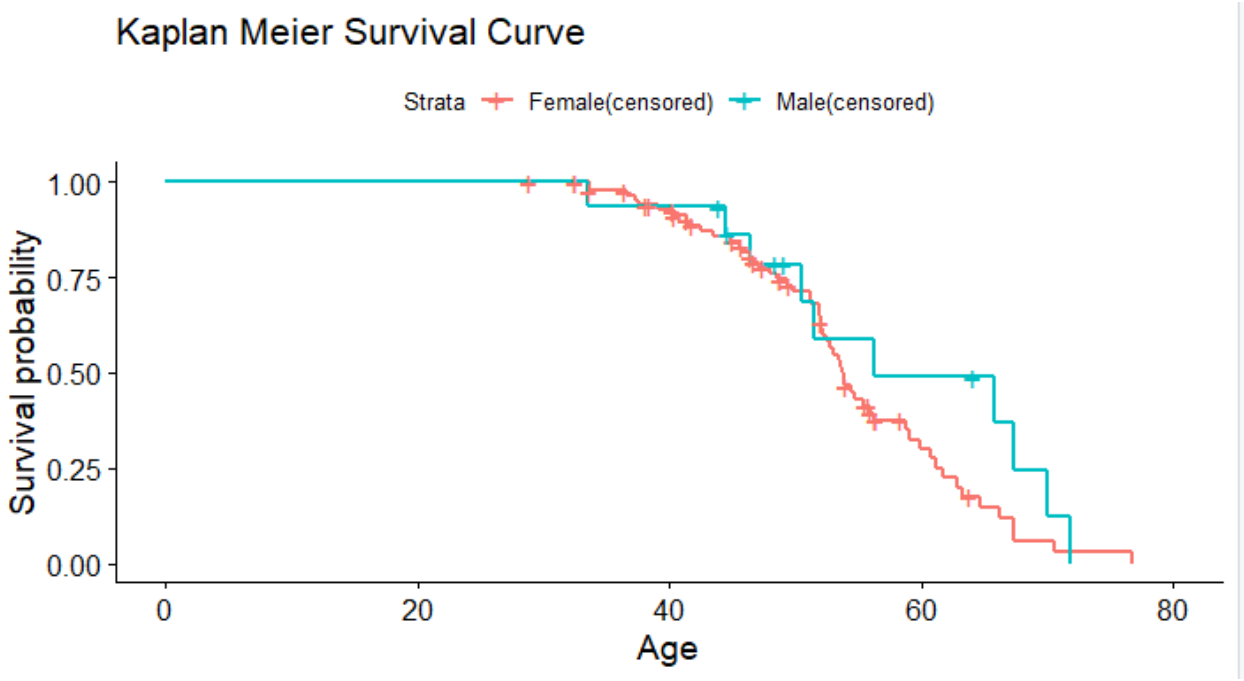

FIGURE 2. The Kaplan-Meier survival curves stratified by gender.

In the data set, there are 36 male and 276 female patients where the "ticks" shown in the legend above represents censored observations. In Figure 2, we plotted the Kaplan-Meier survival curves stratified by gender. The earliest death occurs at about age 34 for males and 31 for females. The 50% survival in males is around age 55 whereas in females it is around age 75. Since the survival curve for female patients lies consistently above that for male patients, females survive longer; however, judging by the appearance, there seems to be no significant difference in survival curves. To verify there are insignificant differences in survival curves for male vs. female, we carry out the log-rank test. From the log rank test output shown in Table 3A of appendix B, the test statistic is $\chi^2 = 1.2$ with degrees of freedom $n - 1 = 1$ where $n = 2$ is the number of curves being compared. The P-value is 0.3 and since it was greater than $\alpha = 0.05$, we fail to reject the null hypothesis, concluding that there is no significant difference in gender survival curves.

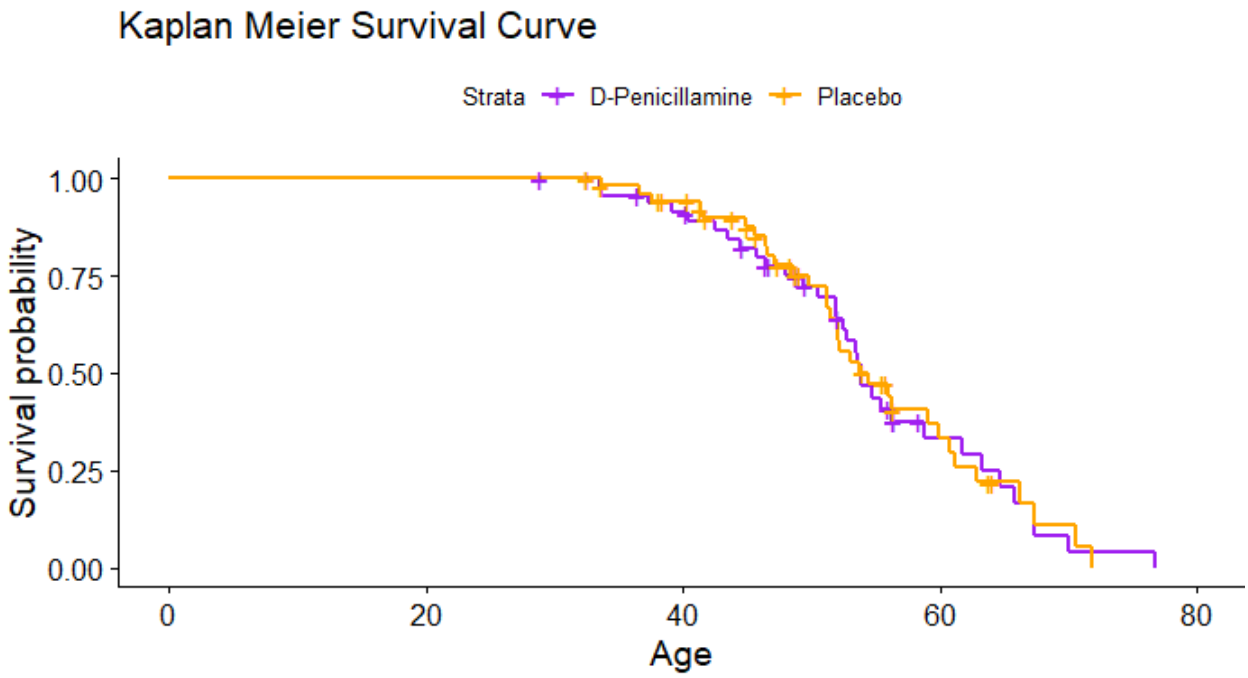

FIGURE 3. The Kaplan-Meier survival curves for D-Penicillamine vs. Placebo Patients.

In Figure 3, we plotted the survival curves for patients taking D-Penicillamine (the treatment group) versus those taking a placebo (the control group). From the appearance of the curves, we can see that the curve D-Penicillamine group had a slightly longer survival time; however, compared to the placebo group, both curves show insignificant differences in survival length despite intersecting at several time points. To justify the claim of insignificant differences in survival probability, the log rank test was carried out. From the log rank test results shown in Table 3B of Appendix B, we achieve a test statistic is $\chi^2 = 0.1$ with a corresponding P-value of 0.08. Therefore, we fail to reject the null hypothesis at the $\alpha = 0.05$ level of significance, and conclude that there is no significant difference between survival curves for the D-Penicillamine and the control group patients.

2.2.3. The Nelson-Aalen Estimator

The Nelson-Aalen estimator is an alternative estimator to the survival function in case of censored data. The overall survival curve depicted in Figure 4 below. The numerical results of the Nelson-Aalen estimator are summarized in Table 2 of Appendix B.

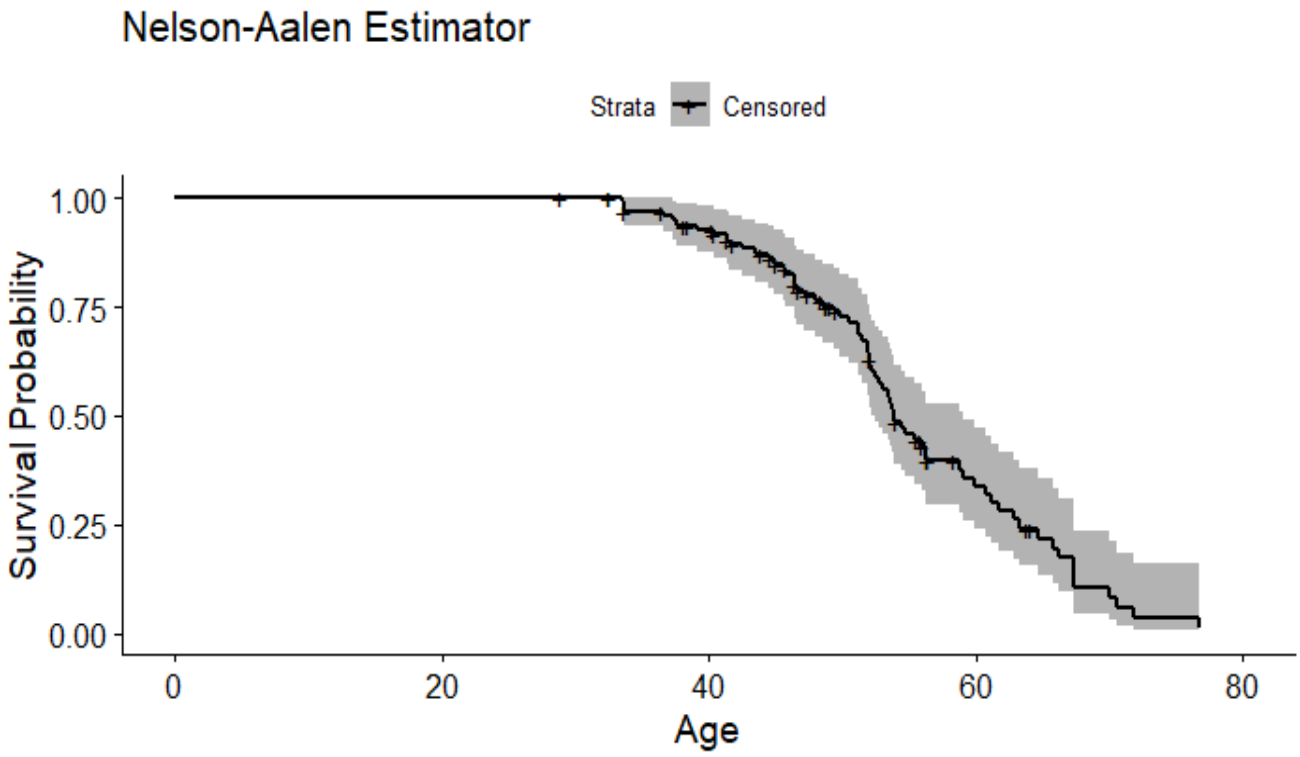

FIGURE 4. The Nelson-Aalen survival curve.

From the table and the graph, there is a 100% survival up to age 30.9, after which patients start dying. The results (see Appendix B, Table 2) shows that more than 50% of the cohort died prior to reaching age 61. At the age of 67 is where 30% of the cohort are still alive but have a high risk of dying within a few months. The curve exhibits slow but steady downward trend which gets steeper as the patients' age progresses. At the age range of 50-80 is where the graph is the steepest, as many patients die of old age. Comparing the results Nelson-Aalen estimator

with that of the Kaplan-Meier one, both estimators produced similar, but slightly deviating, survival probability estimates for patients over time.

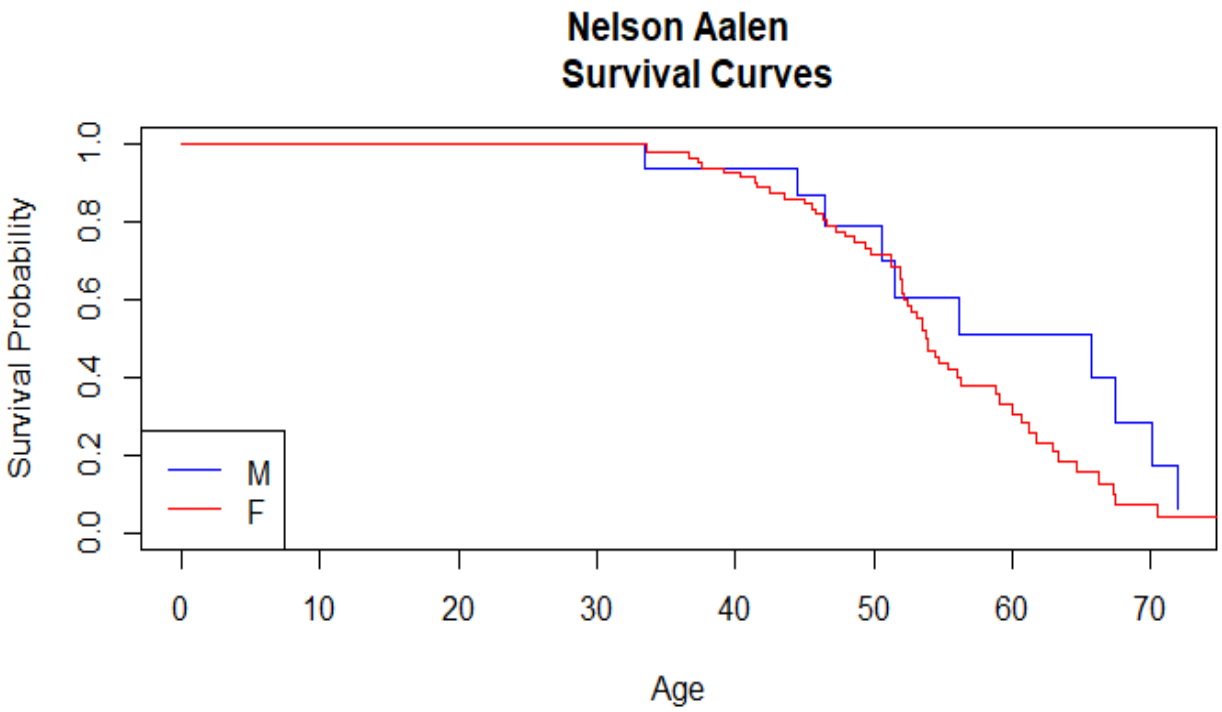

FIGURE 5. The Nelson-Aalen survival curves stratified by gender.

Figure 5 shows the Nelson-Aalen survival curves stratified by gender. From the graph, both curves for males and females exhibit a downward trend with several instances where the survival curves intersected. Likewise, the overall size of the survival curve for males was slightly larger than that of females, indicating that males had a higher survival length.

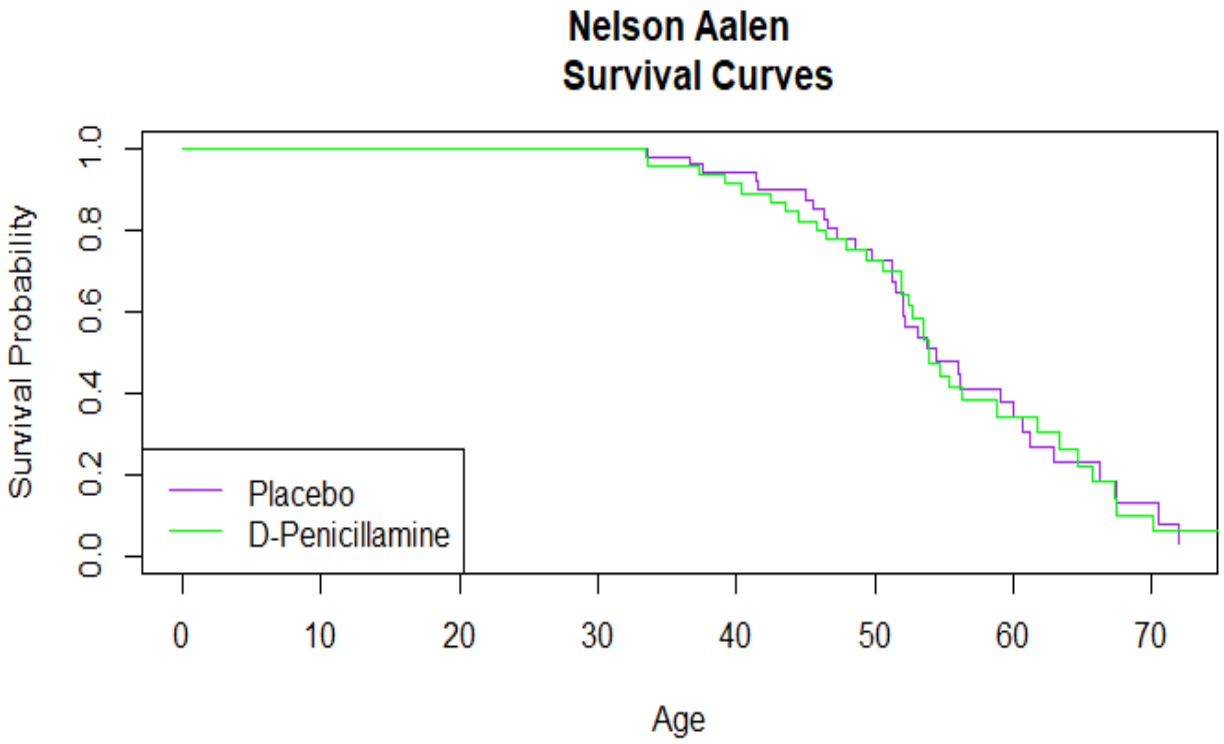

FIGURE 6. The Nelson-Aalen estimated survival curves for D-Penicillamine vs. Placebo patients.

The graph shown in Figure 6 depicts the Nelson-Aalen survival curves for patients given D-Penicillamine vs. those who were given a placebo. From the graph, we see that the survival curves exhibit a similar behavior, not deviating from each other by much. This is indicative of no significant difference in patients' hazard of dying between the two groups.

2.2.4. Weibull Estimator of Survival Function

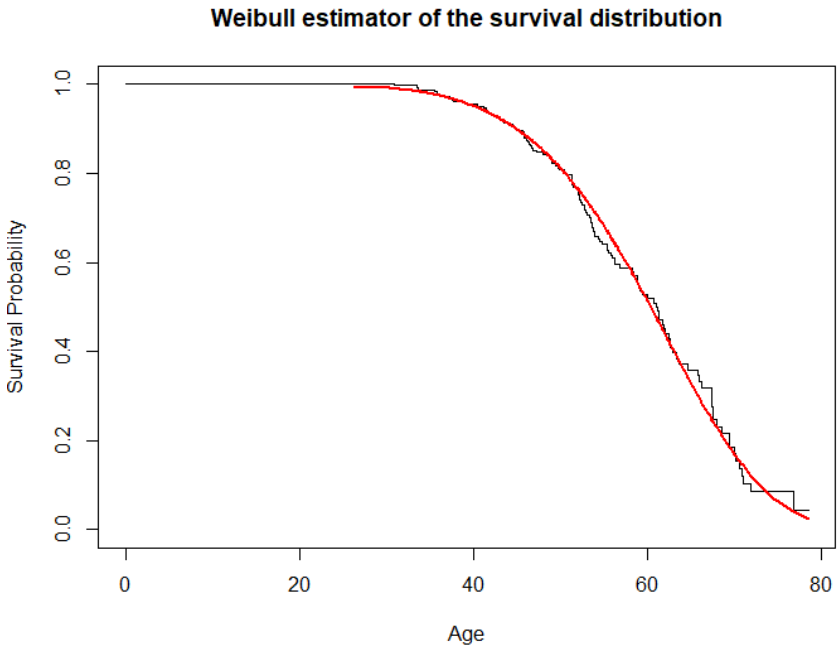

FIGURE 7. Weibull estimator of survival function.

Figure 7 depicts the survival function for the Weibull distribution. From the graph, the Weibull curve remains almost constant up until the age of 30. After age 30, the curve starts decaying slowly, then rapidly near the end since many patients die of old age. Therefore, we can conclude the Weibull parametric model is an appropriate fit for the data. The formula for the survival curves, as well as the estimates for the parameters are shown in the next section.

2.2.5. Weibull Regression Model

From the output of the Weibull regression model (see Appendix B, Table 2), the significant predictors at the $\alpha = 0.05$ level of significance were: edema, age, serum bilirubin concentration, albumin concentration, copper, AST concentration, prothrombin time, and histologic stage of disease.

From there, we fit a reduced Weibull model by re-running the model using only the significant predictors, and obtain the output given in Table 3 in Appendix B. The scale parameter of the distribution is estimated as $\hat{\sigma} = 0.614$ where the shape parameter is equal to $\hat{\alpha} = \frac{1}{\hat{\sigma}} = 1.63$. The estimated parameter λ can be written as:

$$\hat{\lambda} = \exp\left(\frac{-\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k\right)}{\hat{\sigma}}\right) = \exp\left(-\frac{1}{0.614}(11.095 - 0.562 \cdot edema\right.$$

$$-0.020 \cdot age - 0.050 \cdot bili + 0.461 \cdot albumin - 0.002 \cdot copper - 0.003 \cdot ast - 0.176$$

$$\left. \cdot protime - 0.249 \cdot stage)\right).$$

The fitted survival function is

$$\hat{S}(t) = \exp\left(-\hat{\lambda} t^{\hat{\alpha}}\right) = \exp\left(-\exp\left(-\frac{11.095 + \cdots - 0.249 \cdot stage}{0.614}\right) t^{1.63}\right), t \geq 0.$$

To justify whether the Weibull model is a good fit for the data, we conduct the deviance test. The test statistic is equal to

$$deviance = -2(\ln L(\beta_0, \sigma) - \ln L(\beta_0, \ldots, \beta_k, \sigma))$$

$$= -2\left(-1188.753 - (-967.3627)\right) = 442.7803.$$

The number of degrees of freedom is the same as the number of predictors in the fitted model, that is, $df = 8$. Under the null hypothesis, the deviance follows a chi-squared distribution with 8 degree of freedom, where the P-value is computed as

$$\mathbb{P}(\chi^2(8) > 442.7803) \ll 0.05.$$

Since the P-value was significantly less than $\alpha = 0.05$, we accept the alternative hypothesis, and thus conclude that the Weibull model fits the data well.

Next, we give the interpretation of the significant estimated regression coefficients. Recall that the interpretation is done in terms of the estimated hazard function with the estimated regression coefficients $\hat{\beta}^* = -\frac{\hat{\beta}}{\hat{\sigma}} = -\frac{\hat{\beta}}{0.614}$.

1. For a one-unit increase in each stage of edema, there is a $(\exp(0.562/0.614) - 1) \cdot 100\% = 149.75\%$ increase in the estimated hazard.
2. For a one-year increase in age, there is a $(\exp(0.020/0.614) - 1) \cdot 100\% = 3.31\%$ increase in the estimated hazard.
3. For a one-milligram increase in the bilirubin concentration per deciliter of blood, there is a $(\exp(0.05/0.614) - 1) \cdot 100\% = 8.48\%$ increase in the estimated hazard.
4. For a one-milligram increase in albumin concentration per deciliter of blood, there is a $(\exp(-0.461/0.614) - 1) \cdot 100\% = -52.80\%$ change in the estimated hazard, that is, a 52.80% decrease.
5. For each microgram increase in copper per day, there is a $(\exp(0.002/0.614) - 1) \cdot 100\% = 0.32\%$ increase in the estimated hazard.
6. For each unit increase in AST concentration per milliliter, there is a $(\exp(0.003/0.614) - 1) \cdot 100\% = 0.49\%$ increase in the estimated hazard.
7. For each second increase of the Prothrombin time, there is a $(\exp(0.176/0.614) - 1) \cdot 100\% = 33.20\%$ increase in the estimated hazard.
8. For each one-unit increase in the historic stage of the disease, there is a $(\exp(0.249/0.614) - 1) \cdot 100\% = 50.01\%$ increase in the estimated hazard.

Next, we use the fitted Weibull model to predict the probability of survival of a 56-year-old patient with no edema present, with a bilirubin concentration of 1.1 mg/dl, albumin

concentration of 4.4 mg/dl, copper concentration of 54 $\mu g/day$, AST concentration of 113.5 U/liter, a Prothrombin time of 10.4 seconds, and who is in the third stage. Therefore, we calculate

$$\lambda^0 = \exp(-\frac{1}{0.614}(11.095 - 0.562(0) - 0.020(56) - 0.050(1.1) + 0.461(4.4)$$

$$-0.002(54) - 0.003(113.5) - 0.176(10.4) - 0.249(3))),$$

and the predicted survival probability of this patient at 1925 days:

$$\hat{S}(1925) = \exp(-(\lambda^0)(1925)^{1.63}) = 0.8955692.$$

Thus, at 1925days, we can see that the survival probability of this patient is around 0.90.

2.2.6. Cox Proportional Hazards Model

Table 4 of Appendix B contains the output of fitting the Cox proportional hazards model. At the $\alpha = 0.05$ level of significance, the significant predictors were edema, age, serum bilirubin concentration, albumin concentration, copper, AST concentration, prothrombin time, and histologic stage of disease.

From there, we fitted a reduced model by re-running the model specifying only the significant predictors and achieved a reduced Cox output which can be found in Table 5 of Appendix B.

Before estimating the survival function of the Cox model, we must first estimate the baseline function $\hat{S}_0(t)$ through a step function $\bar{S}(t)$ equivalent to

$$\bar{S}(t) = \left[\hat{S}_0(t)\right]^{\exp(\hat{\beta}_1\bar{x}_1+\cdots+\hat{\beta}_k\bar{x}_k)}$$

where $\bar{x}_1, \ldots, \bar{x}_k$ is our sample means of the significant predictors. The purpose of this step function is to model the survival function of an "average" individual by which the values of all predictors are equal to the sample means. Typically, the estimates of both the step function and

the sample means of the predictors can be achieved using a specific R command (see Tables 6, Appendix B).

From the baseline survival, we can now estimate the fitted Cox survival function $\hat{S}_T(t)$ as

$$\hat{S}_T(t) = [\bar{S}(t)]^{\exp(\hat{\beta}_1(x_1-\bar{x}_1)+\cdots+\hat{\beta}_k(x_k-\bar{x}_k))} = [\bar{S}(t)]^{\hat{r}}$$

where $\hat{r} = \exp(0.832(edema - 0.111) + 0.033(age - 50) + 0.085(bili - 3.26)$

$$-0.787(albumin - 3.52) + 0.003(copper - 97.6) + 0.005(ast - 123.0)$$

$$+ 0.268(protime - 10.7) + 0.405(stage - 3.03)).$$

From the reduced Cox model, the fitted significant regression coefficients yield the following interpretation.

1. For a one-unit increase in each stage of edema, there is a $(\exp(0.832) - 1) \cdot 100\% = 129.8\%$ increase in the estimated hazard.
2. For a one-year increase in age, there is a $(\exp(0.033) - 1) \cdot 100\% = 3.36\%$ increase in the estimated hazard.
3. For a one-milligram increase in the bilirubin concentration, there is a $(\exp(0.085) - 1) \cdot 100\% = 8.8\%$ increase in the estimated hazard.
4. For a one-milligram increase in albumin concentration, there is a $(\exp(-0.788) - 1) \cdot 100\% = -54.53\%$ change in the estimated hazard, that is a decrease of 54.53%.
5. For each microgram increase in copper, there is a $(\exp(0.003) - 1) \cdot 100\% = 0.30\%$ increase in the estimated hazard.
6. For each unit increase in AST concentration, there is a $(\exp(0.005) - 1) \cdot 100\% = 0.501\%$ increase in the estimated hazard.

7. For each second increase of the Prothrombin time, there is a $(\exp(0.268) - 1) \cdot 100\% = 30.7\%$ increase in the estimated hazard.
8. For each one-unit increase in the historic stage of the disease, there is a $(\exp(0.405) - 1) \cdot 100\% = 50\%$ increase in the estimated hazard.

Using the same information from the example involving the Weibull model (see Section 2.2.5), we will now use our fitted Cox proportional hazards model to predict the survival of a patient at about time $t = 1925$ days. We compute

$$\begin{aligned}\hat{r}^0 = \exp\{&0.832(0 - 0.111) + 0.033(56 - 50) + 0.085(1.1 - 3.26) - 0.787(4.4 - 3.52) \\ &+ 0.003(54 - 97.6) + 0.005(113.5 - 123.0) + 0.268(10.4 - 10.7) \\ &+ 0.405(3 - 3.03)\} = \exp(-1.041062) = 0.35308,\end{aligned}$$

and

$$\hat{S}(1925) = [\bar{S}(1925)^0]^{\hat{r}^0} = [0.760]^{\hat{r}^0} = [0.760]^{(0.35308)} = 0.9076.$$

Thus, at 1925 days, the predicted probability for this particular patient hovers at around 91%, very close to our estimate achieved from the Weibull model.

CHAPTER 3

LONGITUDINAL DATA ANALYSIS

3.1. Regressions for Longitudinal Data

Longitudinal data are defined as measurements collected on the same individuals at several time points. In the medical field, most of the collected data are collected longitudinally from medical records or during clinical trials. The specificity of longitudinal data is that repeated measurements within each individual are expected to be correlated; therefore, a regression model should reflect potential correlations within each individual and no correlation between different individuals at any time points, same or not.

Below we present the theoretical framework for the random slope and intercept models for the response variables with normal, gamma, binary logistic, and Poisson distributions.

3.1.1 Normally Distributed Response

To model this potential correlation within each individual, we can fit a mixed-effects model (or longitudinal model, or random slope and intercept model). This model is defined as $y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j + \varepsilon_{ij}$ where the measurements on the ith individual (subject), $i = 1, \ldots, n$, are collected at time points $t_j, j = 1, \ldots, p,$ and $x_1, x_2, \ldots, x_k$ denote the predictors (which may vary with time). The term u_1 is the random intercept and u_2 is the random slope. Both $u_i \sim N\left(0, \sigma_{u_i}^2\right), i = 1,2,$ and the random error $\varepsilon \sim N(0, \sigma^2)$. It is also assumed that $Cov(u_{1i}, u_{2i}) = \sigma_{u_1 u_2}$ and $Cov(u_{1i}, u_{2i'}) = 0$ for $i \neq i'$. The slope and intercept are independent of the errors. The observed response y_{ij} on the ith subject at the jth time point is a normally distributed random variable with mean $\mu = E\left(y_{ij}\right) = \beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j$ and variance $Var\left(y_{ij}\right) = \sigma_{u_1}^2 + 2\sigma_{u_1 u_2} t_j + \sigma_{u_2}^2 t_j^2 + \sigma^2$. The

responses for different individuals at any time point (the same or not) are uncorrelated, that is, $Cov(y_{ij}, y_{i'j'}) = 0, i \neq i'$. Observations for the same individual over time are correlated, $Cov(y_{ij}, y_{ij'}) = \sigma_{u1}^2 + \sigma_{u_1u_2}(t_j + t_{j'}) + \sigma_{u_2}^2 t_j t_{j'}$, where $j \neq j'$.

The fitted model is written as $\hat{E}(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$, with the estimated parameters $\hat{\sigma}_{u_1}^2$, $\hat{\sigma}_{u_2}^2, \hat{\sigma}_{u_1u_2}$, and $\hat{\sigma}^2$. The estimated regression coefficients yield the following interpretation. If $x_1$is numeric, then $\hat{\beta}_1$ represents the change in the estimated mean response for one-unit increase in x_1, provided all the other predictors are unchanged. Indeed,

$$\hat{E}(y|x_1 + 1) - \hat{E}(y|x_1) = \hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$$
$$-(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t) = \hat{\beta}_1.$$

If x_1 is a 0-1 predictor variable, then $\hat{\beta}_1$ is interpreted as a difference in the estimated mean response for $x_1 = 1$ and $x_1 = 0$, provided the other predictors stay fixed. This is justified because $\hat{E}(y|x_1 = 1) - \hat{E}(y|x_1 = 0) = \hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t - (\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t) = \hat{\beta}_1$.

From the fitted model, the predicted response y^0 for a set of predictors $x_1^0, \ldots, x_k^0, t^0$ is equal to $y^0 = \hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0$.

3.1.2 Model Goodness-of-Fit Check

To test how well the fitted model fits the data, a goodness-of-fit deviance test is employed. In this test the null hypothesis is that the null model fits the data better where the null model contains only fixed-effect predictors and no random-effect ones. The alternative hypothesis states that the fitted model (with mixed-effect terms) has a better fit. The test statistic is called deviance and is calculated as

$$deviance = -2(\ln L(null\ model) - \ln L(fitted\ model)).$$

Under H_0, the test statistic follows a χ^2- distribution with the number of degrees of freedom equal to the difference in the number of parameters used in both models. Namely, the full model contains $k + 2$ fixed-effect regression coefficients plus 4 sigmas, whereas the null model contains only the $k + 2$ betas plus one sigma. Therefore, the number of degrees of freedom in this case is 3. The fitted model has a good fit if the P-value is smaller than 0.05, and the alternative is accepted.

3.1.3 Generalized Estimating Equations Model

An alternative method to model longitudinal data is with Generalized Estimating Equations (GEE) models. In GEE models there are no random-effect terms and no random error. The distribution of the response variable is assumed known, the mean is modeled related to the linear regression term with fixed-effects only, and the variance-covariance structure is pre-specified. The theory is as follows.

Let $x_{1ij}, \dots, x_{kij}$ denote the longitudinal observations of predictors for each individual $i, i = 1, \dots, n,$ at time $t_j, j = 1, \dots, p,$ and let y_{ij} denote the response for the ith individual at the jth time point. The mean and variance of y_{ij} are equal to $\mu_{ij} = E(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j$ and $Var(y_{ij}) = V(\mu_{ij})$ where $V(.)$ is the variance function. Next, the covariance structure of correlated responses for a given individual $i, i = 1, \dots n,$ is modeled by a $p \times p$ matrix denoted by

$$\mathbf{A}_i = \begin{pmatrix} V(\mu_{i1}) & 0 & \cdots & 0 \\ 0 & V(\mu_{i2}) & \cdots & 0 \\ 0 & 0 & \cdots & \cdots \\ 0 & 0 & \cdots & V(\mu_{ip}) \end{pmatrix}.$$

Observations between individuals are independent. Next, we let $\mathbf{R}_i(\boldsymbol{\alpha})$ represent the working correlation matrix of the repeated responses for the ith subject where $\boldsymbol{\alpha}$ represents a

vector of unknown parameters, equal for all subjects within the study. Then, the covariance matrix for the vector of responses $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})'$ is equal to

$$\mathbf{V_i}(\boldsymbol{\alpha}) = \mathbf{A_i}^{1/2} \cdot \mathbf{R_i}(\boldsymbol{\alpha}) \cdot \mathbf{A_i}^{1/2}.$$

The regression coefficients $\beta_0, \dots, \beta_{k+1}$ and the vector of parameters $\boldsymbol{\alpha}$ are estimated numerically from the data by solving the generalized estimating equations:

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}\right)_{(k+2)\times p} [\mathbf{V}_i(\widehat{\boldsymbol{\alpha}})]^{-1}_{p\times p} (\boldsymbol{y_i} - \boldsymbol{\mu}_i)_{p\times 1} = \mathbf{0}_{(k+2)\times 1}$$

where $\boldsymbol{\mu_i} = (\mu_{i1}, \dots, \mu_{ip})'$ is the vector of mean responses, and $\widehat{\boldsymbol{\alpha}}$ is the method-of-moments estimator of the vector of parameters.

Remark: Five commonly used structures for the working correlation matrix $\mathbf{R_i}(\boldsymbol{\alpha})$ for a GEE are: Unstructured, Toeplitz, autoregressive, compound symmetric (exchangeable), and independent.

- Unstructured matrix with all different off-diagonal entries with all off-diagonal entries, having a total of $p(p-1)/2$ unknown parameters

$$\mathbf{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_{12} & \alpha_{13} & \cdots & \alpha_{1p} \\ \alpha_{12} & 1 & \alpha_{23} & \cdots & \alpha_{2p} \\ \alpha_{13} & \alpha_{23} & 1 & \cdots & \alpha_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{1p} & \alpha_{2p} & \alpha_{3p} & \cdots & 1 \end{pmatrix}.$$

- Toeplitz matrix with identical entries on each descending diagonal, having a total of $p-1$ unknown parameters

$$\mathbf{R_i}(\boldsymbol{\alpha}) = \begin{pmatrix} 1 & \alpha_1 & \alpha_2 & \cdots & \alpha_{p-1} \\ \alpha_1 & 1 & \alpha_1 & \cdots & \alpha_{p-2} \\ \alpha_2 & \alpha_1 & 1 & \cdots & \alpha_{p-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{p-1} & \alpha_{p-2} & \alpha_{p-3} & \cdots & 1 \end{pmatrix}.$$

- Autoregressive matrix with $\alpha^{|i-j|}$ in the $ijth$ position, yielding a total of one unknown parameter

$$\mathbf{R_i(\alpha)} = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{p-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{p-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{p-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha^{p-1} & \alpha^{p-2} & \alpha^{p-3} & \cdots & 1 \end{pmatrix}.$$

- Compound symmetric or exchangeable matrix with all identical off-diagonal elements, yielding a total of one unknown parameter

$$\mathbf{R_i(\alpha)} = \begin{pmatrix} 1 & \alpha & \alpha & \cdots & \alpha \\ \alpha & 1 & \alpha & \cdots & \alpha \\ \alpha & \alpha & 1 & \cdots & \alpha \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha & \alpha & \alpha & \cdots & 1 \end{pmatrix}.$$

- Independent identity matrix with no unknown parameters

$$\mathbf{R_i(\alpha)} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

To determine which model is the best fit for the data, the quasi-likelihood under the independence (QIC) model based on the function

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{p} \int_{y_{ij}}^{\mu_{ij}} \frac{y_{ij} - u}{V(u)} du$$

is computed for each type of working correlation matrix of the repeated responses. The QIC is a goodness-of-fit measure that is used to select the best-fitted working correlation structure. After computing the QIC for all the types of the correlation matrix structures, the model with the smallest QIC is used as the model of best fit. In the case where two or more models are tied with having the lowest QIC (i.e., share the same Q value), then either of those models has the best fit.

The fitted GEE model is written as $\hat{E}(y) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t$, with the estimated working correlation matrix $\widehat{\mathbf{R}}(\hat{\alpha})$. Estimated regression coefficients are interpreted similar to how it is done in the mixed-effects model (see Section 3.1.1).

3.1.4. Application of Normal Response

TABLE 2. Description of Variables in Blood Pressure Dataset

Name	Description	Type	Values
BLOOD_PRESSURE	This is our predictor variable, the systolic blood pressure (in mm/hg).	Numeric	Varies based on input
GENDER	Male or Female	Categorical Binary	M or F
ACTIVITY	The daily activity level of patients. Measured on a scale of 1-10 where 1= Not active and 10=Very active	Categorical Numeric	1=Not active 6=Moderately active 10=Very active
SODIUM	Level of dietary sodium consumed by patient daily	Categorical Numeric	1=Low Sodium 2=Moderate Sodium 3=High Sodium
HISTORY	Family history of high blood pressure	Binary Numeric	0= No family history of high blood pressure 1= Patient's family has a history of high blood pressure

TABLE 2. Continued

Name	Description	Type	Values
CATEGORY	Patients with a Systolic mm/Hg of less than 120 are classified as having normal blood pressure. Those with a systolic blood pressure of 120-129 mm/Hg are classified as having elevated blood pressure. Those with a systolic blood pressure of 130-139 mm/Hg are classified as having stage 1 blood pressure (Hypertension) Those with a systolic blood pressure of greater than 140 mm/Hg are classified as having stage 2 blood pressure (Hypertension) Those with a systolic blood pressure of greater than 180 mm/Hg are classified as having hypertensive crisis. Therefore, emergency treatment is required	Categorical Numeric	1= Normal blood pressure 2= Elevate blood pressure 3= Hypertension Stage 1 4= Hypertension Stage 2 5= Hypertensive crisis
WEEK	At the end of each week, patients visited the clinic and had their blood pressures recorded	Numeric	Ranges from 1 week to 6 weeks

The Blood Pressure dataset (abbreviated as “bp”) is a simulated dataset consisting of $n = 45$ patients of varying blood pressures. The purpose of this dataset was to test the effectiveness of a new pill on lowering patients’ systolic blood pressure levels. The scenario for this example is as follows. Prior to entering the clinical study, the researchers gave each patient a questionnaire that asks about the patients’ physical activity level, sodium intake level, and whether their families had a history of high blood pressure. The results from these questions can

provide the researchers more clues about the effectiveness of the clinical trial. After that, the patients' systolic blood pressure was recorded once a week during the six-week study.

3.1.5. Application (Normal response)

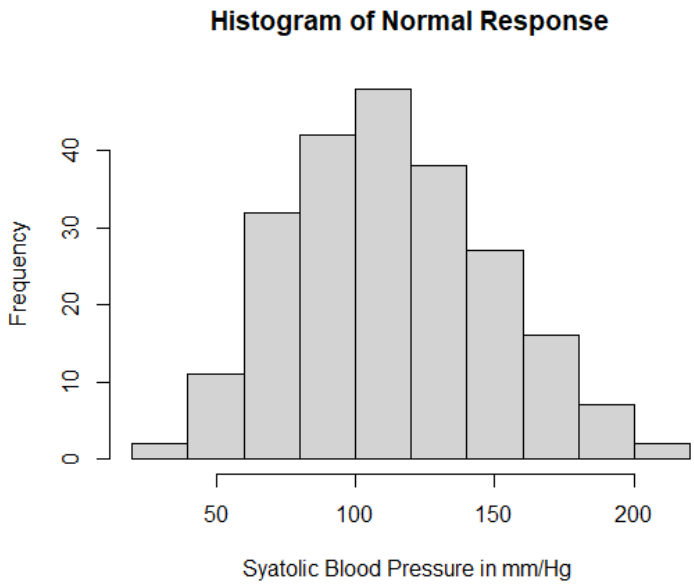

FIGURE 8. Histogram of normal response.

Figure 8 depicts the density histogram of the patients' systolic blood pressure. From the appearance of graph, the response is symmetric about the mean, indicating that the data nearing the mean occurs more frequent than data far from the mean. Thus, the response follows a normal distribution very nicely. To justify our claim that the response is symmetric about the mean, the Shapiro-Wilk normality test shown (see Table 1A in Appendix D) is employed. From the results of the test, we observed that because $p = 0.1338 > \alpha$ at the $\alpha = 0.05$ level of significance, we conclude that the response is indeed normally distributed.

Table 2A of Appendix D depicts the random slope and intercept output for the normal response. From the output, it appears that the patients' activity level, low sodium diet, patients whose families had a history of high blood pressure, the patients' blood pressure category, and

the week the patients visited the clinic were deemed very significant predictors. Therefore, our fitted random slope and intercept model can be written as

$$\hat{E}(Blood\ Pressure) = 111 - 5.8Male - 1.4Activity - 20.4Sodium(Level1)$$
$$-3.6Sodium(Level2) - 8.5HistoryOfHighBloodPressure + 18.8Category - 9.9Time.$$

To determine whether a null model has a better fit against the fitted model, we employ the deviance test. During the test, we specify the null model as a standard generalized linear model, and the fitted model as a random slope and intercept model. The results obtained by the deviance test (see Appendix D, Table 1B), indicate that since the P-value was exponentially small, we accept the alternative hypothesis and therefore conclude that, compared to the null model, the fitted model fits the data better.

Our interpretation of the significant predictors is as follows. The patients' activity was measured on a scale of 1-10, with 10 being very active and 1 being very sedentary. For each unit increase in the activity scale, there was about a 1.42 mm/hg decrease in blood pressure, on average. Secondly, a low sodium diet played a critical role in lowering blood pressure concentration. Those who ate a low sodium diet decreased their estimated average blood pressure by about 20.43 mm/hg. This result is consistent with our common belief that a low sodium is indeed effective in lowering blood pressure levels. Furthermore, the estimated mean blood pressure level for patients whose families had a history of high blood pressure was 8.53mm/hg less than those whose families never had a history of high blood pressure. A plausible explanation for this occurrence was due to the effectiveness of the therapy session. With respect to each category in blood pressure, we observed the average blood pressure for each patient increase by about 18.83 mm/hg for each increase in blood pressure stage. Lastly, there is an estimated average of 9.92 mm/hg reduction in blood pressure every week.

Finally, we put our fitted model to the test by calculating the average systolic blood pressure by the end of week 4 of a female patient who is moderately active (i.e., Activity=6), eats a low sodium diet (i.e., sodium level=1), has no family history of high blood pressure, and is categorized as having Elevated Blood Pressure (i.e., Category=2). Using the fitted random slope and intercept model, the predicted blood pressure for this patient is

$$Blood\ Pressure^0 = 111 - 1.4(6) - 20.4(1) - 3.6(0) - 8.5(0) + 18.8(2) - 5.8(0) - 9.9(4)$$
$$= 80.2\ \frac{mm}{hg}.$$

Next, we fit a generalized estimating equations (GEE) model for the normal response using the unstructured, autoregressive, exchangeable, and independent working correlation matrices. From the outputs shown in Tables 2B, 2C, 2D, and 2E in Appendix D, since the exchangeable GEE model has the lowest QIC out of the four, we conclude that the model with the exchangeable correlation matrix has the best fit. Therefore, we use this model for interpretation and prediction.

Thus, the fitted generalized estimating equation model with the exchangeable working correlation matrix is

$\hat{E}(Blood\ Pressure) = 120.18 - 6.38Male - 1.09Activity - 13.21Sodium(level1)$
$-2.54Sodium(level2) - 10.64HistoryOfHighBloodPressure + 14.61Category -$
$9.92Week,$ and

$$\widehat{\boldsymbol{R}}(\widehat{\boldsymbol{\alpha}} = 0.369) = \begin{pmatrix} 1 & \hat{\alpha} & \hat{\alpha} & \cdots & \hat{\alpha} \\ \hat{\alpha} & 1 & \hat{\alpha} & \cdots & \hat{\alpha} \\ \hat{\alpha} & \hat{\alpha} & 1 & \cdots & \hat{\alpha} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{\alpha} & \hat{\alpha} & \hat{\alpha} & \cdots & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0.369 & 0.369 & 0.369 \\ 0.369 & 1 & 0.369 & 0.369 \\ 0.369 & 0.369 & 1 & 0.369 \\ 0.369 & 0.369 & 0.369 & 1 \end{pmatrix}.$$

At the $\alpha = 0.05$ level of significance, patients whose families had a history of high blood pressure, the patients' blood pressure category, and the week the patients visited the clinic

were deemed very significant predictors. Therefore, our interpretation of the significant estimated regression coefficients is as follows. The estimated mean blood pressure level for patients whose families had a history of high blood pressure was 10.64mm/hg less than those whose families never had a history of high blood pressure. Further, we observed the average blood pressure for each patient increase by about 14.61 mm/hg for each increase in blood pressure stage. Lastly, there is an estimated average of 9.92 mm/hg reduction in blood pressure every week.

Putting our fitted GEE model to the test, using the same example, our predicted blood pressure for this patient by the end of the fourth week is

$$Blood\ Pressure^0 = 120.18 - 6.38(0) - 1.09(6) - 13.21(1) - 2.64(0)$$

$$-10.64(0) + 14.61(2) - 9.92(4) = 89.97\ \frac{mm}{hg}.$$

Thus, from our prediction, we can conclude that from the clinical trial performed on this specific patient, this person is predicted to have normal blood pressure by the end of the fourth week.

3.2. Regressions for Gamma Response

3.2.1. Theoretical Framework

In a longitudinal setting, gamma regression is appropriate to use if the response variable y_{ij} in a dataset follows a right-skewed distribution (i.e., has a long right tail). In that case, the response is written as $y_{ij} \sim \Gamma(\alpha, \beta)$ with the probability density function equal to $f(y_{ij}) = \frac{y_{ij}^{\alpha-1}}{\Gamma(\alpha)\beta^{\alpha}} \exp\left(-\frac{y_{ij}}{\beta}\right)\ \alpha, \beta, y > 0,$ where the expected value of y_{ij} is $\mu_{ij} = E(y_{ij}) = \alpha\beta$, with α and β being the shape and scale parameters. To model the expected value of the response as it relates to the linear combination of explanatory variables, a log-link function is used. For fixed

values of the random intercept u_{1i} and slope u_{2i}, we can write $\ln(\mu_{ij}) = \ln E(y_{ij}) = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j$. The random intercepts u_{1i}'s are independent $N(0, \sigma_{u_1}^2)$ random variables, the random slopes u_{2i}'s are independent $N(0, \sigma_{u_2}^2)$ random variables, and the covariance between u_{1i} and u_{2i} is $\sigma_{u_1 u_2}$.

It is customary to write the fitted model as $\hat{E}(y) = \hat{\alpha}\hat{\beta} = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)$ where all beta parameters along with the random effects parameters $\sigma_{u_1}^2, \sigma_{u_1 u_2}$, and $\sigma_{u_2}^2$ are estimated from the data through the maximum-likelihood method.

Consequently, the parameters $\beta_0, \ldots, \beta_{k+1}$ and α are unknown and are then estimated by the method of maximum likelihood. The estimates of the regression coefficients $\beta_0, \ldots, \beta_{k+1}$ yield the following interpretation. From the model, if a predictor variable x_1 is numeric, then the change in the estimated mean response for a unit increase in x_1, provided all other predictors stay unchanged, is equal to

$$\frac{\hat{E}(y|x_1 + 1) - \hat{E}(y|x_1)}{\hat{E}(y|x_1)}$$

$$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right) - \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}$$

$$= \exp\left(\hat{\beta}_1\right) - 1.$$

Equivalently, $\left(\exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\%$ represents the percentage change in estimated mean response for a unit increase in x_1.

If x_1 is a 0-1 predictor variable, then the percent ratio of the estimated mean response $\hat{E}(y)$ for $x_1 = 1$ and $x_1 = 0$, provided the other predictors stay unchanged, is equal to

$$\frac{\hat{E}(y|x_1 = 1)}{\hat{E}(y|x_1 = 0)} \cdot 100\% = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)} \cdot 100\% = \exp\left(\hat{\beta}_1\right) \cdot 100\%.$$

From the fitted model, the predicted response y^0 for a set of predictors $x_1^0, \dots, x_k^0, t^0$ is equal to $y^0 = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0\right)$.

Further, the GEE model for the gamma-distributed response estimates the mean response through the function $\hat{E}(y) = \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)$, and the best fitted structure of the working correlation matrix is the one with the smallest QIC value. Like in the fitted random slope and intercept model, the beta parameters along with the parameters of the working correlation matrix are estimated through the maximum-likelihood estimation. Interpretation of estimated regression coefficients is done the same way as above.

3.2.2. Data Description Cancer (Gamma response)

TABLE 3. Description of variables in Cancer dataset

Name	Description	Type	Values
Oral_cond	This is our response variable, predicting the oral condition of patients	Numeric	Varies based on input
SEX	Male or Female	Binary Categorical	M or F
TRT	Patients were randomly assigned to a treatment and control group	Binary Categorical	0=Tx (Aloe Juice) 1=Cx (Placebo)
AGE	Age of patients	Numeric	Ranges from 26 to 86 years old
WEIGHT	Weight of patients (in lbs)	Numeric	Ranges from 120 lbs to 300 lbs

TABLE 3. Continued

Name	Description	Type	Values
STAGE	This is the initial cancer stage of the patients prior to entering the study. There are four stages in cancer: Stage 1= Typically a small cancer or tumor that has not grown deeply within the tissues; Stage 2= Larger cancers or tumors have grown more deeply into nearby tissue. The cancer may have spread to the lymph nodes, but not to other parts of the body; Stage 3= The tumor may have grown to a specific size and likely have spread to adjacent lymph nodes, organs, or tissues; Stage 4= Serious cancer condition where the cancer has spread from origin to distant parts of the body.	Multinomial Categorical	1= Stage 1 2= Stage 2 3= Stage 3 4= Stage 4
WEEKS	The Oral Condition of patients were measured every two weeks during the 6-week study. The Oral Condition is measured on a scale of 1-25 where an oral condition between the range of 15-25 represents having excellent oral health while an oral condition of 1 represents the worst possible oral health.	Numeric	Varies over time

The cancer dataset is a subset of data for a longitudinal study of the oral condition of cancer patients at the Mid-Michigan Medical Center. The primary goal for this dataset was to determine the effectiveness of the treatment (aloe juice) against a placebo in improving the oral condition of patients. The Oral Condition is measured on a scale of 1-25 where an oral condition

between the range of 15-25 represents having excellent oral health, the range of 10-15 represents normal oral health, and anything below 10 represents bad oral health. The oral condition of patients was measured every other week for a total of six weeks in this longitudinal study. During the study, the researchers found out that the oral health distribution was more right skewed, with the majority of patients falling within the 5-10 range, as opposed to the 20-25 range. Thus, a Gamma regression would appropriately model the response.

This sample dataset originally contained $n = 25$ patients with neck cancer, but extra patients were simulated for illustrative purposes.

3.2.3. Application

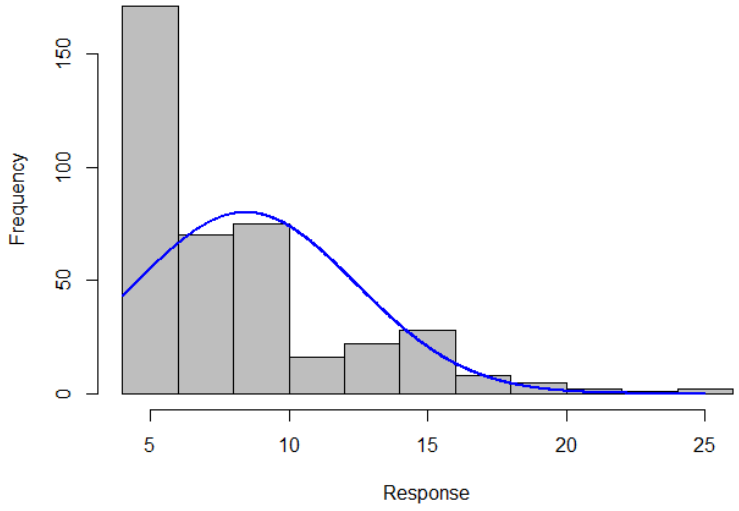

FIGURE 9. Histogram for oral condition.

From the histogram, the distribution of oral condition has a right-skewed distribution. To verify this claim, the Shapiro-Wilk normality test was run (see Appendix D, Table 3A). The P-value of the test is less than 0.05, leading to conclusion that the response is not normally distributed.

Next, we fit a random slope and intercept model that models the response, the oral condition of patients, and achieved an output which is shown in Table 4 of Appendix D. From the output, at the $\alpha = 0.05$ level of significance, we observed that the treatment (aloe juice) along with the number of weeks were significant predictors in determining the oral condition of the patients. The rest of the predictors were insignificant as they had P-values greater than $\alpha = 0.10$. The fitted random slope and intercept gamma model can be written as:

$$\hat{E}(Oral\ Condition) = \exp\ (1.643 - 0.1231Male + 0.3223Tx - 0.0015Age + 0.0013Weight + 0.045Stage + 0.083Weeks).$$

To check goodness-of-fit of the mode, we ran the deviance test. In this test, we specify our null model as a standard generalized linear model and our fitted model as that of a random slope and intercept model. From the deviance test results (see Appendix D, Table 3B), since the P-value was exponentially small, we concluded that the fitted model for the gamma response was better compared to the null.

Thus, our interpretation of the significant predictors is as follows. First, the estimated mean oral condition for patients in the treatment group is $\exp(0.3223) \cdot 100\% = 138.03\%$ of that for the patients in the control group. Next, for every week in the study, the estimated average oral condition of patients increases by $\exp(0.083) \cdot 100\% = 108.68\%$.

Using the fitted random slope and intercept model, our goal now is to predict the expected *initial* (week=0) oral condition of a 68-year-old female patient weighing 168 lbs., who is randomly assigned to the treatment group, and who is in the first cancer stage. The predicted value is:

$$Oral\ Condition^0 = \exp(1.643 - 0.1231(0) + 0.3223(1) - 0.001568 + 0.0013(168) + 0.045(1) + 0.083(0)) = \exp(2.22713) = 9.27.$$

Next, we fit a generalized estimating equations (GEE) model shown in Section 3.1.2 for the gamma response using the autoregressive, unstructured, exchangeable, and independent working correlation matrices. From the outputs shown in tables 4B, 4C, 4D, and 4E of Appendix D, since the unstructured GEE model had the lowest QIC out of the four, we conclude that the model with the unstructured working correlation matrix was the best-fitted model. It is written as

$$\hat{E}(Oral\ Condition) = \exp(1.516 - 0.126Male + 0.664Tx - 0.0013Age + 0.00068Weighin + 0.02Stage + 0.069Weeks),$$ and

$$\widehat{\mathbf{R}}_{\mathbf{i}}(\hat{\boldsymbol{\alpha}}) = \begin{pmatrix} 1 & \hat{\alpha}_{12} & \hat{\alpha}_{13} & \cdots & \hat{\alpha}_{1p} \\ \hat{\alpha}_{12} & 1 & \hat{\alpha}_{23} & \cdots & \hat{\alpha}_{2p} \\ \hat{\alpha}_{13} & \hat{\alpha}_{23} & 1 & \cdots & \hat{\alpha}_{3p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{\alpha}_{1p} & \hat{\alpha}_{2p} & \hat{\alpha}_{3p} & \cdots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 1.0973 & 0.3948 & -0.0734 \\ 1.0973 & 1 & 0.5157 & 0.2773 \\ 0.3948 & 0.5157 & 1 & 0.5379 \\ -0.0734 & 0.2773 & 0.5379 & 1 \end{pmatrix}.$$

The interpretation of the significant estimated regression coefficients is as follows. First, we observed that estimated cancer for males was $\exp(-0.126) \cdot 100\% = 88.2\%$ of that for females. Second, the estimated average oral condition for patients in the treatment group was $\exp(0.664) \cdot 100\% = 194\%$ of that in the control group. Lastly, for every two weeks, the estimated average oral condition of patients changes by about $\exp(0.069) \cdot 100\% = 107.14\%$.

For our fitted GEE model with the unstructured working correlation matrix, using the same example, our predicted oral condition is

$$Oral\ Condition^0 = \exp(1.516 - 0.126(0) + 0.664(1) - 0.0013(68) + 0.00068(168) + 0.021 + 0.069(0)) = \exp(2.23) = 9.30.$$

From the results of the predictions, we observe that at $Weeks = 0$ this patient is expected to have an oral condition of 9.27 from the fitted random slope and intercept model and 9.30 from the fitted GEE model. The results obtained through both fitted models were very close.

3.3. Regressions for Binary Response

3.3.1. Theoretical Framework

Binary logistic regression with random slope and intercept is used to model longitudinal data where the response variable assumes values 0 or 1. For subject $i, i = 1, \dots, n,$ at time t_j , $j = 1, \dots, p,$ let $\pi_{ij} = P(y_{ij} = 1)$. Note that π_{ij} is also the mean of y_{ij}. Indeed, $E(y_{ij}) = 1 \cdot \pi_{ij} + 0 \cdot (1 - \pi_{ij}) = \pi_{ij}$. The random slope and intercept model for a binary response can be written as:

$$\pi_{ij} = E(y_{ij}) = \frac{\exp(\beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j)}{1 + \exp(\beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j)}.$$

An alternative form of the model is:

$$\frac{\pi_{ij}(u)}{1 - \pi_{ij}(u)} = \exp\{\beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j\}.$$

Here $u_{1i}'s \sim N(0, \sigma_{u_1}^2)$ are the random intercepts and $u_{2i}'s \sim N(0, \sigma_{u_2}^2)$ are the random slopes for $i = 1, \dots, n, j = 1, \dots, p$. The covariance between u_{1i} and u_{2i} is $\sigma_{u_1 u_2}$. The parameters of this model $\beta_0, \dots, \beta_{k+1}, \sigma_{u_1}^2, \sigma_{u_2}^2,$ is $\sigma_{u_1 u_2},$ and σ_u^2 are estimated numerically by maximum likelihood estimation. The fitted mean response in this model can be written as:

$$\hat{E}(y) = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)},$$

or equivalently,

$$\frac{\hat{\pi}}{1 - \hat{\pi}} = \exp\{\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\}.$$

The ratio $\frac{\hat{\pi}}{1-\hat{\pi}}$ represents the estimated odds in favor of $y = 1$. The estimated regression coefficients yield the following interpretation in terms of the estimated odds. If x_1 is numeric, $\left(\exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\%$ represents the estimated percent change in the odds for a one-unit increase in x_1, given that all the other predictors remain fixed. This can be seen by writing

$$\frac{\frac{\hat{\pi}|_{x_1+1}}{1-\hat{\pi}|_{x_1+1}} - \frac{\hat{\pi}|_{x_1}}{1-\hat{\pi}|_{x_1}}}{\frac{\hat{\pi}|_{x_1}}{1-\hat{\pi}|_{x_1}}} \cdot 100\%$$

$$= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1(x_1+1) + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t) - \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)} \cdot 100\%$$

$$= \left(\exp\left(\hat{\beta}_1\right) - 1\right) \cdot 100\% .$$

If x_1 is a 0-1 variable, then $\exp\left(\hat{\beta}_1\right) \cdot 100\%$ can be interpreted as the estimated ratio of odds for $x_1 = 1$ and that for $x_1 = 0$, under the condition that the other predictors are held constant. We demonstrate this by writing

$$\frac{\left(\frac{\hat{\pi}|_{x_1=1}}{1-\hat{\pi}|_{x_1=1}}\right)}{\left(\frac{\hat{\pi}|_{x_1=0}}{1-\hat{\pi}|_{x_1=0}}\right)} \cdot 100\% = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)} \cdot 100\% = \exp\left(\hat{\beta}_1\right) \cdot 100\%.$$

From the fitted model, for values of predictor variables $x_1^0, \ldots, x_k^0$, and t^0, the predicted probability π is:

$$\pi^0 = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0\right)} .$$

Furthermore, the generalized estimating equations model for the binary response in longitudinal setting has the mean response $\hat{E}(y) = \frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t\right)}$ and unstructured, autoregressive, exchangeable, or independent working correlation matrix.

3.3.2. Data Description Anthrax (Binary response)

TABLE 4. Description of variables in Anthrax dataset

Name	Description	Type	Values
remission_from_anthrax	The binary response variable, measuring whether the person has any symptoms of Anthrax.	Binary	Either 0 or 1 0=No 1=Yes
age	The age of patients recorded in the study.	Categorical Binary	Ranges from 21 years old to 80 years old
medicine	Patients were randomly assigned to either a treatment group (Tx) or a control group (Cx). The treatment group received medicine (Antitoxin) while the control group received an unknown placebo drug.	Binary Numeric	0=Tx (Antitoxin) 1=Cx (Placebo)
gender	Gender of patients	Numeric	Male(M) or Female(F)
risk	The risk (i.e. chance) of contacting anthrax. Abbreviated on a scale of 1(very low risk) to 5 (very high risk)	Binary Numeric	Ranges from 120 lbs to 300 lbs
contacted	Patient had possible contact with someone or something that showed symptoms of anthrax prior to entering the clinical trial.	Binary categorical	Either Y or N
Months	Patients were recorded once every month for 12 months to see if they have symptoms of anthrax. 0=No presence of anthrax 1=Presence of anthrax	Numeric	Either 0 or 1

The Anthrax dataset is a simulated longitudinal dataset with the purpose of determining whether Antitoxin (Tx) is effective against anthrax, a skin infection caused by bacteria commonly found in soil. The simulated dataset contains $n = 100$ patients of various ages. Prior

to entering the study, the patients were asked if they had contact with either a human or an animal that showed possible signs of Anthrax. If so, then this information could provide researchers evidence about the effectiveness of Antitoxin. Next, patients were randomly assigned to either a treatment (Tx) group or control (Cx) group. The treatment group received Antitoxin while the control group patients received a placebo. In follow up survey, the researchers contacted each patient once a month for any symptoms of Anthrax. The results were recorded each month for a total of 12 months.

3.3.3. Application

In this data set, the response variable measured was presence or absence of anthrax symptoms. The histogram of the response variable is given below (see Figure 10).

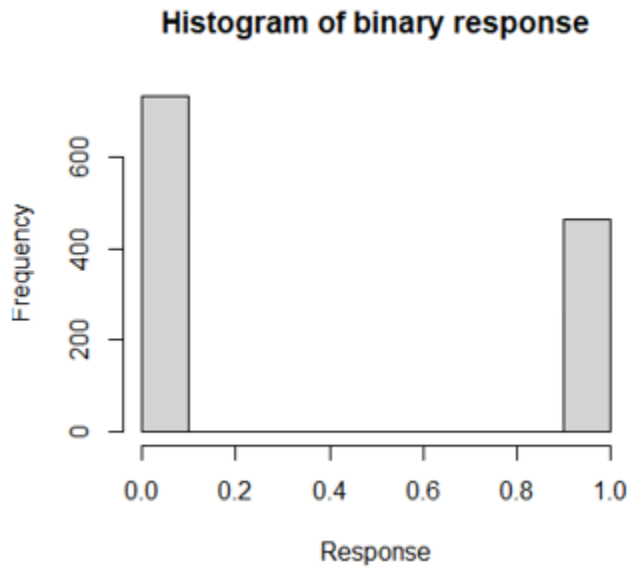

FIGURE 10. Histogram of Binary Response.

Next, we fitted a random slope and intercept binary logistic model. The output can be found in Table 6A in Appendix D. From the output, at the $\alpha = 0.05$ level of significance, the Antitoxin, prior exposure to Anthrax, and month of inspection were all deemed significant

predictors. Out of all the significant predictors, we noticed the Antitoxin was by far the most significant one, yielding a very minuscule P-Value of $p = 5.00 \cdot 10^{-15}$. Consequently, we can therefore conclude that Antitoxin was indeed very effective against anthrax. Following our analysis, we can write the fitted model as:

$$\frac{\hat{\pi}}{1-\hat{\pi}} = \exp(1.875 - 0.00033Age - 1.661Tx + 0.028Male - 0.056Risk - 0.310 \cdot NoPriorContact - 0.210Month).$$

Next, we conduct the deviance test with the null model being the ordinary binary logistic model. The P-value for this test is very small (see Table 5 in Appendix D), leading to conclusion that the longitudinal model has a better fit.

The interpretation of the significant regression coefficients is as follows. The estimated odds in favor of anthrax for patients in the treatment group is $\exp(-1.661) \cdot 100\% = 19\%$ of that for patients in the control group (meaning that the treatment is effective). In addition, the estimated odds of having anthrax for patients who had no prior contact with this disease before entering the clinical trial are $\exp(-0.310) \cdot 100\% = 73.34\%$ of those for patients who were exposed to the disease in the past. Lastly, the estimated odds in favor of anthrax change by $(\exp(-0.211) - 1)) \cdot 100\% = -19.03\%$ every month, that is, decrease every month by 19.03%.

Finally, we would like to predict the probability of a certain patient showing remission from anthrax by month 6. Suppose this patient is a 29-year-old male farmer who is at high risk of contracting the disease. His family, who are also farmers, has had a history of contracting anthrax (i.e., showed previous exposure to this disease). Suppose also, the patient was randomly

assigned to the treatment group. The predicted probability of remission from anthrax for this patient is

$$\hat{\pi}^0 = \frac{\exp\,(r^0)}{1 + \exp\,(r^0)}$$

where $r^0 = 1.875 - 0.00033(29) - 1.661(1) + 0.028(1) - 0.056(5) - 0.310(1) - 0.210(6) = -1.61757$. Thus,

$$\hat{\pi}^0 = \frac{\exp\,(-1.61757)}{1 + \exp\,(-1.61757)} = 0.16554.$$

Next, we fit a generalized estimating equations model for the binary logistic response using the autoregressive, exchangeable, and independent working correlation matrices. The unstructured model was not able to converge, so that was omitted from consideration. Tables 6A, 6B, and 6C in Appendix D present the outputs of the GEE models for this response. Since all three GEE models shared the same QIC of 1441, we conclude that all three of the models can be a good fit for this data set. Therefore, we will arbitrarily choose the independent working correlation matrix GEE model as our fitted model.

From the output, the fitted GEE model is written as

$$\frac{\hat{\pi}}{1 - \hat{\pi}} = \exp(1.403 + 0.00077Age - 1.23Tx + 0.028Male - 0.041Risk - 0.383 \cdot NoPriorContact - 0.163Month).$$

The working correlation matrix is the identity matrix

$$\mathbf{R_i} = \begin{pmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}.$$

The interpretation of the significant regression coefficients is as follows. The estimated odds in favor of remission from anthrax for patients in the treatment group are $\exp(-1.23) \cdot 100\% = 29.23\%$ of those for patients in the control group, meaning that the treatment is efficient. Secondly, the estimated odds for those who had no prior contact with this disease before entering the clinical trial are $\exp(-0.383) \cdot 100\% = 68.18\%$ of those patients who were exposed to the disease in the past. Lastly, the estimated odds of anthrax change by $(\exp(-0.163) - 1) \cdot 100\% = -15.05\%$ each month, that is, decrease every month by 15.05%.

Using the fitted GEE model with the independent working correlation matrix, we predict the probability of remission from anthrax for the patients described previously in this section. The predicted value is $r^0 = 1.403 + 0.00077(29) - 1.23(1) + 0.028(1) - 0.041(5) - 0.383 \cdot (1) - 0.163(6) = -1.34267$, and

$$\hat{\pi}^0 = \frac{\exp(r^0)}{1 + \exp(r^0)} = \frac{\exp(-1.34267)}{1 + \exp(-1.34267)} = 0.207071.$$

3.4. Regressions for Poisson Response

3.4.1. Theoretical Framework

In a random slope and intercept Poisson regression model the response variable y follows a Poisson distribution with the probability mass function $P(y) = \frac{\lambda^y e^{-\lambda}}{y!}$, $y = 0, 1, 2, \ldots$. For the ith individual at time t_j, predictors $x_{1ij}, \ldots, x_{kij}$ and fixed values u_{1i} and u_{2i} of the random intercept and slope, respectively, the parameter λ_{ij} is written as $\lambda_{ij} = \exp\big(\beta_0 + \beta_1 x_{1ij} + \cdots + \beta_k x_{kij} + \beta_{k+1} t_j + u_{1i} + u_{2i} t_j\big)$. The random intercepts u_{1i}'s are independent $N(0, \sigma_{u_1}^2)$ random variables, the random slopes u_{2i}'s are independent $N(0, \sigma_{u_2}^2)$ random variables, and the covariance between u_{1i} and u_{2i} is $\sigma_{u_1 u_2}$.

The fitted model has the estimated rate $\hat{\lambda} = \hat{E}(y) = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_n x_k + \hat{\beta}_{k+1} t)$. The parameters of the model are $\beta_0, \ldots, \beta_{k+1}, \sigma^2_{u_1}, \sigma^2_{u_2}$, and $\sigma_{u_1 u_2}$, which are estimated by the maximum-likelihood method.

From the fitted model, the estimates of the regression coefficients yield the following interpretation. For a numeric predictor x_1, the estimated change in rate when x_1 increases by one unit, while all the other predictors are held fixed, is equal to:

$$\frac{\hat{\lambda}|_{x_1+1} - \hat{\lambda}|_{x_1}}{\hat{\lambda}|_{x_1}}$$

$$= \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1(x_1 + 1) + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t) - \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)}$$

$$= \exp(\hat{\beta}_1) - 1.$$

Thus, $(\exp(\hat{\beta}_1) - 1) \cdot 100\%$ is also equivalently interpreted as the estimated percent change in rate when x_1 increases by one unit, given all the other predictors are fixed.

If the predictor x_1 is a 0-1predictor, then the ratio of the estimated rates when $x_1 = 1$ and when $x_1 = 0$ is equal to:

$$\frac{\hat{\lambda}|_{x_1=1}}{\hat{\lambda}|_{x_1=0}} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)}{\exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 0 + \cdots + \hat{\beta}_k x_k + \hat{\beta}_{k+1} t)} = \exp(\hat{\beta}_1).$$

Hence, $\exp\{\hat{\beta}_1\} \cdot 100\%$ is equivalently interpreted as the percent ratio of estimated rates when $x_1 = 1$ and $x_1 = 0$, given that the other predictors remain constant.

From the fitted model, for a given set of predictors $x_1^0, \ldots, x_k^0, t^0$, the predicted response is found as $y^0 = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_1^0 + \cdots + \hat{\beta}_k x_k^0 + \hat{\beta}_{k+1} t^0)$.

Further, the generalized estimating equations approach models the response variable as a Poisson random variable with mean $E(y) = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + \beta_{k+1} t)$ and an

unstructured, autoregressive, exchangeable, or independent working correlation matrix. As before, the best-fitted model has the smallest value of the QIC criterion.

3.4.2. Data Description Cigarettes (Poisson Response)

TABLE 5. Description of Variables in Cigarette Dataset

Name	**Description**	**Type**	**Values**
N_CIGARETTES	This is our response variable, number of cigarettes smoked.	Numeric	Varies based on input
SEX	Sex of patient	Binary Categorical	M or F
TRT	Patients were randomly assigned to either a treatment group (Tx) or a control group (Cx). The treatment group received medicine (Chantix) while the control group received an unknown placebo drug.	Binary Categorical	Tx or Cx
AGE	Age of patients	Numeric	Ranges from age 21 to age 80
Weight	Weight of patients (in lbs)	Numeric	Ranges from 101 pounds to 297 pounds
Intention	Did the patient intend to quit smoking prior to entering this clinical trial? • 0=No • 1=Yes	Binary Categorical	Either 0 or 1
Addiction.Status	A numeric scale from 1 to 5 that measures the level of patient cigarette addiction • 1= Minimal level of addiction • 5= Maximum level of addiction. Medicine is mandatory.	Numeric	Ranges from 1 to 5

TABLE 5. Continued

Name	Description	Type	Values
Month	At the end of each month, patients were asked how many cigarettes they smoked while additionally taking the drug assigned with their group. The data were recorded every month for 6 months to see if the number of cigarettes smoked per month changed with the drug.	Numeric	Number of cigarettes smoked varies by month

The Cigarette dataset is a simulated dataset similar in structure to the Cancer dataset shown in Section 3.1.3. It is a longitudinal dataset where the dependent variable models count data, the number of cigarettes smoked per month. The purpose of this dataset was to determine the effectiveness of Chantix, a prescription medicine drug developed to help people stop smoking, against an unknown placebo drug. The dataset contains $n = 100$ simulated patients of varying levels of cigarette addiction, ranging from a scale 1 to 5 where 1 represents no addiction and 5 represents an extreme addiction to cigarettes. In this scenario, a patient with an addiction score of 4 or 5 would highly benefit from either seeing a doctor or receiving Chantix, as opposed to a patient with a score of 1. Prior to entering the study, the researchers asked the patients if they had any intention to stop smoking, or not. Patients with the desire to quit smoking would likely benefit from the study compared to those without any desire to stop smoking. Next, the patients were randomly assigned into either a treatment or a control group, with the treatment group receiving Chantix, and the control group receiving a placebo drug. Furthermore, the patients were asked to take one pill of the drug assigned by their group daily and monitor the number of cigarettes smoked. Following the patients' cigarette use, the researchers asked each patient the

total amount of cigarettes they smoked at the end of each month. The number of cigarettes each patient smoked was recorded at the end of each month for a total of six months during the study.

3.4.3. Application

In this data set, the response variable, the number of cigarettes smoked every month, follows a Poisson distribution. To verify that claim, we plotted the histogram of the response variable.

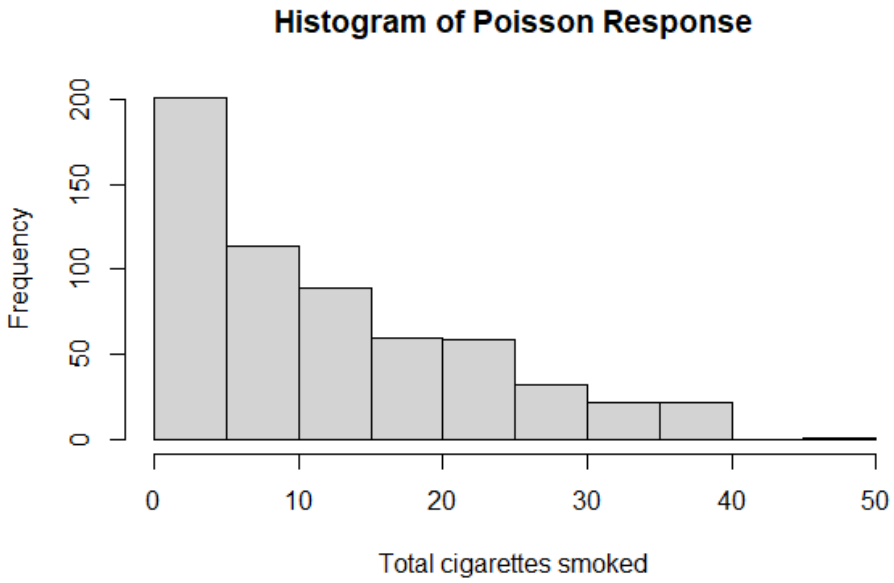

FIGURE 11. Histogram of Poisson response.

Figure 11 depicts the histogram for the number of cigarettes smoked per month. We can see that the distribution resembles the Poisson probability mass function, and so we fit the Poisson random slope and intercept model. The output can be found in Appendix D, Table 8A. At the $\alpha = 0.05$ level of significance, we observed that the treatment group, level of addiction for patients, and the months were all deemed very significant predictors. All other predictors were very insignificant as they had very large P-values.

Thus, from the output, we can write the fitted model where the response models the number of cigarettes smoked every month as

$$\hat{\lambda} = \hat{E}(\text{y}) = \exp(0.970 + 0.058Male + 0.247Tx - 0.00091Age - 0.00029Weight + 0.090Intention + 0.508AddictionStatus - 0.230Month).$$

Next, we employ the deviance test to compare the model fit of the null model against the fitted random slope and intercept model. From the output shown in Table 7 of Appendix D, since the P-value was less than 0.05, we reject the null model and conclude that the fitted model fits the data better.

Next, we interpret the significant estimated regression coefficients. From the fitted model, observe that those within the treatment group showed a $\exp(0.247) \cdot 100\% = 128\%$ decrease in cigarette consumption, more than double compared to those of the control group. Also, For each level increase in cigarette addiction, the estimated average total amount of cigarettes each patient smoked increases by $(\exp(0.508) - 1) \cdot 100\% = 66.20\%$. Lastly, each month, the estimated average total amount of cigarettes each patient smoked changes by $(\exp(-0.23) - 1) \cdot 100\% = -20.55\%$, that is, every month, the estimated average number of cigarettes smoked by each patient decreases by about 20.55%.

Next, our goal now is to predict the number of cigarettes smoked by a 47-year-old female weighing 150 lbs. This person has developed a serious nicotine addiction ever since she started smoking cigarettes at the age of 21 and is desperate to quit. Therefore, the researchers categorized her nicotine addiction as “5”, very high levels of addiction.

From our given data, we will now predict the number of cigarettes this person smoked over the course of the study.

$$\lambda^0 = \exp(0.970 + 0.058Male^0 + 0.247Tx^0 - 0.00091Age^0 - 0.00029Weight^0 + 0.090Intention^0 + 0.508AddictionStatus^0 - 0.230Month^0)$$

At month 3: $\lambda^0 = \exp(0.970 + 0.058(0) + 0.247(1) - 0.00091(21) - 0.00029(150) + 0.090(1) + 0.508(5) - 0.230(3)) = 22.0738$ cigarettes.

At month 6: $\lambda^0 = \exp\big(0.998 - 0.066(0) - 0.001(150) - 0.058(1) + 0.00039(150) + 0.533(5) - 0.189(6) - 0.06(0)\big) = 11.0717$ cigarettes.

Next, we fit a generalized estimating equations model shown in Section 3.1.2. for the Poisson response using the autoregressive, exchangeable, independent, and unstructured working correlation matrices (refer to Appendix D, Tables 8B, 8C, 8D, and 8E for outputs). Since the output for the autoregressive had the smallest QIC value of -25573.84, we will choose the autoregressive model as our fitted GEE model for both our interpretation and prediction.

Thus, from the output table (see Appendix D, Table 8B) the fitted model has the estimated rate

$$\hat{\lambda} = \hat{E}(\mathrm{y}) = \exp(1.351 + 0.108Male - 0.42Tx - 0.001Age - 0.00022Weight - 0.026Intention + 0.486AddictionStatus - 0.121Month),$$

and the estimated working correlation matrix for $p = 6$ months is equal to

$$\widehat{\mathbf{R}}_{\mathbf{i}}(\hat{\boldsymbol{\alpha}} = \mathbf{0.635}) = \begin{pmatrix} 1 & \hat{\alpha} & \hat{\alpha}^2 & \cdots & \hat{\alpha}^{p-1} \\ \hat{\alpha} & 1 & \hat{\alpha} & \cdots & \hat{\alpha}^{p-2} \\ \hat{\alpha}^2 & \hat{\alpha} & 1 & \cdots & \hat{\alpha}^{p-3} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \hat{\alpha}^{p-1} & \hat{\alpha}^{p-2} & \hat{\alpha}^{p-3} & \cdots & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 0.635 & (0.635)^2 & \cdots & (0.635)^5 \\ 0.635 & 1 & 0.635 & \cdots & (0.635)^4 \\ (0.635)^2 & 0.635 & 1 & \cdots & (0.635)^3 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ (0.635)^5 & (0.635)^4 & (0.635)^3 & \cdots & 1 \end{pmatrix}.$$

We will now give our interpretation of the significant predictors. The estimated average number of cigarettes smoked by the patients within the treatment group is $\exp(-0.419) \cdot 100\% = 65.77\%$ of those for the control group. Also, as the level of addiction increases, the estimated average number of cigarettes each patient smoked increases by $(\exp(0.486) - 1) \cdot 100\% = 62.58\%$. Finally, each month, the estimated average number of cigarettes each patient smoked changes by $(\exp(-0.121) - 1) \cdot 100\% = -11.40\%$. That is, every month, the estimated average number of cigarettes smoked by each patient decreases by about 11.40%.

Using the same example, under the autoregressive GEE model, we will now predict the number of cigarettes this patient will smoke at months 3 and 6. We have

$$\hat{\lambda}^0 = \exp(1.351 + 0.108Male^0 - 0.42Tx^0 - 0.001Age^0 - 0.00022Weight^0 - 0.026Intention^0 + 0.486AddictionStatus^0 - 0.121Month^0).$$

At month 3: $\lambda^0 = \exp(1.351 + 0.108 \cdot 0 - 0.42 \cdot 1 - 0.001 \cdot 21 - 0.00022 \cdot 150 - 0.026 \cdot 1 + 0.486 \cdot 5 - 0.121 \cdot 3) = 18.5042$ cigarettes.

At month 6: $\lambda^0 = \exp(1.351 + 0.108 \cdot 0 - 0.42 \cdot 1 - 0.001 \cdot 21 - 0.00022 \cdot 150 - 0.026 \cdot 1 + 0.486 \cdot 5 - 0.121 \cdot 6) = 12.8713$ cigarettes.

To conclude, from month 3 to month 6, the patient is predicted to show a steady decrease in monthly cigarette consumption. Therefore, we conclude that Chantix would indeed be helpful in lowering the monthly cigarette use for this patient.

CHAPTER 4

DATA MONITORING

4.1 Background Information

4.1.1 Determining the Length of a Clinical Trial

In a clinical trial, the sample size is the total number of subjects enrolled in the trial. A larger sample size is necessary to deliver a more detailed and accurate information about the efficacy of the tested product. The minimum required sample size should be determined prior to commencement of a clinical trial. The steps that have to be followed in calculation of the minimum required sample size are shown below.

Step 1: the endpoint, known as the measure of the target outcome, of a clinical trial is defined. There are three types of endpoints. The first endpoint is called a pre-specified percentage change from the baseline value of a medical measurement. The second endpoint is a pre-specified actual change from the baseline value in some medical characteristic. The last endpoint is a pre-specified rate of a certain adverse event. This is defined as the ratio between the total number of events and the total trial time for all subjects.

Step 2: To model the endpoint, a certain family of distributions is used, where its parameters are estimated from the data.

Step 3: The null and alternative hypotheses for the endpoint are identified. The type of hypothesis test (two-tailed, left-tailed, or right-tailed) used depends on the nature of the experiment.

Step 4: The probability of a type I error, known as the significance level of the test, is determined. The probability of type I error is the probability of accepting the alterative hypothesis H_1, given the null hypothesis H_0 is true.

Step 5: The probability of a type II error along with the minimum detectable difference, if appropriate, is determined. The probability of a type II error is the probability of failing to reject H_0 when a specific alternative hypothesis is true. For instance, $H_1: \mu_t - \mu_c = \delta$ is true, where δ is the minimum detectable difference between the mean responses in the two groups.

Within the clinical trial setting, a power function of a test (i.e., $1 - \beta$) is frequently used to determine the probability of rejecting the null hypothesis, given that the alternative is valid.

We will now present a numeric example illustrating the power analysis to determine the required sample size when two means are compared. To determine the required sample size for the test of $H_0: \mu_t = \mu_c$ against $H_0: \mu_t > \mu_c$, we assume that $\bar{x}_t \sim N(\mu_t, \frac{\sigma^2}{n})$ and $\bar{x}_c \sim N(\mu_c, \frac{\sigma^2}{n})$. We also assume that σ is known (from previous studies, says), and that the probabilities of type I and type II errors, respectively α and β, are pre-determined. The value of β is given for a specific alternative $H_0: \mu_t - \mu_c = \delta$ where δ is fixed. Therefore, under the null hypothesis $Z = \frac{\bar{x}_t - \bar{x}_c}{\sigma\sqrt{\frac{2}{n}}} \sim N(0,1)$, and we can write the acceptance region for the null hypothesis as

$\{Z < k\} = \left\{ \frac{\bar{x}_t - \bar{x}_c}{\sigma\sqrt{\frac{2}{n}}} < k \right\} = \left\{ \bar{x}_t - \bar{x}_c < k\sigma\sqrt{\frac{2}{n}} \right\}$ for some critical value k. On the other hand, if a specific alternative $H_1: \mu_t - \mu_c = \delta$ holds, then $\bar{x}_t - \bar{x}_c \sim N\left(\delta, 2\frac{\sigma^2}{n}\right)$.

Further, the values of k and n can be found from the equations for the probabilities of type I and type II errors. The equations are:

i. $1 - \alpha = P\big(Z < k \big| Z \sim N(0,1)\big) = \Phi(k)$

and

ii. $$\beta = P\left(\bar{x}_t - \bar{x}_c < k\sigma\sqrt{\frac{2}{n}} \middle| \bar{x}_t - \bar{x}_c \sim N\left(\delta, \frac{2\sigma^2}{n}\right)\right) = \Phi\left(k - \frac{\delta}{\sigma\sqrt{\frac{2}{n}}}\right)$$

where Φ is defined as the cumulative distribution function of a standard normal distribution. From the first equation, the critical value of the acceptance region $k = \Phi^{-1}(1-\alpha)$. Substituting it into the second equation yields

$$\Phi^{-1}(\beta) = k - \frac{\delta}{\sigma\sqrt{\frac{2}{n}}} = \Phi^{-1}(1-\alpha) - \frac{\delta}{\sigma\sqrt{\frac{2}{n}}}.$$

Solving for n we obtain the required sample size (per group),

$$n = 2\left(\frac{\sigma}{\delta}\right)^2 \left(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\right)^2.$$

In practice, n should be taken as the smallest integer exceeding this calculated value, that is,

$$n = \left\lceil 2\left(\frac{\sigma}{\delta}\right)^2 \left(\Phi^{-1}(1-\alpha) - \Phi^{-1}(\beta)\right)^2 \right\rceil,$$

which results in the probability of a Type II error being slightly smaller than the specified value.

To give a numeric example, suppose $\sigma = 17.4, \alpha = 0.05,\ \beta = 0.25$, and $\delta = 8$. Plugging the values into the above expression yields the required sample size per group, $n = \left\lceil 2\left(\frac{17.4}{8}\right)^2 \left(\Phi^{-1}(1-0.05) - \Phi^{-1}(0.25)\right)^2 \right\rceil = \lceil 50.89541 \rceil = 51$. That is, 51 patients should be enrolled in each group in the clinical trial to achieve the power of the test of at least $1-\beta = 0.75$. The actual probability of type II error in this case will be

$$\beta' = \Phi\left(\Phi^{-1}(0.95) - \frac{8}{17.4\sqrt{\frac{2}{51}}}\right) = 0.249244,$$

and thus, the actual power of the test will be $1-\beta = 0.750756$.

4.1.2 Interim Data Monitoring

Interim data monitoring in clinical trials is a type of data analysis performed while the trial is still in progress to determine if the trial should continue or not. A clinical trial might be terminated earlier if there is enough evidence to justify the claim that a tested product is either superior, or worse than its standard counterpart. The decision on either to conduct a full-length trial of the product's efficacy or stop early is determined by the amount of confidence the researchers have in the tested product. If the researchers have high confidence in the product's ability to succeed, interim data monitoring is an appropriate solution because there is a high chance of early termination of the trial. Below, we present two major statistical methods for calculating interim sample sizes.

4.1.3 Classical Group Sequential Testing

Classical group sequential testing is used in a randomized trial involving two groups (treatment and control). The procedure is as follows.

Once data for n subjects in each group become available, an interim analysis is conducted on the $2n$ subjects. From there, the two groups are compared statistically. If the null hypothesis H_0 is rejected in favor of the alternative H_1, then the trial is terminated, and the conclusion is achieved. On the other hand, if the null H_0 is kept, then the trial continues until data for another set of $2n$ subjects become available. Then a statistical test on $4n$ subjects is conducted. Similarly, if the alternative hypothesis H_1 is accepted, then the trial is terminated. Otherwise, the trial repeats continuously with periodic evaluations until N sets of $2n$ subjects become available and will terminate once the null hypothesis is rejected, favoring the alternative.

For each of the N interim statistical tests, we denote the probability of type I error by α'. The number of interim tests N must be determined a priori. Therefore α' and n can be found if

the overall probabilities of type I and type II errors, α and β, respectively, are specified. They solve the equations:

i. $P(\text{at least one interim test rejects } H_0 \mid H_0 \text{ is true}) = \alpha$
ii. $P(\text{at least one interim test accepts } H_1 \mid H_1 \text{ is true}) = 1 - \beta$.

We will now present an example illustrating sequential testing. Recall from the power analysis example discussed previously, our null and alternative hypotheses of interest are specified as $H_0: \mu_t = \mu_c$ vs. $H_1: \mu_t - \mu_c = \delta$ where our stated parameters are $\alpha = 0.05$, $\beta = 0.25$, $\delta = 8$, and $\sigma = 17.4$ (see Section 4.1.1). For this test to be conducted, an estimated sample size of $n' = 51$ patients per group is required.

Now, suppose we consider a group sequential case with $N = 2$, meaning that we test at most two times, first time on $2n$ subjects (n subjects per group), the second time, if needed, we test on $4n$ subjects ($2n$ subjects per group). To show how this testing works, we introduce $\bar{x}_t^{(i)}$ and $\bar{x}_c^{(i)}$ as the respective group sample means in the ith set of $2n$ subjects, $i = 1, 2$. Now, denote by

$\bar{x}_t = \frac{\bar{x}_t^{(1)} + \bar{x}_t^{(2)}}{2}$ and $\bar{x}_c = \frac{\bar{x}_c^{(1)} + \bar{x}_c^{(2)}}{2}$ the group sample means in the combined set of $4n$ subjects.

We perform the first statistical test of H_0 against H_1 at significance level α' on the initial set of $2n$ subjects where, under the null hypothesis, $\bar{x}_t^{(1)} - \bar{x}_c^{(1)} \sim N\left(0, \frac{2\sigma^2}{n}\right)$ with the probability of the acceptance region equal to $1 - \alpha' = P\left(\frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{\sqrt{\frac{2\sigma^2}{n}}} < k\right) = \Phi(k)$. Thus, if we know the critical value k, we can compute $\alpha' = 1 - \Phi(k)$.

If the trial is not stopped at the first testing, it continues until $4n$ subjects are accrued. The difference in sample means for the set of $4n$ subjects can be calculated as

$$\bar{x}_t - \bar{x}_c = \frac{\bar{x}_t^{(1)} + \bar{x}_t^{(2)}}{2} - \frac{\bar{x}_c^{(1)} + \bar{x}_c^{(2)}}{2} = \frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{2} + \frac{\bar{x}_t^{(2)} - \bar{x}_c^{(2)}}{2}.$$

Since $\frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{2} \sim N\left(0, \frac{\sigma^2}{2n}\right)$ and $\frac{\bar{x}_t^{(2)} - \bar{x}_c^{(2)}}{2} \sim N\left(0, \frac{\sigma^2}{2n}\right)$, we deduce that $\bar{x}_t - \bar{x}_c \sim N\left(0, \frac{\sigma^2}{n}\right)$. We want the acceptance region for the second test also have the critical value k (and thus, the significance level α'). From the definitions of α and β, we can specify two equations for k and n as

$$1 - \alpha = P\left(\frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{\sqrt{\frac{2\sigma^2}{n}}} < k, \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{\sigma^2}{n}}} < k\right)$$

where $\bar{x}_t^{(1)} - \bar{x}_c^{(1)} \sim N\left(0, \frac{2\sigma^2}{n}\right)$ and $\bar{x}_t - \bar{x}_c \sim N\left(0, \frac{\sigma^2}{n}\right)$, and

$$\beta = P\left(\frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{\sqrt{\frac{2\sigma^2}{n}}} < k, \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{\sigma^2}{n}}} < k\right)$$

where $\bar{x}_t^{(1)} - \bar{x}_c^{(1)} \sim N\left(\delta, \frac{2\sigma^2}{n}\right)$ and $\bar{x}_t - \bar{x}_c \sim N\left(\delta, \frac{\sigma^2}{n}\right)$.

The former equation above can be simplified to

$$1 - \alpha = \mathbb{P}\left(Z_1 < k, Z_1 + Z_2 < k\sqrt{2}\right)$$

where $Z_1 = \frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)}}{\sqrt{\frac{2\sigma^2}{n}}}$ and $Z_2 = \frac{\bar{x}_t^{(2)} - \bar{x}_c^{(2)}}{\sqrt{\frac{2\sigma^2}{n}}}$ specify the random variables, which are independent and follow a standard normal distribution. The latter equation becomes

$$\beta = P\left(Z_3 + \frac{\delta}{\sqrt{\frac{2\sigma^2}{n}}} < k, Z_3 + Z_4 + \frac{2\delta}{\sqrt{\frac{2\sigma^2}{n}}} < k\sqrt{2}\right)$$

where $Z_3 = \frac{\bar{x}_t^{(1)} - \bar{x}_c^{(1)} - \delta}{\sqrt{\frac{2\sigma^2}{n}}}$ and $Z_4 = \frac{\bar{x}_t^{(2)} - \bar{x}_c^{(2)} - \delta}{\sqrt{\frac{2\sigma^2}{n}}}$ are independent standard normal random variables.

These equations have to be solved numerically for specific values of α and β. Going back to our example (see Section 4.1.1), we specify $\sigma = 17.4, \alpha = 0.05,\ \beta = 0.25$, and $\delta = 8$. Writing these equations as double integrals, we obtain

$$0.95 = \int_{-\infty}^{k} \int_{-\infty}^{k\sqrt{2}-z_1} \frac{1}{2\pi} e^{-\frac{(z_1^2+z_2^2)}{2}} dz_2 dz_1 = \int_{-\infty}^{k} \phi(z)\Phi(k\sqrt{2}-z)dz\ ,$$

and

$$0.25 = \int_{-\infty}^{k-\frac{\delta\sqrt{n}}{\sqrt{2}\sigma}} \int_{-\infty}^{k\sqrt{2}-\frac{\delta\sqrt{2n}}{\sigma}-z_1} \frac{1}{2\pi} e^{-\frac{(z_1^2+z_2^2)}{2}} dz_2 dz_1 = \int_{-\infty}^{k-\frac{\delta\sqrt{n}}{\sqrt{2}\sigma}} \phi(z)\Phi\left(k\sqrt{2}-\frac{\delta\sqrt{2n}}{\sigma}-z\right)dz$$

where ϕ and Φ denote the pdf and cdf of a standard normal distribution, respectively.

Solved numerically (see Appendix E), $k = 1.88$ which corresponds to $\alpha' = 1 - \Phi(1.88) = 0.03005404,$ and $n = 28.66759$ or, rounding up, $n = 29$ subjects.

In conclusion, the interim group size is 29 patients, meaning that instead of accruing 51 patients in each group and testing the null and alternative hypotheses at the $\alpha = 0.05$ level of significance, the group sequential method with $N = 2$ dictates that investigators test the hypotheses with $n = 29$ subjects per group at the $\alpha' = 0.03$ level of significance. If the first test is inconclusive, a second test using the same level of significance α' with a group size of $2n = 58$ subjects is imposed.

An advantage of using sequential testing is that there is a good chance of stopping the trial earlier. If researchers have strong confidence that a certain product could succeed, they would likely choose the sequential testing method as there is a good chance of stopping the trial after data have been collected and analyzed for $n = 29$ subjects, rather than waiting until 51 subjects are accrued for nonsequential testing. However, a disadvantage of using sequential testing is that if a product does worse than expected, the trial must continue until $2n = 58$

subjects are accrued for each group. Consequently, resulting in a trial longer than that of the nonsequential monitoring (51 subjects).

The example we presented illustrates sequential testing for $N = 2$. A larger number of tests may be used. Then the quantities k and n can be found as numeric solutions of two equations:

$$1 - \alpha = P\left(\bigcap_{m=1}^{N}\{Z_1 + \cdots + Z_m < k\sqrt{m}\}\right)$$

and

$$\beta = P\left(\bigcap_{m=1}^{N}\left\{Z_1 + \cdots + Z_m + \frac{m\,\delta}{\sqrt{\frac{2\sigma^2}{n}}} < k\sqrt{m}\right\}\right)$$

where $Z_1, \ldots, Z_N$ denote independent $N(0,1)$ random variables.

4.2 Bayesian Sequential Procedure

4.2.1 Bayes Theorem

Define a partition of a sample space S as the union of mutually exclusive events. Assume that for some positive integer k, the events $B_1, B_2, \ldots, B_k$ satisfy two conditions:

i. $S = B_1 \cup B_2 \cup \ldots \cup B_k$ (their union is all of the sample space), and

ii. $B_i \cap B_j = \emptyset$ for $i \neq j$ (they are mutually exclusive).

Consider an event A, and suppose we know the probabilities $P(B_i), i = 1, \ldots, k$, of each of the event in the partition, and each conditional probability $P(A|\, B_i), i = 1, \ldots, k$, of A occurring, given each of the event in the partition. Suppose the event A has occurred. Then, the probabilities of B_j for each fixed $j, j = 1, \ldots, k$, can be updated according to the Bayes' formula:

$$P(B_j|A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)}.$$

4.2.2 Prior and Posterior Distributions

Bayesian statistics is the system for describing uncertainty of a future event using the mathematical language of probability. Under the scenario of uncertainty, Bayesians would start with their existing beliefs, known as priors, and update their priors using data analysis to give posterior beliefs. These posterior beliefs can then be used for decisions.

In the setting of Bayesian statistics, to estimate the parameters of interest, one would need to have some prior knowledge of the experiment. This knowledge is modeled as a distribution incorporating a Bayesian's subjective beliefs about the unknown parameter θ prior to examining the data. After gathering the data, the prior distribution is updated using Bayes' theorem to obtain the posterior distribution which is basically a probability distribution that represents the updated beliefs about the estimated parameters after observing the data. The posterior density function is derived as follows.

Assume the data are represented by a random vector $\boldsymbol{X} = (X_1, \dots, X_n)$ from a probability distribution that depends on an unknown parameter $\theta \in \Omega$ that needs to be estimated. Knowing that θ is a fixed parameter, Bayesians model it as a random variable Θ that follows a certain prior probability distribution over the set Ω. We write the prior pdf of Θ as $\pi(\theta),\ \theta \in \Omega$.
Next, we assume that the variables $X_1, \dots, X_n$ are independent, and have identical pdf $f(x|\theta)$, given $\Theta = \theta$. The likelihood function of $X_1, \dots, X_n$, given $\Theta = \theta$, can be written as

$$L(x_1, x_2, \dots, x_n|\theta) = f(x_1|\theta) \cdot f(x_2|\theta) \cdots f(x_n|\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

In this notation, the conditional pdf of Θ given data $x_1, \dots, x_n$ is equal to

$$f_{\Theta}(\theta \mid x_1, \dots, x_n) = \frac{L(x_1, x_2, \dots, x_n|\theta) \cdot \pi(\theta)}{\int_{\Omega} L(x_1, x_2, \dots, x_n|\theta) \cdot \pi(\theta) d\theta} = \frac{\prod_{i=1}^{n} f(x_i|\theta) \cdot \pi(\theta)}{\int_{\Omega} \prod_{i=1}^{n} f(x_i|\theta) \cdot \pi(\theta) d\theta}.$$

This is called the posterior distribution of Θ since it represents our knowledge about the parameter Θ after observing the data. Note that the posterior distribution is proportional to the product of the likelihood function and the prior distribution:

$$f_{\Theta}(\theta \mid x_1, \dots, x_n) \propto L(x_1, x_2, \dots, x_n|\theta) \cdot \pi(\theta).$$

The proportionality is defined as equality up to a multiplicative normalizing constant $\left(\int_{\Omega} L(x_1, x_2, \dots, x_n|\theta) \cdot \pi(\theta) d\theta\right)^{-1}$, not depending on θ.

Within a clinical trial setting, a Bayesian sequential procedure is used to model the clinical endpoint as a random variable Θ where the prior density of Θ, $\pi(\theta) = f_{\Theta}(\theta)$, can be chosen in multiple ways. If there is strong belief that a tested product would succeed, an enthusiastic prior is chosen. An enthusiastic prior assumes that an alternative hypothesis (i.e., $H_1: \Theta \in \Omega_1$) is more likely to hold than the null hypothesis (i.e., $H_0: \Theta \in \Omega_0$) where Ω_0 and Ω_1 partition Ω. Alternatively, if researchers are skeptical about the tested product and assume that $P(H_1) \leq P(H_0)$, then a skeptical prior can be used to model a prior distribution.

To use the Bayesian inference, we first compute the posterior density of Θ given the data are observed. From the Bayes' theorem, the posterior density is equal to

$$f_{\Theta}(\theta|data) = \frac{f(data|\Theta = \theta)\pi(\theta)}{\int_{\Omega} f(data|\Theta = \theta)\pi(\theta) d\theta}.$$

Finally, to arrive at the decision of either rejecting or accepting (failing to reject) the null hypothesis, the posterior probability of the null hypothesis is calculated as

$$P(H_0|data) = \int_{\Omega_0} f_{\Theta}(\theta|data) d\theta,$$

and the decision is made according to the following rule:

i. If $P(H_1|data) < 0.05$, the trial is stopped and the product is not marketed.
ii. If $P(H_1|data) > 0.95$, the trial is stopped and the product is marketed.
iii. If $0.05 \leq P(H_1|data) \leq 0.95$, the trial continues.

4.2.3 Comparing Conjugate vs. Nonconjugate Priors

Note that in the expression for the posterior density, calculating the denominator is a computationally-intensive task. To avoid it, it is often convenient to use conjugate priors, priors chosen in such a way that the posterior density would have the same algebraic form as the prior. It is a very cost-efficient approach since in this case there is no need to calculate the integral (the normalizing constant). Conjugate priors come from a class of priors that are conjugate to the class of likelihood functions. That is, being conjugate is a class property. Below we will consider some concrete examples of conjugate priors.

Alternatively, nonconjugate priors can be considered. In some situations, for instance, researchers don't possess any prior knowledge about the parameter, so using a uniform prior would be advisable. In effect, using a nonconjugate prior would mean letting the likelihood function of the observed data play a major role in forming the posterior distribution. If for example, a new medication was tested in animals only, and there is no prior knowledge if in humans it performs better or worse, then the correct approach would be to rely on a prior distribution that may appropriately shape the posterior distribution.

Below, we present examples using Poisson, normal, and binomial distributions for the data in clinical settings, and compare the use of conjugate and nonconjugate priors for the parameters.

4.3 Poisson Inference

4.3.1. Poisson-Gamma Example

A clinical trial is conducted to test the performance of a new heart valve where the endpoint of the trial is the rate of certain valve-related complication. The rate of a complication is defined as the total number of cases divided by the total amount of years accumulated by all patients in the trial. This is known as "patient-years". For example, if there were 9 cases in 500 patient years, then the rate of complication would be 0.018 or 1.8%. From this example, the complication rate R for the new heart valve was compared to its historic value of $R_h = 0.012$. The primary endpoint is an endpoint that is used in power analysis to pre-determine the required sample size of a trial. From all possible valve-related complications, endocarditis (inflammation of heart lining) is chosen as the primary endpoint because it is the rarest and takes the longest time to be detected.

Next, we perform a hypothesis test to determine if the new valve performed better or worse than the historical one. We specify the null hypothesis, indicating that the new valve performed worse than the historic one, as $H_0: R \geq 2R_h = 0.024$ against the alternative hypothesis $H_1: R < 2R_h = 0.024$. If the null hypothesis is not rejected, then we arrive at the conclusion that the new valve performed worse than the historic one and should not be used.

An assumption is made that the number of endocarditis events (specified as N) during time T follows a Poisson distribution with mean $\lambda = RT$ and probability mass function

$$P(N = n) = \frac{(RT)^n e^{-RT}}{n!}, n = 0,1,2 \ldots.$$

Now we model R as a random variable, and choose a prior distribution from the class of distributions that is conjugate to the class of Poisson distributions. To this end, we view the

above probability mass function as a function of R and note that it has the algebraic form of a gamma distribution. It means that gamma priors are conjugate to Poisson likelihood functions. We specify the prior distribution of R as $\Gamma(\alpha, \beta)$ with density

$$\pi(r) = \frac{r^{\alpha-1}\exp\left(-\frac{r}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}}, \quad r, \alpha, \beta > 0.$$

It is not difficult to deduce that the posterior distribution of R given that n endocarditis evens have been observed during time t is again a gamma distribution with parameters $n + \alpha$ and $\left(t + \frac{1}{\beta}\right)^{-1}$. We show the derivation below.

$$f_R(r|n,t) \propto f(t,n|r)\pi(r) = \frac{(rt)^n e^{-rt}}{n!} \cdot \frac{r^{\alpha-1}\exp\left(-\frac{r}{\beta}\right)}{\Gamma(\alpha)\beta^{\alpha}} \propto r^{n+\alpha-1}e^{-rt-\frac{r}{\beta}} = r^{(n+\alpha)-1}e^{-r/(t+\frac{1}{\beta})^{-1}},$$

which is the algebraic form of the gamma distribution with the said parameters.

Next, the posterior probability that the alternative hypothesis H_1 is correct is computed as

$$P(H_1|data) = P(R < 0.024|n,t) = \int_0^{0.024} f_R(r|n,t)dr$$

$$= \frac{\left(t + \frac{1}{\beta}\right)^{n+\alpha}}{\Gamma(n+\alpha)} \int_0^{0.024} r^{n+\alpha-1} \cdot \exp\left(-r\left(t + \frac{1}{\beta}\right)\right) dr.$$

The decision to accept or reject the alternative hypothesis is based on the following criterion. If $P(H_1|data) < 0.05$, the alternative hypothesis is rejected. On the other hand, if $P(H_1|data) > 0.95$, the alternative hypothesis is accepted. Otherwise, the trial continues.

To specify the parameters α, β, we note that the gamma density is unimodal and right-skewed, meaning there is one peak and a long right tail. For these distributions, the mean, median, mode inequality holds in the form $mode < median < mean$. This implies that $P(R < mode) < P(R < median) < P(R < mean)$, or since $P(R < median) = 0.5$ by

definition, we obtain $P(R < mode) < 0.5 < P(R < mean)$. From our example mentioned previously, if investigators wanted to use enthusiastic prior, they would set the mean of the prior distribution equal to 0.024. From the mean, this gives them the opportunity to specify any desired prior probability of the true H_1 larger than 0.5. Thus, the probability of the accepting the true alternative hypothesis can be written as $0.5 < P(R < mean) = \mathbb{P}(R < 0.024) = P(H_1)$. On the contrary, if the researchers wanted to impose a skeptical prior, they would set the mode of the prior distribution equal to 0.024, resulting in $P(H_1) = P(R < 0.024) = P(R < mode) < 0.5$, and so $P(H_1)$ can take on any value below 0.5.

The gamma distribution with parameters α, β has a mean equal to $\alpha\beta$ and mode equal to $(\alpha - 1)\beta$. From the information, the parameters α and β are computed numerically from the equations:

$$\alpha\beta = 0.024\ (Enthusiastic\ prior),$$

$$(\alpha - 1)\beta = 0.024\ (Skeptical\ prior),$$

and

$P(H_1) = P(R < 0.024) = \int_0^{0.024} \pi(r)dr = \int_0^{0.024} \frac{r^{\alpha-1}\exp(-\frac{r}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} dr$ = value specified by the researchers.

To give a numeric example, suppose the researchers used a skeptical prior, where the probability of the true alternative is specified as $P(H_1) = 0.3$, then the parameters α and β of the prior density satisfy the equations

$$(\alpha - 1)\beta = 0.024, \text{ and}$$

$$0.3 = \int_0^{0.024} \frac{r^{\alpha-1}\exp(-\frac{r}{\beta})}{\Gamma(\alpha)\beta^{\alpha}} dr.$$

After the parameters are evaluated, we will be looking for *n* such that the posterior probability of the alternative

$$P(H_1|data) = P(R < 0.024|n,t) = \int_0^{0.024} f_R(r|n,t)dr$$

$$= \frac{\left(t+\frac{1}{\beta}\right)^{n+\alpha}}{\Gamma(n+\alpha)} \int_0^{0.024} r^{n+\alpha-1} \cdot \exp\left(-r\left(t+\frac{1}{\beta}\right)\right) dr$$

is either less than 0.05 or greater than 0.95. The results of the Bayesian stopping rules are shown in Section 4.2.3.

4.3.2. Poisson Inverse-Gamma Example

Now, let us assume a nonconjugate prior distribution chosen from a family of distributions similar to Poisson. In this case, that would be the inverse-gamma distribution where the prior distribution of R is specified respectively as $\Gamma^{-1}(\alpha,\beta)$ with density $\pi(r) = \frac{\beta^\alpha}{\Gamma(\alpha)} r^{-\alpha-1} \exp\left(-\frac{\beta}{r}\right),\ r,\alpha,\beta > 0$. From the prior distribution, the posterior distribution of R given that *n* endocarditis evens have been observed within *t* patient-years yields a density function that is derived as follows.

$$f_R(r|n,t) \ \propto f(t,n|r)\pi(r) = \frac{(rt)^n e^{-rt}}{n!} \cdot \frac{\beta^\alpha\, r^{-\alpha-1} \exp\left(-\frac{\beta}{r}\right)}{\Gamma(\alpha)} \propto r^{n-\alpha-1} e^{-rt-\frac{\beta}{r}}.$$

Since this posterior density is not of a recognizable algebraic form, we need to compute the normalizing constant. That is, we need to calculate the entire expression

$$f_R(r|n,t) = \frac{r^{n-\alpha-1}\, e^{-(rt+\beta/r)}}{\int_0^\infty r^{n-\alpha-1}\, e^{-(rt+\beta/r)} dr}\ .$$

Next, the posterior probability that the alternative hypothesis H_1 is correct is computed as

$$P(H_1|data) = P(R < 0.024|n,t) = \int_0^{0.024} f_R(r|n,t)dr$$

$$= \frac{\int_0^{0.024} r^{n-\alpha-1}\, e^{-(rt+\beta/r)}\, dr}{\int_0^{\infty} r^{n-\alpha-1}\, e^{-(rt+\beta/r)} dr}.$$

The inverse-gamma distribution with parameters α, β has a mean equal to $\frac{\beta}{\alpha-1}$ and mode equal to $\frac{\beta}{\alpha+1}$. From this information, the parameters α and β are estimated from the equations shown below as

$$\frac{\beta}{\alpha - 1} = 0.024\ (Enthusiastic\ prior),$$

$$\frac{\beta}{\alpha + 1} = 0.024\ (Skeptical\ prior),$$

and

$P(H_1) = P(R < 0.024) = \int_0^{0.024} \pi(r)dr = \int_0^{0.024} \frac{\beta^\alpha}{\Gamma(\alpha)} r^{-\alpha-1} \exp\left(-\frac{\beta}{r}\right) dr$ where $P(H_1)$ is specified by the researchers.

To give a numeric example, suppose the researchers used a skeptical prior, where the probability of the true alternative is specified as $P(H_1) = 0.3$, then the parameters α and β of the prior density satisfy the equations.

$$\frac{\beta}{\alpha + 1} = 0.024 \quad \text{and}$$

$$0.3 = \int_0^{0.024} \frac{\beta^\alpha}{\Gamma(\alpha)} r^{-\alpha-1} \exp\left(-\frac{\beta}{r}\right) dr\,.$$

After α and β are calculated, we will be looking for *n* such that the posterior probability of the alternative $P(H_1|data) = P(R < 0.024|n,t) = \int_0^{0.024} f_R(r|n,t)dr$

$$= \frac{\int_0^{0.024} r^{n-\alpha-1} e^{-\left(rt+\frac{\beta}{r}\right)} dr}{\int_0^{\infty} r^{n-\alpha-1} e^{-\left(rt+\frac{\beta}{r}\right)} dr}$$

is either less than 0.05 or greater than 0.95. The results are shown in Section 4.4.3.

4.3.3. Poisson Conjugate vs. Nonconjugate Stopping Results

According to FDA, the minimum length of the trial without the Bayesian monitoring is 800 patient-years (Grunkemeier, G.L., Johnson, D.M., & Naftel, D.C. 1994). To illustrate how the Bayesian approach works in this case, we suppose that the investigators decide to conduct interim Bayesian analyses at $t = 400, 500,$ and 600 patient-years. For each value of t, we compute the required sample size n, for which the posterior probability of the true alternative is below 0.5 or above 0.95. These are the stopping values for the analyses. In the tables below, we present numerical values for the conjugate case (gamma prior) and nonconjugate case (inverse-gamma prior).

TABLE 6. Results of Bayesian Monitoring in Poisson-Gamma Example

t	*n*	$P(H_1\|n,t)$	*t*	*n*	$P(H_1\|n,t)$
400	**3**	**0.97713**	400	14	0.07695
	4	0.94766		**15**	**0.04573**
	5	0.89733		16	0.02592
500	**5**	**0.97127**	500	16	0.11349
	6	0.94193		17	0.07319
	7	0.89535		**18**	**0.04529**

TABLE 6. Continued

t	n	$P(H_1\|n,t)$	t	n	$P(H_1\|n,t)$
600	**7**	**0.96708**	600	19	0.10411
	8	0.93866		20	0.06882
	9	0.89562		**21**	**0.04388**

From the results of Table 6, at 400 patient-years, if $n \leq 3$, the trial is stopped, and the heart valve is marketed; furthermore, if $n \geq 15$, the trial is stopped, and valve is not marketed. On the other hand, if $4 \leq n \leq 14$, the trial continues until 500 patient-years are accrued. At 500 patient-years, if $n \leq 5$, the trial is stopped, and the heart valve is marketed; additionally, if $n \geq 18$, the trial is stopped, and the valve is not marketed. Nonetheless, if $6 \leq n \leq 17$, the trial continues until 600 patient-years are accrued. At 600 patient-years, if $n \leq 7$, the trial ends and the valve is marketed; moreover, if $n \geq 21$, the trial concludes, and the valve fails to be marketed. Otherwise, if $8 \leq n \leq 20$, the trial continues until 800 patient-years, when it is eventually stopped and the maximum-likelihood test is carried out.

TABLE 7. Results of Bayesian Monitoring in Poisson-Inverse Gamma Example

t	n	$P(H_1\|n,t)$	t	n	$P(H_1\|n,t)$
400	**1**	**0.96892**	400	15	0.06413
	2	0.94761		**16**	**0.03959**
	3	0.91598		17	0.02339
500	2	0.98170	500	18	0.06132
	3	**0.96853**		**19**	**0.03908**

	4	0.94831		20	0.02398
600	**5**	**0.96837**	600	21	0.05795
	6	0.94920		**22**	**0.03783**
	7	0.92178		23	0.02389

From the results of Table 7, at 400 patient-years, if $n \leq 1$, the trial is stopped, and valve is not marketed. On the other hand, if $n \geq 16$, the trial is stopped and the valve is marketed; however, if $2 \leq n \leq 15$, the trial continues until 500 patient-years are accrued. At 500 patient-years, if $n \leq 3$, the trial is stopped, and the heart valve is not marketed. Additionally, if $n \geq 19$, the trial is stopped and the valve is marketed; on the contrary, if $4 \leq n \leq 18$, the trial continues until 600 patient years are accrued. Lastly, at 600 patient-years, if $n \leq 5$, the trial ends and the valve is marketed; moreover, if $n \geq 22$, the trial concludes, and the valve fails to be marketed. Otherwise, if $6 \leq n \leq 21$, the trial continues until 800 patient-years, when it is eventually stopped and requires the maximum likelihood test to be carried out.

Comparing the stopping rules of the conjugate vs. nonconjugate priors, we observe that it is harder to stop the trial for superiority as well as inferiority of the tested valve for the nonconjugate prior as fewer adverse events are required to market the value and more adverse events are required to stop the trial and not market the valve.

4.4 Normal Inference

4.4.1. Normal-Normal Example

We illustrate the application of a normal prior density when the data are normally distributed. Suppose a clinical trial is conducted to determine if a new drug was efficient in lowering blood pressure for patients suffering from hypertension. To do so, the researchers first

specify the endpoint of the trial as the percentage reduction in diastolic blood pressure. The new drug is given to the treatment group, while the control group receives a placebo. The true mean percentage of the reduction in blood pressure is specified as μ_t and μ_c, in the treatment and control groups, respectively. The researchers test $H_0: \mu_t \leq \mu_c$ against $H_1: \mu_t > \mu_c$. If the alternative is accepted, it would indicate that the true mean percentage of the reduction in blood pressure is greater in the treatment group and thus the new drug is effective.

Let $X_t \sim N(\mu_t, \sigma^2)$ and $X_c \sim N(\mu_c, \sigma^2)$ be the random variables representing respectively blood pressure reduction in the treatment and control group patients. The variance σ^2 is assumed known (from previous studies). Suppose there are n patients in each group. The distribution of the difference in sample means is $\bar{X}_t - \bar{X}_c \sim N(\mu_t - \mu_c, 2\sigma^2/n)$. Next, we specify a conjugate normal prior for the difference in means as $\mu_t - \mu_c \sim N(\delta_0, \sigma_0^2)$ where the pdf is equal to $\pi(\delta) = (2\pi\sigma_0^2)^{-1/2}\exp\left(-\frac{(\delta-\delta_0)^2}{2\sigma_0^2}\right),\ -\infty < \delta < \infty$.

From this information, the posterior is derived as follows

$$f_{\mu_t-\mu_c}(\delta \mid n, \bar{x}_t, \bar{x}_c) \propto f(n, \bar{x}_t, \bar{x}_c \mid \delta)\pi(\delta)$$

$$= (2\pi)^{-1/2}(2\sigma^2/n)^{-1/2}\exp\left\{\frac{-(\bar{x}_t - \bar{x}_c - \delta)^2}{4\sigma^2/n}\right\} \cdot (2\pi\sigma_0^2)^{-1/2}\exp\left\{\frac{-(\delta-\delta_0)^2}{2\sigma_0^2}\right\}$$

$$\propto \exp\left\{-\frac{n\delta^2 - 2n(\bar{x}_t - \bar{x}_c)\delta}{4\sigma^2} - \frac{\delta^2 - 2\delta_0\delta}{2\sigma_0^2}\right\}$$

$$= \exp\left\{-\frac{1}{2} \cdot \frac{2\sigma^2 + n\sigma_0^2}{2\sigma^2\sigma_0^2} \cdot \left(\delta^2 - 2 \cdot \frac{n(\bar{x}_t - \bar{x}_c)\sigma_0^2 + 2\delta_0\sigma^2}{2\sigma^2 + n\sigma_0^2} \cdot \delta\right)\right\}$$

which has the algebraic form of a normal distribution with mean

$$\frac{n(\bar{x}_t - \bar{x}_c)\sigma_0^2 + 2\delta_0\sigma^2}{2\sigma^2 + n\sigma_0^2} = \left(\delta_0 \cdot \frac{1}{\sigma_0^2} + (\bar{x}_t - \bar{x}_c) \cdot \frac{n}{2\sigma^2}\right) / \left(\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}\right)$$

and variance

$$\left(\frac{2\sigma^2 + n\sigma_0^2}{2\sigma^2\sigma_0^2}\right)^{-1} = \left(\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}\right)^{-1}.$$

It remains to determine reasonable values of δ_0 and σ_0^2, the parameters of the prior distribution. These can be elicited by asking investigators explicitly what they think the most likely value of $\mu_t - \mu_c$ is. The investigators should also specify $P(H_1)$. From here, σ_0 is derived as the solution of the equation

$$P(H_1) = P(\mu_t - \mu_c > 0) = \int_0^\infty (2\pi\sigma_0^2)^{-1/2} \exp\left\{\frac{-(x-\delta_0)^2}{2\sigma_0^2}\right\} dx.$$

Making the substitution $z = (x - \delta_0)/\sigma_0$, we obtain

$$P(H_1) = \int_{-\delta_0/\sigma_0}^\infty (2\pi)^{-1/2} \exp\left(-\frac{z^2}{2}\right) dz = 1 - \Phi\left(-\frac{\delta_0}{\sigma_0}\right).$$

From here,

$$\sigma_0 = \frac{-\delta_0}{\Phi^{-1}\big(1 - P(H_1)\big)}.$$

Once $\sigma, P(H_1)$, and δ_0 are elicited, we compute σ_0, and then fix group size n and search for values of $\bar{x}_t - \bar{x}_c$ that make the posterior probability of H_1 either above 0.95 or below 0.05. The expression for the posterior probability is

$$P(H_1 | n, \bar{x}_t - \bar{x}_c) = P(\mu_t - \mu_c > 0 \mid n, \bar{x}_t - \bar{x}_c)$$

$$= P\left(Z > -\frac{n(\bar{x}_t - \bar{x}_c)\sigma_0^2 + 2\delta_0\sigma^2}{2\sigma^2 + n\sigma_0^2} \Big/ \sqrt{\left(\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}\right)^{-1}}\right)$$

$$= 1 - \Phi\left(\frac{-\left(\frac{n(\bar{x}_t - \bar{x}_c)}{2\sigma^2} + \frac{\delta_0}{\sigma_0^2}\right)}{\sqrt{\frac{1}{\sigma_0^2} + \frac{n}{2\sigma^2}}}\right).$$

Next, we present a numeric example. Suppose from similar studies done in the past, the standard deviation is estimated as $\sigma = 17.4$. Suppose that researchers have confidence in the tested product and use an optimistic prior with $P(H_1) = 0.8$, and approximate the most likely value of $\mu_t - \mu_c$ by $\delta_0 = 8$. Using this information, we calculate

$$\sigma_0 = \frac{-8}{\Phi^{-1}\big(1-(0.8)\big)} = 9.5.$$

Suppose the researchers decided to conduct an interim Bayesian analysis when $n = 30$ patients per group are accrued. The trial is stopped if the posterior probability of H_1 is either less than 0.05 or larger than 0.95. The stopping rules are summarized in Table 8 in Section 4.4.3.

4.4.2. Normal-Cauchy Example

For the same scenario, assume that the prior distribution does not follow a conjugate normal distribution but instead follows a nonconjugate Cauchy distribution. Thus we have that the distribution of the difference in sample means is $\bar{X}_t - \bar{X}_c \sim N(\mu_t - \mu_c, 2\sigma^2/n)$ and assume that the prior distribution of $\mu_t - \mu_c$ is $Cauchy(\delta_0, \sigma_o^2)$ with the probability density function $\pi(\delta) = \frac{1}{\pi\sigma_0\left[1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2\right]}$, $-\infty < \delta < \infty$. The posterior density is proportional to

$$f_{\mu_t-\mu_c}(\delta \mid n, \bar{x}_t, \bar{x}_c) \propto f(n, \bar{x}_t, \bar{x}_c \mid \delta)\pi(\delta)$$

$$= (2\pi)^{-1/2}(2\sigma^2/n)^{-1/2} \exp\left\{\frac{-(\bar{x}_t - \bar{x}_c - \delta)^2}{4\sigma^2/n}\right\} \cdot \frac{1}{\pi\sigma_0\left[1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2\right]}$$

$$\propto \frac{\exp\left\{\frac{-\big(\delta-(\bar{x}_t-\bar{x}_c)\big)^2}{\frac{4\sigma^2}{n}}\right\}}{1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2}.$$

The normalizing constant $\left(\int_{-\infty}^{\infty} \frac{\exp\left\{ \frac{-\left(\delta-(\bar{x}_t-\bar{x}_c)\right)^2}{\frac{4\sigma^2}{n}} \right\}}{1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2} d\delta \right)^{-1}$ has to be computed numerically.

The parameter δ_0 is specified as an a priori most likely value of $\mu_t - \mu_c$. In addition, the prior probability of accepting the alternative hypothesis $P(H_1)$ can also be elicited. From here, σ_0 is computed as the solution of the equation

$$P(H_1) = P(\mu_t - \mu_c > 0) = \int_0^{\infty} \frac{1}{\pi\sigma_0\left[1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2\right]}\, d\delta\,.$$

Making the substitution $z = \frac{\delta-\delta_0}{\sigma_0}$, we obtain

$$P(H_1) = \int_{-\frac{\delta_0}{\sigma_0}}^{\infty} \frac{1}{\pi(1+z^2)}\, dz = \frac{1}{\pi}\arctan(z)\,|_{z=-\frac{\delta_0}{\sigma_0}}^{z=\infty} = \frac{1}{\pi}\left(\frac{\pi}{2} - \arctan\left(-\frac{\delta_0}{\sigma_0}\right)\right).$$

Hence,

$$\sigma_0 = \frac{-\delta_0}{\tan\left(\frac{\pi}{2} - \pi\cdot P(H_1)\right)}.$$

The posterior probability of H_1 can be found numerically according to the formula

$$P(H_1|n,\bar{x}_t - \bar{x}_c) = P(\mu_t - \mu_c > 0 \mid n,\bar{x}_t - \bar{x}_c) = \frac{\int_0^{\infty} \frac{\exp\left\{ \frac{-\left(\delta-(\bar{x}_t-\bar{x}_c)\right)^2}{\frac{4\sigma^2}{n}} \right\}}{1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2} d\delta}{\int_{-\infty}^{\infty} \frac{\exp\left\{ \frac{-\left(\delta-(\bar{x}_t-\bar{x}_c)\right)^2}{\frac{4\sigma^2}{n}} \right\}}{1+\left(\frac{\delta-\delta_0}{\sigma_0}\right)^2} d\delta}.$$

In our numerical example with $P(H_1) = 0.8$ and $\delta_0 = 8$, we compute

$$\sigma_0 = \frac{-8}{\tan\left(\frac{\pi}{2} - \pi \cdot 0.8\right)} = 5.812.$$

The trial is stopped if the posterior probability of H_1 is either less than 0.05 or larger than 0.95. The stopping rules for $\sigma = 17.4$ and $n = 30$ are summarized in Table 8 in Section 4.4.3.

4.4.3. Normal Conjugate vs. Nonconjugate Stopping Results

For the numerical example with the conjugate prior, the stopping rules are given in Table 8 below.

TABLE 8. Results of Bayesian Monitoring in Normal-Normal Example

n	$\bar{x}_t - \bar{x}_c$	$P(H_1 \mid n, \bar{x}_t - \bar{x}_c)$	n	$\bar{x}_t - \bar{x}_c$	$P(H_1 \mid n, \bar{x}_t - \bar{x}_c)$
30	**-10.0**	**0.04918**	30	6.2	0.94600
	-9.9	0.05127		6.3	0.94817
	-9.8	0.05342		**6.4**	**0.95028**

The interpretation of the Bayesian monitoring procedure is as follows: when the sample size n has reached 30 in each group, if $\bar{x}_t - \bar{x}_c \leq -10$, then the trial is stopped, and the drug is not marketed due to its inability to reduce blood pressure. On the other hand, if $\bar{x}_t - \bar{x}_c \geq 6.4$, then the trial is stopped, and the drug is marketed. Otherwise, if $-10 < \bar{x}_t - \bar{x}_c < 6.4$, then the trial continues until the required sample size of 51 patients per group are accrued, at which point the trial is stopped and the standard z-test is carried out. The estimate of the required sample size of 51 patients per group for a non-Bayesian monitoring is computed via power analysis (see Section 4.1.3). The stopping rules for the Cauchy prior are given in Table 9.

TABLE 9. Results of Bayesian Monitoring in Normal-Cauchy Example

n	$\bar{x}_t - \bar{x}_c$	$P(H_1 \mid n, \bar{x}_t - \bar{x}_c)$	n	$\bar{x}_t - \bar{x}_c$	$P(H_1 \mid n, \bar{x}_t - \bar{x}_c)$
30	**-10.2**	**0.04872**	30	5.0	0.94974
	-10.1	0.05114		**5.1**	**0.95169**
	-10.0	0.05365		5.2	0.95357

The interpretation of the Bayesian monitoring procedure is as follows: when the sample size n has reached 30 in each group, if $\bar{x}_t - \bar{x}_c \leq -10.2$, then the trial is stopped, and the drug is not marketed; however, if $\bar{x}_t - \bar{x}_c \geq 5.1$, then the trial is stopped, and the drug is marketed. Otherwise, if $-10.1 \leq \bar{x}_t - \bar{x}_c \leq 5.0$, the continues until the required sample size of 51 patients per group are accrued, at which point the trial is stopped and the standard z test is carried out.

Comparing the results of the conjugate vs. nonconjugate priors, we conclude that it is harder to stop the trial earlier and not market the drug with nonconjugate prior (since the threshold is lower (-10.2 as opposed to -10.0 with the conjugate prior). However, it is easier to stop the trial and market the drug with nonconjugate prior since the acceptance value of 5.1 is lower than 6.4, which is required with the conjugate prior.

4.5 Binomial Inference

4.5.1 Binomial-Beta Example

A clinical trial is conducted in which N patients with heart arrhythmia are implanted defibrillators. The researchers involved with this study are interested in testing whether the chance of false positive alarms by the defibrillators within the first year of use is low. To do so, we specify $X \sim Bin(N, p)$ where X is the number of false positives and p is the probability of a false positive. The null and alternative hypotheses of interest are specified as

$$H_0: p \geq p_0 \text{ vs. } H_1: p < p_0 .$$

Since X has a binomial distribution, we can specify a $Beta(\alpha, \beta)$ prior on p, which is conjugate to the binomial distribution. The posterior distribution of p after observing the number of false alarms $X = x$ is derived as follows.

$$f_p(p \mid x) \propto P_X(x, p) \cdot \pi(p)$$

$$\binom{N}{x} p^x (1-p)^{N-x} \cdot \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} \propto \; p^{x+\alpha-1}(1-p)^{N-x+\beta-1}.$$

Thus, the posterior distribution is $Beta(x + \alpha, N - x + \beta)$.

The values of α and β may be determined from two equations, for the prior probability of H_1, and the most likely value of p (i.e., the mode of the prior distribution). The equations are:

$$P(H_1) = P(p < p_0) = \int_0^{p_0} \frac{p^{\alpha-1}(1-p)^{\beta-1}}{B(\alpha, \beta)} dp,$$

and

$$Mode = \frac{\alpha - 1}{\alpha + \beta - 2}$$

where $Mode < p_0$, to ensure a unique solution for the estimates of α and β.

To present a numeric example, suppose there are $N = 110$ patients and the probability of a false alarm $H_1: p < 0.25$ is being tested. Suppose the researchers use a skeptical prior with the probability of the true alternative equal to $P(H_1) = 0.40$, and the elicited value of the mode is 0.23. Solving numerically the two equations above, we get that the estimated values for α and β are 1.96 and 4.22, respectively. The posterior probability of the alternative is

$$P(H_1|x) = P(p < p_0|x) = \int_0^{p_0} \frac{p^{x+\alpha}(1-p)^{N-x+\beta-1}}{B(x + \alpha, N - x + \beta)} dp.$$

The trial is stopped if this probability is less than 0.05 or in excess of 0.95. The stopping rules are shown in Table 10 Section 4.5.3.

4.5.2 Binomial-Truncated Normal

Using the same information from the previous example, suppose the researchers choose to specify a nonconjugate truncated normal prior on p with the probability density function

$$\pi(p) = \frac{(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right)}{\int_0^1 (2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) dp}, \qquad 0 \le p \le 1.$$

The posterior distribution of p after observing the number of false alarms x is

$$f_p(p \mid x) = \frac{p^x(1-p)^{N-x} \cdot \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right)}{\int_0^1 p^x(1-p)^{N-x} \cdot \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) dp}$$

where the mean μ and standard deviation σ can be arbitrarily chosen.

The posterior probability of the alternative has the expression

$$P(H_1|x) = P(p < p_0 \,|x) = \frac{\int_0^{p_0} p^x(1-p)^{N-x} \cdot \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) dp}{\int_0^1 p^x(1-p)^{N-x} \cdot \exp\left(-\frac{(p-\mu)^2}{2\sigma^2}\right) dp}.$$

The trial is stopped if this probability is smaller than 0.05 or larger than 0.95. The stopping rules are shown in the next section in Table 11.

4.5.3. Binomial Conjugate vs. Nonconjugate Stopping Results

In the numerical example with the beta conjugate prior, the stopping rules are as given in Table 10 below.

TABLE 10. Results of Bayesian Monitoring in Binomial-Beta Example

x	$P(H_1\|x)$	x	$P(H_1\|x)$
19	**0.967199**	35	0.052273
20	0.946275	**36**	**0.038359**

If $x \leq 19$, the trial is stopped, indicating that the number of false positives is low, and thus the defibrillators can be sold to the public. On the contrary, if $x \geq 35$, the trial is stopped due to a high number of false positives, and thus the defibrillators are not marketed. Otherwise, if $20 \leq x \leq 35$, the trial continues.

For the nonconjugate truncated normal prior, the stopping rule is summarized in Table 11 that follows.

TABLE 11. Results of Bayesian Monitoring in Binomial-Truncated Normal Example

x	$P(H_1\|x)$	x	$P(H_1\|x)$
19	**0.953627**	34	0.054265
20	0.92602	**35**	**0.035085**

Remark: The calculations in table assumes the researchers arbitrary specify $\mu = \frac{1}{2}$ and $\sigma = \frac{1}{4}$ for the posterior distribution

We then give our interpretation of the results shown in Table 6. If $x \leq 19$, the trial is stopped, indicating that the number of false positives is low, and thus the defibrillators can be sold to the public. On the contrary, if $x \geq 35$, the trial is stopped due to a high number of false

positives, and thus the defibrillators are not marketed. Likewise, if $20 \leq x \leq 34$, the trial continues.

Comparing the results of the conjugate to that of the nonconjugate model, we observe equal lower bound of 19 false positives to stop the trial and market defibrillators. However, use of nonconjugate prior allows stopping the trial earlier and not marketing defibrillators if 35 false positive incidences are observed as opposed to 36 with conjugate prior, making it easier to stop the trial with nonconjugate prior. A possible explanation for this phenomenon is observe that with the posterior of the nonconjugate model, because we restricted the parameter p to the interval $[0,1]$ where μ and σ are arbitrary constants, the nonconjugate model yielded a posterior that behaved somewhat like a beta distribution. This is the case with the Binomial-Beta example shown above, as this example resulted in a beta posterior distribution. Therefore, we can conclude that both models performed almost equally well, but the null H_0 was easier to accept for the nonconjugate model.

CHAPTER 5

CONCLUSION

This thesis provided a broad overview of statistical techniques applicable to medical data, in survival analysis, longitudinal regression data analysis, and Bayesian monitoring of clinical trials. In Chapter 2, we modeled survival data using a variety of techniques (such as Kaplan-Meier survival curve, Cox proportional hazards model, Weibull regression, etc.).

In Chapter 3, we modeled longitudinal data using random slope and intercept models for longitudinal responses with normal, gamma, binary, and Poisson distributions on simulated datasets. Next, to test how well each model fits the data, we employed a goodness-of-fit deviance test and found out that all of our fitted models were definitely suitable for our data. We used the fitted models for interpretation of estimated significant regression coefficients and for prediction of response for a fixed values of predictor variables. From there, we alternatively fitted generalized estimating equations (GEE) models for each of our datasets, choosing the optimal one according to the QIC criterion. We also interpreted the estimated significant regression coefficients and used the fitted GEE models for prediction.

The focus in Chapter 4 was on providing an overview of interim data monitoring in clinical trials using the classical sequential testing approach and the Bayesian sequential procedure. The Bayesian procedure was illustrated with three clinical trial settings involving end-point variables with Poisson, normal, and binomial distributions. Conjugate as well as nonconjugate prior distributions of the parameters were analyzed and results compared. We saw through the numerical examples, that the use of conjugate priors achieves computationally easier and faster results, and moreover, the stopping rules are less strenuous than when nonconjugate priors are utilized.

Further work can be done on the examples mentioned throughout the thesis. Most notably in survival analysis, we can use the accelerated failure time model (AFT model) to serve as an alternative to the commonly used proportional hazards model, competing risk models, and general frailty models. We can possibly explore the use of some machine learning methods on survival data, such as decision trees, random forests, and so on.

APPENDICES

APPENDIX A

R CODE FOR SURVIVAL OUTPUT

```
#Survival analysis finalized codes
#primary_biliary_cirrhosis
#dataset source:
#https://www.kaggle.com/jixing475/mayo-clinic-primary-biliary-cirrhosis-
data?select=pbc.csv
library(survival)
library(readxl)
library(dplyr)
library(msm)#this package is used to fit a weibull/expontential survival fxn
library(flexsurv)
library(mice)
library(survMisc)
library(KMsurv)
library(SurvRegCensCov)
library(ggplot2)
library(pammtools)
library(survminer)

#first, "Status=2" refers to those who are dead, therefore we  should censor
the column
#It appears that row 313 and above are all empty, therefore we will delete
them
pbc <- read_excel("C:/Schoolwork_files/my excel
files/primary_biliary_cirrhosis.xlsx")
pbc<-pbc[!(pbc$trt=="NA"),]
pbc<-pbc[1:312,]
pbc$status[pbc$status==1]<-0
pbc$status[pbc$status==2]<-1


#We are looking at the time between registration(age) until event occurs
#**Using a Kaplan Meier Curve
event<-survfit(Surv(age,status)~1, conf.type="none", data=pbc)
summary(event)
surv_object<-Surv(time=pbc$age,event=pbc$status)
general_kaplan_meier <- survfit(surv_object ~ 1, data = pbc)

#GGPLOT of a general KM curve
ggsurvplot(general_kaplan_meier, data = pbc, xlab="Age", ylab="Survival
probability", legend.labs=c("Censoring"),palette=c("blue"), censor.size=5)
#Summary Stats for general KM
summary(general_kaplan_meier)

#K-M gender survival rate
gendervector<-table(pbc$sex)
gendervector[names(gendervector)=="m"] #36
gendervector[names(gendervector)=="f"]#276
gender.surv<-survfit(surv_object~sex,data=pbc)
summary(gender.surv)
ggsurvplot(gender.surv,legend.labs=c("Female(censored)","Male(censored)"),
xlab="Age", title="Kaplan Meier Survival Curve")
```

```
#K-M Drug survival curve
drug.surv<-survfit(surv_object~trt,data=pbc)
summary(drug.surv)
ggsurvplot(drug.surv,legend.labs=c("D-Penicillamine","Placebo"),xlab="Age",
palette=c("purple","orange"),title="Kaplan Meier Survival Curve")

#Running LR test to verify significance differences
survdiff(Surv(age,status)~sex,data=pbc) #not significant
survdiff(Surv(age,status)~trt,data=pbc) #Drug is somewhat significant

#General Nelson aalen
n.a<-survfit(coxph(Surv(age,status)~1, data=pbc), type="aalen")
summary(n.a) #Summary statistics for NA estimator

#Plotting a general NA estimator
ggsurvplot(n.a, data = pbc,xlab="Age", ylab="Survival
Probability",palette=c("black"),title="Nelson-Aalen Estimator",
legend.labs="Censored", censor.size=4)

#Nelson aalen comparing males vs females
data_genderM<-pbc[pbc$sex=="m",]
data_genderF<-pbc[pbc$sex=="f",]

#Summary statistics for NA stratified by male and female
n.a.m<-survfit(coxph(Surv(age,status)~1, data=data_genderM), type="aalen")
summary(n.a.m)

n.a.f<-survfit(coxph(Surv(age,status)~1, data=data_genderF), type="aalen")
summary(n.a.f)

#Plotting NA estimator for males/females
plot(n.a.m,col=c("blue"), xlab="Age", ylab="Survival Probability",
main="Nelson Aalen Survival Curves", conf.int=FALSE)
legend("bottomleft", c("M","F"), lty=1, col=c("blue","red"))
lines(n.a.f, col="red", conf.int=FALSE)

#Nelson aalen summary statistics comparing Placebo vs D-Penicillamine
data_controldrug<-pbc[pbc$trt==2,] #2 is placebo
data_treatmentdrug<-pbc[pbc$trt==1,] #1 is trt

n.a.control<-survfit(coxph(Surv(age,status)~1, data=data_controldrug),
type="aalen")
summary(n.a.control)

n.a.trt<-survfit(coxph(Surv(age,status)~1,data=data_treatmentdrug),
type="aalen")
summary(n.a.trt)

#Plotting NA estimator for Treatment vs control drug
```

```
plot(n.a.control,col=c("purple"), xlab="Age", ylab="Survival
Probability",main="Nelson Aalen Survival Curves",conf.int=FALSE)
legend("bottomleft", c("Placebo","D-Penicillamine"), lty=1,
col=c("purple","green"))
lines(n.a.trt,col="green", conf.int=FALSE)


#Fitting a parametric survival model for weibull & exp dist
wei_surv <- flexsurvreg(Surv(age, status)~1, data=pbc, dist="weibull")
plot(wei_surv, ci=FALSE, conf.int=FALSE, ylab="Survival Probability",
xlab="Age", main="Weibull estimator of the survival distribution")

#Try fitting a weibull regression model on  all variables
weibull_full_model<- survreg(Surv(time,status)~trt+ascites+hepato
+spiders+edema+age+sex+bili+chol+albumin+copper+alk.phos+ast+trig+platelet+pr
otime+stage, data=pbc,dist="weibull")
summary(weibull_full_model)

#checking model fit
weibull_intercept_model<-flexsurvreg(Surv(time, status)~1, data=pbc,
dist="weibull")
print(deviance<- -2*(logLik(weibull_intercept_model)-
logLik(weibull_full_model)))
print(p.value<- pchisq(deviance, df=8, lower.tail=FALSE)) # Significantly
less than 0.05

#reducing the model to those with P-values less than 5%
weibull_reduced<-survreg(Surv(time,status)~edema+age+bili+albumin+copper+
ast+protime+stage, data=pbc,dist="weibull")
summary(weibull_reduced)


#Cox Proportional hazard model on all variables;
cox=coxph(Surv(time,status)~trt+ascites+hepato+spiders+edema+age+sex
+bili+chol+albumin+copper+alk.phos+ast+trig+platelet+protime+stage,data=pbc)
summary(cox)
basehaz(cox, centered=TRUE)

#reducing the model to those with P-values less than 10%
library(rgl)
library(fields)
cox_reduced<-
coxph(Surv(time,status)~edema+age+bili+albumin+copper+ast+protime
+stage, data=pbc)
summary(cox_reduced)

#The basehaz code estimates the baseline hazard for various times for this
#reduced model
basehaz(cox_reduced, centered = TRUE)
```

```
#estimates the baseline survival function computed at the mean values of
#predictors
base.surv<-survfit(cox_reduced)
summary(base.surv)
#Estimating the sample means of reduced model significant predictors
#NOTE: there is a missing value in the copper column so that value is omitted
from calculations
summarise(pbc, edema = mean(edema),
          age = mean(age),
          bili = mean(bili),
          albumin = mean(albumin),
          copper = mean(copper,na.rm=TRUE),
          ast = mean(ast),
          protime=mean(protime),
          stage=mean(stage))
```

APPENDIX B

SURVIVAL OUTPUT FROM R

Table 1A. Log rank test results on gender survival

```
> survdiff(Surv(age,status)~sex,data=pbc) #not significant
Call:
survdiff(formula = Surv(age, status) ~ sex, data = pbc)

        N Observed Expected (O-E)^2/E (O-E)^2/V
sex=f 276      103     98.2     0.235      1.17
sex=m  36       22     26.8     0.861      1.17

 Chisq= 1.2  on 1 degrees of freedom, p= 0.3
```

Table 1B. Log rank test results on effectiveness of drugs

```
> survdiff(Surv(age,status)~trt,data=pbc) #Drug is somewhat significant
Call:
survdiff(formula = Surv(age, status) ~ trt, data = pbc)

        N Observed Expected (O-E)^2/E (O-E)^2/V
trt=1 158       65     74.5      1.20      3.03
trt=2 154       60     50.5      1.77      3.03

 Chisq= 3  on 1 degrees of freedom, p= 0.08
```

Table 2. Weibull output on all predictors

```
> summary(weibull_full_model)

Call:
survreg(formula = Surv(time, status) ~ trt + ascites + hepato +
    spiders + edema + age + sex + bili + chol + albumin + copper +
    alk.phos + ast + trig + platelet + protime + stage, data = pbc,
    dist = "weibull")
                Value Std. Error     z       p
(Intercept)  1.12e+01   1.21e+00  9.22 < 2e-16
trt          6.39e-02   1.30e-01  0.49  0.6242
ascites     -1.14e-01   2.32e-01 -0.49  0.6230
hepato      -1.79e-02   1.52e-01 -0.12  0.9060
spiders     -2.75e-02   1.49e-01 -0.19  0.8532
edema       -6.95e-01   2.35e-01 -2.96  0.0031
age         -1.80e-02   7.10e-03 -2.53  0.0114
sexm        -2.05e-01   1.93e-01 -1.07  0.2861
bili        -4.61e-02   1.46e-02 -3.15  0.0016
chol        -2.84e-04   2.71e-04 -1.05  0.2957
albumin      3.93e-01   1.82e-01  2.15  0.0314
copper      -1.56e-03   7.18e-04 -2.17  0.0297
alk.phos    -5.07e-06   2.42e-05 -0.21  0.8342
ast         -2.66e-03   1.19e-03 -2.23  0.0255
trig         5.28e-04   7.98e-04  0.66  0.5087
platelet    -5.20e-04   7.20e-04 -0.72  0.4700
protime     -1.52e-01   6.50e-02 -2.34  0.0193
stage       -2.76e-01   1.06e-01 -2.60  0.0092
Log(scale)  -4.95e-01   7.68e-02 -6.44 1.2e-10

Scale= 0.61

Weibull distribution
Loglik(model)= -967.4   Loglik(intercept only)= -1053.2
        Chisq= 171.68 on 17 degrees of freedom, p= 1.3e-27
Number of Newton-Raphson Iterations: 6
n=276 (36 observations deleted due to missingness)
```

Table 3. Reduced Weibull model on significant predictors

```
> summary(weibull_reduced)

Call:
survreg(formula = Surv(time, status) ~ edema + age + bili + albumin +
    copper + ast + protime + stage, data = pbc, dist = "weibull")
               Value Std. Error     z       p
(Intercept) 11.094503   0.999553 11.10 < 2e-16
edema       -0.562153   0.186190 -3.02 0.00253
age         -0.019662   0.005766 -3.41 0.00065
bili        -0.049793   0.010771 -4.62 3.8e-06
albumin      0.460951   0.152516  3.02 0.00251
copper      -0.001888   0.000559 -3.38 0.00073
ast         -0.002927   0.001021 -2.87 0.00416
protime     -0.176038   0.056421 -3.12 0.00181
stage       -0.248649   0.083250 -2.99 0.00282
Log(scale)  -0.487743   0.070695 -6.90 5.2e-12

Scale= 0.614
```

Table 4. Cox Proportional hazard model output

```
> summary(cox)
Call:
coxph(formula = Surv(time, status) ~ trt + ascites + hepato +
    spiders + edema + age + sex + bili + chol + albumin + copper +
    alk.phos + ast + trig + platelet + protime + stage, data = pbc)

  n= 276, number of events= 111
   (36 observations deleted due to missingness)

               coef  exp(coef)   se(coef)      z Pr(>|z|)
trt      -1.242e-01  8.832e-01  2.147e-01 -0.579  0.56290
ascites   8.833e-02  1.092e+00  3.872e-01  0.228  0.81955
hepato    2.552e-02  1.026e+00  2.510e-01  0.102  0.91900
spiders   1.012e-01  1.107e+00  2.435e-01  0.416  0.67760
edema     1.011e+00  2.749e+00  3.941e-01  2.566  0.01029 *
age       2.890e-02  1.029e+00  1.164e-02  2.482  0.01305 *
sexm      3.656e-01  1.441e+00  3.113e-01  1.174  0.24022
bili      8.001e-02  1.083e+00  2.550e-02  3.138  0.00170 **
chol      4.918e-04  1.000e+00  4.442e-04  1.107  0.26829
albumin  -7.408e-01  4.767e-01  3.078e-01 -2.407  0.01608 *
copper    2.490e-03  1.002e+00  1.170e-03  2.128  0.03337 *
alk.phos  1.048e-06  1.000e+00  3.969e-05  0.026  0.97893
ast       4.070e-03  1.004e+00  1.958e-03  2.078  0.03767 *
trig     -9.758e-04  9.990e-01  1.333e-03 -0.732  0.46414
platelet  9.019e-04  1.001e+00  1.184e-03  0.762  0.44629
protime   2.324e-01  1.262e+00  1.061e-01  2.190  0.02850 *
stage     4.545e-01  1.575e+00  1.754e-01  2.591  0.00958 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 5. Reduced Cox Proportional hazards model by 5% significance

```
> summary(cox_reduced)
Call:
coxph(formula = Surv(time, status) ~ edema + age + bili + albumin +
    copper + ast + protime + stage, data = pbc)

  n= 310, number of events= 124
   (2 observations deleted due to missingness)

              coef  exp(coef)   se(coef)      z Pr(>|z|)
edema     0.8322747  2.2985413  0.3107014  2.679  0.00739 **
age       0.0326821  1.0332220  0.0094460  3.460  0.00054 ***
bili      0.0849733  1.0886880  0.0186276  4.562 5.07e-06 ***
albumin -0.7879031  0.4547975  0.2551801 -3.088  0.00202 **
copper    0.0027640  1.0027678  0.0009231  2.994  0.00275 **
ast       0.0046992  1.0047103  0.0016574  2.835  0.00458 **
protime   0.2676418  1.3068790  0.0911334  2.937  0.00332 **
stage     0.4052377  1.4996589  0.1355840  2.989  0.00280 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


        exp(coef) exp(-coef) lower .95 upper .95
edema      2.2985     0.4351    1.2502    4.2259
age        1.0332     0.9678    1.0143    1.0525
bili       1.0887     0.9185    1.0497    1.1292
albumin    0.4548     2.1988    0.2758    0.7499
copper     1.0028     0.9972    1.0010    1.0046
ast        1.0047     0.9953    1.0015    1.0080
protime    1.3069     0.7652    1.0931    1.5625
stage      1.4997     0.6668    1.1497    1.9561

Concordance= 0.852  (se = 0.017 )
Likelihood ratio test= 194.1  on 8 df,   p=<2e-16
Wald test            = 205.5  on 8 df,   p=<2e-16
Score (logrank) test = 305.3  on 8 df,   p=<2e-16
```

Table 6. Estimates of sample means for significant Cox Predictors under the reduced model

```
> #Estimating the sample means of reduced model significant predictors
> #NOTE: there is a missing value in the copper column so that's excluded from calculations
> summarise(pbc, edema = mean(edema),
+           age = mean(age),
+           bili = mean(bili),
+           albumin = mean(albumin),
+           copper = mean(copper,na.rm=TRUE),
+           ast = mean(ast),
+           protime=mean(protime),
+           stage=mean(stage))
# A tibble: 1 x 8
  edema   age  bili albumin copper   ast protime stage
  <dbl> <dbl> <dbl>   <dbl>  <dbl> <dbl>   <dbl> <dbl>
1 0.111  50.0  3.26    3.52   97.6  123.    10.7  3.03
```

Table 7. Baseline Cox Survival Approximation at about t=1925

```
#estimates the baseline survival function computed at the mean values of predictors
base.surv<-survfit(cox_reduced)
summary(base.surv)
```

```
> #estimates the baseline survival function computed at the mean values of predictors
> base.surv<-survfit(cox_reduced)
> summary(base.surv)
Call: survfit(formula = cox_reduced)

 time n.risk n.event survival  std.err lower 95% CI upper 95% CI
   41    310       1    0.999 0.000652        0.998        1.000
   51    309       1    0.999 0.001095        0.996        1.000
   71    308       1    0.998 0.001444        0.995        1.000
   77    307       1    0.997 0.001746        0.993        1.000
  110    306       1    0.996 0.002029        0.992        1.000
  130    305       1    0.995 0.002292        0.991        1.000
  131    304       1    0.994 0.002546        0.989        0.999
  140    303       1    0.993 0.002839        0.988        0.999
```

1741	168	1	0.787	0.028496	0.733	0.844
1786	161	1	0.780	0.029041	0.725	0.839
1827	158	1	0.773	0.029586	0.718	0.834
1847	155	1	0.767	0.030132	0.710	0.828
1925	151	1	0.760	0.030684	0.702	0.822
2055	141	1	0.752	0.031333	0.693	0.816
2081	140	1	0.745	0.031946	0.685	0.810
2090	139	1	0.738	0.032539	0.676	0.804
2105	138	1	0.730	0.033104	0.668	0.798
2224	127	1	0.722	0.033732	0.659	0.792
2256	123	1	0.714	0.034345	0.650	0.785

APPENDIX C

LONGITUDINAL ANALYSIS R CODES

```
#Longitudial code for Normal,Gamma,Binary, & Poisson

library(readxl)
library(nlme)
library(MuMIn)
library(lme4)
library(dplyr)
library(geepack)
library(reshape2)
library(rcompanion)
######################## NORMAL RESPONSE ################
############################################################
#Blood pressure longitudinal study on simulated data

bp <- read_excel("C:/Schoolwork_files/my excel files/bp.xlsx")

#creating long-form data set
longform.data.bp<- melt(bp, id.vars=c("ID", "GENDER",
"ACTIVITY","SODIUM","HISTORY","CATEGORY"), variable.name="TIME",
value.name="Blood_Pressure")

#sorting long-form data set by id
longform.data.bp<- longform.data.bp[order(longform.data.bp$ID),]

#creating numeric variable for time
time.factor<- ifelse(longform.data.bp$TIME=="WEEK1",1,
                     ifelse(longform.data.bp$TIME=="WEEK2",2,
                            ifelse(longform.data.bp$TIME=="WEEK3",3,
                                   ifelse(longform.data.bp$TIME=="WEEK4",4,5
                                   ))))
#specifying reference categories
longform.data.bp$Blood_Pressure<-as.numeric(longform.data.bp$Blood_Pressure)
longform.data.bp$GENDER<-as.factor(longform.data.bp$GENDER)

#Converting to Factor
longform.data.bp$SODIUM<-as.factor(longform.data.bp$SODIUM)
longform.data.bp$HISTORY<-as.factor(longform.data.bp$HISTORY)
str(longform.data.bp)
history_factor<-relevel(longform.data.bp$HISTORY,ref="Y")
sex_factor<-relevel(longform.data.bp$GENDER,ref="F")
sodium_factor<-relevel(longform.data.bp$SODIUM,ref=3)


#plotting histogram with fitted normal density
hist(longform.data.bp$Blood_Pressure,main = "Histogram of Normal Response",
xlab="Syatolic Blood Pressure in mm/Hg")

#testing for normality of distribution
shapiro.test(longform.data.bp$Blood_Pressure)
```

```
#Fitting random s/i model
summary(fitted.model<- lme(Blood_Pressure ~ sex_factor+ACTIVITY+sodium_factor
+HISTORY+CATEGORY+time.factor, random=~ 1 +time.factor|ID,
control=list(opt="optim"),data=longform.data.bp))

intervals(fitted.model)

#checking model fit
null.model<- glm(Blood_Pressure ~ .,data=longform.data.bp)
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

print(p.value<- pchisq(deviance, df=3, lower.tail=FALSE))


#*****Fitting NORMAL GEE models
#fitting GEE model with autoregressive working correlation matrix
summary(ar1.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor
+ACTIVITY+sodium_factor+HISTORY+CATEGORY+time.factor, data=longform.data.bp,
id=ID, family=gaussian(link="identity"), corstr="ar1"))
QIC(ar1.fitted.model.normal) #93101.69

# #fitting GEE model with unstructured working correlation matrix
summary(uns.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor
+ACTIVITY+sodium_factor+HISTORY+CATEGORY+time.factor, data=longform.data.bp,
id=ID, family=gaussian(link="identity"), corstr="unstructured"))
QIC(uns.fitted.model.normal) #96540.92


#fitting GEE model with exchangeable working correlation matrix
summary(exch.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor
+ACTIVITY+sodium_factor+HISTORY+CATEGORY+time.factor, data=longform.data.bp,
id=ID, family=gaussian(link="identity"), corstr="exchangeable"))
QIC(exch.fitted.model.normal) #92964.38  ******Good fit****


#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor
+ACTIVITY+sodium_factor+HISTORY+CATEGORY+time.factor, data=longform.data.bp,
id=ID, family=gaussian(link="identity"), corstr="independence"))
QIC(ind.fitted.model.normal)#92981


#########################################################################

######################## GAMMA RESPONSE ################

#Cancer Longitudinal analysis
cancer <- read_excel("C:/Schoolwork_files/my excel files/cancer.xlsx")
```

```
###THE MISSING VALUES(DOTS) ARE CONVERTED TO NA***
cancer$WEEK6[cancer$WEEK6 == "."] <- NA
cancer<-na.omit(cancer)
str(cancer)
cancer$WEEK6<-as.numeric(cancer$WEEK6)

#creating long-form data set
longform.data.cancer<- melt(cancer, id.vars=c("ID", "TRT", "AGE","WEIGHT",
"STAGE","SEX"), variable.name="TIME", value.name="oral_cond")

#sorting long-form data set by id
longform.data.cancer<- longform.data.cancer[order(longform.data.cancer$ID),]

#creating numeric variable for time
time.factor<- ifelse(longform.data.cancer$TIME=="WEEK0", 0,
                     ifelse(longform.data.cancer$TIME=="WEEK2",

2,ifelse(longform.data.cancer$TIME=="WEEK4",4,6)))
#specifying reference categories
longform.data.cancer$TRT<-as.factor(longform.data.cancer$TRT)
longform.data.cancer$SEX<-as.factor(longform.data.cancer$SEX)
treatment_factor<- relevel(longform.data.cancer$TRT, ref="Cx")
sex_factor<-relevel(longform.data.cancer$SEX,ref="F")
str(longform.data.cancer)

#plotting histogram with fitted normal density
longform.data.cancer$oral_cond<-as.numeric(longform.data.cancer$oral_cond)
plotNormalHistogram(longform.data.cancer$oral_cond,xlab="Response",main="Gamm
a Response Distribution")

#testing for normality of distribution
shapiro.test(longform.data.cancer$oral_cond) #Significantly less than 0.005



#fitting gamma regression model with random slope and intercept
summary(gamma.fitted.model<- glmer(oral_cond ~ sex_factor+treatment_factor +
AGE+WEIGHT+STAGE+time.factor+(1 + time.factor| ID),
data=longform.data.cancer, family=Gamma(link='log')))

#checking model fit
null.model<- glm(oral_cond ~1,data=longform.data.cancer,
amily=Gamma(link='log'))
gamma.fitted.model.deviance.testing<-glm(oral_cond ~ sex_factor
+treatment_factor+AGE+ WEIGHT+STAGE+time.factor, data=longform.data.cancer,
family=Gamma(link='log'))
```

```
print(deviance<- -2*(logLik(null.model)-
logLik(gamma.fitted.model.deviance.testing)))
print(pvalue<- pchisq(deviance, df=7, lower.tail = FALSE)) #Fitted model
better


#fitting GEE model with autoregressive working correlation matrix
summary(ar1.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor
+treatment_factor+AGE+WEIGHT+STAGE+time.factor, data=longform.data.cancer,
id=ID, family=Gamma(link="log"), corstr="ar1"))
QIC(ar1.fitted.model.gamma) #1102.12

# #fitting GEE model with unstructured working correlation matrix
summary(uns.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor
+treatment_factor+AGE+WEIGHT+STAGE+time.factor, data=longform.data.cancer,
id=ID, family=Gamma(link="log"), corstr="unstructured"))
QIC(uns.fitted.model.gamma) #1097.33 ***BEST FIT***


#fitting GEE model with exchangeable working correlation matrix
summary(exch.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor
+treatment_factor+AGE+WEIGHT+STAGE+time.factor, data=longform.data.cancer,
id=ID, family=Gamma(link="log"), corstr="exchangeable"))
QIC(exch.fitted.model.gamma) #1101.11


#fitting GEE model with independent working correlation matrix
summary(ind.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor
+treatment_factor+AGE+WEIGHT+STAGE+time.factor, data=longform.data.cancer,
id=ID, family=Gamma(link="log"), corstr="independence"))
QIC(ind.fitted.model.gamma)#1111.5


##########################################################################

######################## BINARY RESPONSE ################
#Binary logistic longitidunal on anthrax *Fake(made up)* data
anthrax <- read_excel("C:/Schoolwork_files/my excel
files/anthrax.fake.data.xlsx")

#creating longform dataset
longform.datax<- melt(anthrax, id.vars=c("ID", "age","medicine","gender",
"risk","contacted"), variable.name="monthn",
value.name="remission_from_anthrax")
month<- ifelse(longform.datax$monthn=='month1',1,
               ifelse(longform.datax$monthn=='month2',2,
                      ifelse(longform.datax$monthn=='month3',3,
                             ifelse(longform.datax$monthn=='month4',4,
                                    ifelse(longform.datax$monthn=='month5',5,
```

```
ifelse(longform.datax$monthn=='month6',6,

ifelse(longform.datax$monthn=='month7',7,

ifelse(longform.datax$monthn=='month8',8,

ifelse(longform.datax$monthn=='month9',9,

ifelse(longform.datax$monthn=='month10',10,

ifelse(longform.datax$monthn=='month11',11,12))))))))))
#Specifying factors
longform.datax$contacted<-as.factor(longform.datax$contacted)
contact_factor<- relevel(longform.datax$contacted, ref="Y")
longform.datax$medicine<-as.factor(longform.datax$medicine)
longform.datax$gender<-as.factor(longform.datax$gender)

#fitting generalized random slope and intercept model, binary logistic
summary(fitted.model<- glmer(remission_from_anthrax ~ age+medicine+gender
+risk+contact_factor+month+(1+month|ID), data=longform.datax,
family=binomial(link='logit')))#
hist(longform.datax$remission_from_anthrax,main="Histogram of binary
response",xlab="Response")

#checking model fit by deviance test
fitted.model.deviance.testing<-glm(remission_from_anthrax ~ age+medicine+
gender+risk+contact_factor+ month, data=longform.datax,
family=binomial(link='logit'))

null.model<- glm(remission_from_anthrax ~ 1, data=longform.datax,
family=binomial(link='logit'))

print(deviance<- -2*(logLik(null.model)-
logLik(fitted.model.deviance.testing)))
print(p.value<- pchisq(deviance, df=7, lower.tail = FALSE))

###################################GEE#####################
#fitting GEE model with autoregressive working correlation matrix
summary(ar1.fitted.log.model<- geeglm(remission_from_anthrax ~ age+medicine
+gender+risk+contact_factor+month, data=longform.datax, id=ID,
family=binomial(link="logit"), corstr="ar1"))
QIC(ar1.fitted.log.model) #1441


#fitting GEE model with exchangeable working correlation matrix
summary(exch.fitted.log.model<- geeglm(remission_from_anthrax ~ age+medicine
+gender+risk+contact_factor+month, data=longform.datax, id=ID,
family=binomial(link="logit"), corstr="exchangeable"))
```

```
QIC(exch.fitted.log.model) #1441

#fitting GEE model with independent working correlation matrix
summary(ind.fitted.log.model<- geeglm(remission_from_anthrax ~ age+medicine +
gender+risk+contact_factor+month, data=longform.datax, id=ID,
family=binomial(link="logit"), corstr="independence"))
QIC(ind.fitted.log.model) #1441


# all 3 had the same QIC so it's up to the person to choose which one they
want

#############################################################################

######################### POISSON RESPONSE ################

#Cigerette medicine effectiveness  Longitudinal study *Fake* data
cig <- read_excel("C:/Schoolwork_files/my excel
files/cigarette.longitudinal.xlsx")

#creating long-form data set
longform.data.cig<- melt(cig, id.vars=c("ID", "SEX", "TRT","AGE","WEIGHIN",
"Intention","Addiction.Status"), variable.name="TIME",
value.name="N_CIGARETTES")

#sorting long-form data set by id
longform.data.cig<- longform.data.cig[order(longform.data.cig$ID),]

#specifying reference categories
longform.data.cig$TRT<-as.factor(longform.data.cig$TRT)
longform.data.cig$SEX<-as.factor(longform.data.cig$SEX)
treatment_factor<- relevel(longform.data.cig$TRT, ref="Cx")
sex_factor<-relevel(longform.data.cig$SEX,ref="F")

#creating numeric variable for time
time.factor<- ifelse(longform.data.cig$TIME=="Mo1",1,
                     ifelse(longform.data.cig$TIME=="Mo2",2,
                            ifelse(longform.data.cig$TIME=="Mo3",3,
                                   ifelse(longform.data.cig$TIME=="Mo4",4,

ifelse(longform.data.cig$TIME=="Mo5",5,6
                                                 )))))
#plotting histogram with fitted poisson response density
hist(longform.data.cig$N_CIGARETTES,main ="Histogram of Poisson Response",
xlab="Total cigarettes smoked")

#testing for normality of distribution
shapiro.test(longform.data.cig$N_CIGARETTES)
```

```
#Fitted  Poisson model
summary(fitted.model<- glm(N_CIGARETTES ~ treatment_factor+AGE
+WEIGHIN+Intention+Addiction.Status+time.factor+sex_factor,
family=poisson(link="log"), data=longform.data.cig))


#checking model fit
null.model<- glm(N_CIGARETTES ~ 1, data=longform.data.cig, family =
poisson(link = "log"))
print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))

print(p.value<- pchisq(deviance, df=7, lower.tail=FALSE))

#fitting random slope and intercept Poisson model
summary(fit.pois<-glmer(N_CIGARETTES ~ sex_factor+treatment_factor+AGE
+WEIGHIN+Intention+Addiction.Status+time.factor+ (1+time.factor|ID),
data=longform.data.cig, family=poisson(link="log")))


##################################GEE#############################
#fitting GEE model with autoregressive working correlation matrix

summary(ar1.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor
+treatment_factor+AGE+WEIGHIN+Intention+Addiction.Status+time.factor,
data=longform.data.cig, id=ID, family=poisson(link="log"), corstr="ar1"))
QIC(ar1.fitted.pois.model) #-25573.84 *** BEST FIT ***


#fitting GEE model with exchangeable working correlation matrix
summary(exch.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor
+treatment_factor+AGE+WEIGHIN+Intention+Addiction.Status+time.factor,
data=longform.data.cig, id=ID, family=poisson(link="log"),
corstr="exchangeable"))
QIC(exch.fitted.pois.model) #-25437.9

#fitting GEE model with independent working correlation matrix
summary(ind.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor
+treatment_factor+AGE+WEIGHIN+Intention+Addiction.Status+time.factor,
data=longform.data.cig, id=ID, family=poisson(link="log"),
corstr="independence"))
QIC(ind.fitted.pois.model) #-25552.6

#fitting GEE model with unstructured working correlation matrix
summary(uns.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor
+treatment_factor+AGE+WEIGHIN+Intention+Addiction.Status+time.factor,
data=longform.data.cig, id=ID, family=poisson(link="log"),
corstr="unstructured"))
QIC(uns.fitted.pois.model) #-25552.26
#################################
```

APPENDIX D

LONGITUDINAL OUTPUT AND DATASET SAMPLES

Table 1A. Shapiro Wilk normality test results

```
        Shapiro-Wilk normality test

data:  longform.data.bp$TOTALWEEK
W = 0.99024, p-value = 0.1338
```

Table 1B. Deviance testing results

```
> #checking model fit
> null.model<- glm(Blood_Pressure ~ .,data=longform.data.bp)
> print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
'log Lik.' 231 (df=13)
> print(p.value<- pchisq(deviance, df=3, lower.tail=FALSE))
'log Lik.' 1.03e-49 (df=13)
```

Table 2A. Normal Response random slope and intercept output

```
Linear mixed-effects model fit by REML
 Data: longform.data.bp
   AIC  BIC logLik
  1768 1809   -872

Random effects:
 Formula: ~1 + time.factor | ID
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 16.96  (Intr)
time.factor  9.15  -0.794
Residual     7.51

Fixed effects: Blood_Pressure ~ sex_factor + ACTIVITY + sodium_factor + HISTORY +      CATEGOR
or
               Value Std.Error  DF t-value p-value
(Intercept)    111.0      9.64 179   11.52  0.0000
sex_factorM     -5.8      3.96  38   -1.47  0.1505
ACTIVITY        -1.4      0.68  38   -2.08  0.0445
sodium_factor1 -20.4      8.03  38   -2.54  0.0152
sodium_factor2  -3.6      6.03  38   -0.60  0.5533
HISTORYY        -8.5      3.73  38   -2.29  0.0277
CATEGORY        18.8      2.16  38    8.72  0.0000
time.factor     -9.9      1.41 179   -7.04  0.0000
 Correlation:
               (Intr) sx_fcM ACTIVI sdm_f1 sdm_f2 HISTOR CATEGO
sex_factorM     0.142
ACTIVITY       -0.273  0.013
sodium_factor1 -0.795 -0.316 -0.083
sodium_factor2 -0.638 -0.021 -0.246  0.697
HISTORYY       -0.156 -0.270  0.125  0.082 -0.194
CATEGORY       -0.831 -0.325 -0.116  0.842  0.694 -0.008
time.factor    -0.229  0.000  0.000  0.000  0.000  0.000  0.000

Standardized Within-Group Residuals:
      Min        Q1       Med        Q3       Max
-2.156549 -0.470510  0.000273  0.464679  2.123249

Number of Observations: 225
Number of Groups: 45
```

Table 2B. Normal response GEE model with AR1 correlation matrix

```
> #*****Fitting NORMAL GEE models
> #fitting GEE model with autoregressive working correlation matrix
> summary(ar1.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor+ACTIVITY+sodium_factor+HISTORY+CATE
ORY
+                                          +time.factor,
+                                          data=longform.data.bp,
+                                          id=ID, family=gaussian(link="identity"),
+                                          corstr="ar1"))

Call:
geeglm(formula = Blood_Pressure ~ sex_factor + ACTIVITY + sodium_factor +
    HISTORY + CATEGORY + time.factor, family = gaussian(link = "identity"),
    data = longform.data.bp, id = ID, corstr = "ar1")

 Coefficients:
              Estimate Std.err   Wald Pr(>|W|)
(Intercept)     119.586  10.168 138.32  < 2e-16 ***
sex_factorM      -5.720   5.228   1.20    0.274
ACTIVITY         -1.009   0.807   1.56    0.211
sodium_factor1  -11.785   9.983   1.39    0.238
sodium_factor2   -1.817   6.308   0.08    0.773
HISTORYY        -11.022   4.890   5.08    0.024 *
CATEGORY         14.583   2.303  40.10  2.4e-10 ***
time.factor     -10.004   1.347  55.19  1.1e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)      414      57
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.574   0.074
Number of clusters:   45  Maximum cluster size: 5
```

```
> QIC(ar1.fitted.model.normal) #93101.69
       QIC       QICu Quasi Lik       CIC    params      QICC
 93101.69  93100.18 -46542.09      8.75      8.00  93102.53
> |
```

Table 2C. Normal response GEE model with unstructured working correlation matrix

```
Call:
geeglm(formula = Blood_Pressure ~ sex_factor + ACTIVITY + sodium_factor +
    HISTORY + CATEGORY + time.factor, family = gaussian(link = "identity"),
    data = longform.data.bp, id = ID, corstr = "unstructured")

 Coefficients:
               Estimate Std.err   Wald Pr(>|W|)
(Intercept)     116.471   8.027 210.55  < 2e-16 ***
sex_factorM      -5.981   4.042   2.19    0.139
ACTIVITY         -1.224   0.631   3.77    0.052 .
sodium_factor1  -16.123   7.866   4.20    0.040 *
sodium_factor2   -2.886   4.917   0.34    0.557
HISTORYY         -9.767   3.981   6.02    0.014 *
CATEGORY         16.491   1.787  85.14  < 2e-16 ***
time.factor     -10.151   1.330  58.24  2.3e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = unstructured
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)      429    60.1
  Link = identity

Estimated Correlation Parameters:
           Estimate Std.err
alpha.1:2  0.301593  0.0577
alpha.1:3  0.105241  0.0528
alpha.1:4  0.000876  0.0938
alpha.1:5 -0.095878  0.1367
alpha.2:3  0.282542  0.0993
alpha.2:4  0.212741  0.1245
alpha.2:5  0.121454  0.1701
alpha.3:4  0.594050  0.1026
alpha.3:5  0.753654  0.1243
alpha.4:5  1.641975  0.0717
Number of clusters:   45  Maximum cluster size: 5
>

> QIC(uns.fitted.model.normal) #96540.92
      QIC       QICu Quasi Lik       CIC    params      QICC
 96540.92  96545.94 -48264.97      5.49      8.00  96544.24
>
```

Table 2D. Normal response GEE model with exchangeable working correlation matrix

```
Call:
geeglm(formula = Blood_Pressure ~ sex_factor + ACTIVITY + sodium_factor +
    HISTORY + CATEGORY + time.factor, family = gaussian(link = "identity"),
    data = longform.data.bp, id = ID, corstr = "exchangeable")

 Coefficients:
               Estimate Std.err   Wald Pr(>|W|)
(Intercept)      120.18    9.62 155.90  < 2e-16 ***
sex_factorM       -6.38    5.03   1.61    0.204
ACTIVITY          -1.09    0.79   1.91    0.167
sodium_factor1   -13.21    9.38   1.98    0.159
sodium_factor2    -2.54    5.93   0.18    0.668
HISTORYY         -10.64    4.81   4.90    0.027 *
CATEGORY          14.61    2.14  46.53  9.0e-12 ***
time.factor       -9.92    1.39  50.69  1.1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)      413    56.9
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.369  0.0776
Number of clusters:   45  Maximum cluster size: 5
> QIC(exch.fitted.model.normal) #92964.38  ******Good fit****
      QIC      QICu Quasi Lik       CIC    params      QICC
 92964.38  92960.88 -46472.44      9.75      8.00  92965.21
> |
```

Table 2E. Normal response GEE model with independent working correlation matrix

```
> #fitting GEE model with independent working correlation matrix
> summary(ind.fitted.model.normal<- geeglm(Blood_Pressure ~ sex_factor+ACTIVITY+soc
ORY
+                                          +time.factor,
+                                          data=longform.data.bp,
+                                          id=ID, family=gaussian(link="identity"),
+                                          corstr="independence"))

Call:
geeglm(formula = Blood_Pressure ~ sex_factor + ACTIVITY + sodium_factor +
    HISTORY + CATEGORY + time.factor, family = gaussian(link = "identity"),
    data = longform.data.bp, id = ID, corstr = "independence")

 Coefficients:
               Estimate Std.err   Wald Pr(>|W|)
(Intercept)      120.18    9.62 155.90  < 2e-16 ***
sex_factorM       -6.38    5.03   1.61    0.204
ACTIVITY          -1.09    0.79   1.91    0.167
sodium_factor1   -13.21    9.38   1.98    0.159
sodium_factor2    -2.54    5.93   0.18    0.668
HISTORYY         -10.64    4.81   4.90    0.027 *
CATEGORY          14.61    2.14  46.53  9.0e-12 ***
time.factor       -9.92    1.39  50.69  1.1e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)      413    56.9
Number of clusters:   45  Maximum cluster size: 5
> QIC(ind.fitted.model.normal)#92981
      QIC      QICu Quasi Lik       CIC    params      QICC
    92981     92961    -46472        18         8     92981
> |
```

Table 3A. Normality test for gamma response

```
        Shapiro-Wilk normality test

data:  longform.data.cancer$oral_cond
W = 0.9, p-value <2e-16
```

Table 3B. Deviance testing results for gamma response

```
> #checking model fit
> null.model<- glm(oral_cond ~1,data=longform.data.cancer, family=Gamma(link='log'))
> gamma.fitted.model.deviance.testing<-glm(oral_cond ~ sex_factor+treatment_factor + AGE
+                                          + WEIGHT+STAGE+time.factor,
+                                            data=longform.data.cancer,
+                                          family=Gamma(link='log'))
> print(deviance<- -2*(logLik(null.model)-logLik(gamma.fitted.model.deviance.testing)))
'log Lik.' 294 (df=2)
> print(pvalue<- pchisq(deviance, df=7, lower.tail = FALSE)) #Fitted model better
'log Lik.' 1.16e-59 (df=2)
>
```

Table 4A. Random Slope and intercept Gamma output

```
> #fitting gamma regression model with random slope and intercept
> summary(gamma.fitted.model<- glmer(oral_cond ~ sex_factor+treatment_factor + AGE
+                                    + WEIGHT+STAGE+time.factor+
+                                      (1 + time.factor| ID), data=longform.data.cancer,
+                                    family=Gamma(link='log')))
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) ['glmerMod']
 Family: Gamma  ( log )
Formula: oral_cond ~ sex_factor + treatment_factor + AGE + WEIGHT + STAGE +
    time.factor + (1 + time.factor | ID)
   Data: longform.data.cancer

     AIC      BIC   logLik deviance df.resid
    1139     1183     -558     1117      389

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.8233 -0.3158 -0.0459  0.2862  2.5815

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 ID       (Intercept) 0.06158  0.2482
          time.factor 0.00525  0.0725   -0.74
 Residual             0.01557  0.1248
Number of obs: 400, groups:  ID, 100

Fixed effects:
                   Estimate Std. Error t value Pr(>|z|)
(Intercept)         1.643049   0.198070    8.30  < 2e-16 ***
sex_factorM        -0.123060   0.063315   -1.94    0.052 .
treatment_factorTx  0.322459   0.156171    2.06    0.039 *
AGE                -0.001522   0.001824   -0.83    0.404
WEIGHT              0.000130   0.000655    0.20    0.843
STAGE               0.044580   0.029981    1.49    0.137
time.factor         0.083266   0.012931    6.44  1.2e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Correlation of Fixed Effects:
            (Intr) sx_fcM trtm_T AGE    WEIGHT STAGE
sex_factorM -0.172
trtmnt_fctT -0.512 -0.025
AGE         -0.485  0.272 -0.005
WEIGHT      -0.673 -0.116  0.289 -0.080
STAGE       -0.113 -0.093 -0.227  0.004 -0.158
time.factor -0.175  0.003  0.068  0.002  0.017 -0.019
convergence code: 0



> QIC(ar1.fitted.model.gamma) #1102.12
      QIC      QICu Quasi Lik       CIC    params      QICC
  1102.12   1101.25   -543.62      7.44      7.00   1102.49
```

Table 4B. Gamma GEE model unstructured working correlation matrix

```
> # #fitting GEE model with unstructured working correlation matrix
> summary(uns.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor+treatment_factor + AGE +
+                                           WEIGHT+STAGE+time.factor,
+                                         data=longform.data.cancer,
+                                         id=ID, family=Gamma(link="log"),
+                                         corstr="unstructured"))

Call:
geeglm(formula = oral_cond ~ sex_factor + treatment_factor +
    AGE + WEIGHT + STAGE + time.factor, family = Gamma(link = "log"),
    data = longform.data.cancer, id = ID, corstr = "unstructured")

 Coefficients:
                   Estimate   Std.err   Wald Pr(>|W|)
(Intercept)        1.515796  0.138393 119.96   <2e-16 ***
sex_factorM       -0.125978  0.050097   6.32    0.012 *
treatment_factorTx 0.663875  0.050970 169.65   <2e-16 ***
AGE               -0.001362  0.001434   0.90    0.342
WEIGHT             0.000680  0.000532   1.63    0.201
STAGE              0.019892  0.024285   0.67    0.413
time.factor        0.068572  0.007916  75.04   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = unstructured
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.102 0.00871
  Link = identity

Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2   1.0973  0.0804
alpha.1:3   0.3948  0.0728
alpha.1:4  -0.0734  0.0839
alpha.2:3   0.5157  0.0574
alpha.2:4   0.2773  0.0611
alpha.3:4   0.5379  0.0466
Number of clusters:   100  Maximum cluster size: 4
```

```
> QIC(uns.fitted.model.gamma) #1097.33 ***BEST FIT***
      QIC      QICu Quasi Lik       CIC    params      QICC
  1097.33   1099.21   -542.61      6.06      7.00   1098.28
> |
```

Table 4C. Gamma GEE model with exchangeable working correlation matrix

```
> #fitting GEE model with exchangeable working correlation matrix
> summary(exch.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor+treatment_factor + AGE +
+                                            WEIGHT+STAGE+time.factor,
+                                          data=longform.data.cancer, id=ID,
+                                          family=Gamma(link="log"),
+                                          corstr="exchangeable"))

Call:
geeglm(formula = oral_cond ~ sex_factor + treatment_factor +
    AGE + WEIGHT + STAGE + time.factor, family = Gamma(link = "log"),
    data = longform.data.cancer, id = ID, corstr = "exchangeable")

 Coefficients:
                    Estimate   Std.err   Wald Pr(>|W|)
(Intercept)         1.572971  0.121377 167.95   <2e-16 ***
sex_factorM        -0.122118  0.043442   7.90   0.0049 **
treatment_factorTx  0.459249  0.043778 110.05   <2e-16 ***
AGE                -0.001196  0.001259   0.90   0.3423
WEIGHT              0.000516  0.000475   1.18   0.2772
STAGE               0.029020  0.020623   1.98   0.1594
time.factor         0.080849  0.009158  77.95   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)   0.0942 0.00873
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.356  0.0532
Number of clusters:   100  Maximum cluster size: 4
> QIC(exch.fitted.model.gamma) #1101.11
      QIC      QICu Quasi Lik       CIC    params      QICC
  1101.11   1098.03   -542.02      8.54      7.00   1101.48
> |
```

.

Table 4D. Gamma GEE model with independent working correlation matrix

```
> #fitting GEE model with independent working correlation matrix
> summary(ind.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor+treatment_factor + AGE +
+                                           WEIGHT+STAGE+time.factor,
+                                         data=longform.data.cancer, id=ID,
+                                         family=Gamma(link="log"),
+                                         corstr="independence"))

Call:
geeglm(formula = oral_cond ~ sex_factor + treatment_factor +
    AGE + WEIGHT + STAGE + time.factor, family = Gamma(link = "log"),
    data = longform.data.cancer, id = ID, corstr = "independence")

 Coefficients:
                   Estimate   Std.err   Wald Pr(>|W|)
(Intercept)        1.572973  0.121377 167.95   <2e-16 ***
sex_factorM       -0.122118  0.043442   7.90   0.0049 **
treatment_factorTx 0.459247  0.043778 110.05   <2e-16 ***
AGE               -0.001196  0.001259   0.90   0.3423
WEIGHT             0.000516  0.000475   1.18   0.2772
STAGE              0.029020  0.020623   1.98   0.1594
time.factor        0.080849  0.009158  77.94   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)   0.0942 0.00873
Number of clusters:   100  Maximum cluster size: 4
> QIC(ind.fitted.model.gamma)#1111.5
      QIC      QICu Quasi Lik       CIC    params      QICC
   1111.5    1098.0    -542.0      13.7       7.0    1111.8
```

Table 4E. Gamma GEE model with Autoregressive working correlation matrix

```
> #fitting GEE model with autoregressive working correlation matrix
> summary(ar1.fitted.model.gamma<- geeglm(oral_cond ~ sex_factor+treatment_factor + AGE +
+                                           WEIGHT+STAGE+time.factor,
+                                         data=longform.data.cancer,
+                                         id=ID, family=Gamma(link="log"),
+                                         corstr="ar1"))

Call:
geeglm(formula = oral_cond ~ sex_factor + treatment_factor +
    AGE + WEIGHT + STAGE + time.factor, family = Gamma(link = "log"),
    data = longform.data.cancer, id = ID, corstr = "ar1")

 Coefficients:
                   Estimate   Std.err   Wald Pr(>|W|)
(Intercept)        1.578965  0.119159 175.59   <2e-16 ***
sex_factorM       -0.122954  0.042772   8.26    0.004 **
treatment_factorTx  0.443266  0.042878 106.87   <2e-16 ***
AGE               -0.001201  0.001240   0.94    0.333
WEIGHT             0.000495  0.000467   1.12    0.289
STAGE              0.030782  0.020160   2.33    0.127
time.factor        0.080929  0.009143  78.35   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)   0.0948 0.00898
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.515  0.0538
Number of clusters:   100  Maximum cluster size: 4
> QIC(ar1.fitted.model.gamma) #1102.12
      QIC      QICu Quasi Lik       CIC    params      QICC
  1102.12   1101.25   -543.62      7.44      7.00   1102.49
>
```

Table 5. Deviance test results for Binary response

```
> #checking model fit by deviance test
> fitted.model.deviance.testing<-glm(remission_from_anthrax ~ age + medicine + gender
+                                    +risk+contact_factor+ month,
+                                    data=longform.datax, family=binomial(link='logit'))
> null.model<- glm(remission_from_anthrax ~ 1,
+                  data=longform.datax,
+                  family=binomial(link='logit'))
> print(deviance<- -2*(logLik(null.model)-logLik(fitted.model.deviance.testing)))
'log Lik.' 176 (df=1)
> print(p.value<- pchisq(deviance, df=7, lower.tail = FALSE))
'log Lik.' 1.44e-34 (df=1)
```

Table 6A. Binary random slope and intercept output

```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [
glmerMod]
 Family: binomial  ( logit )
Formula: sideeffects ~ age + medicine + gender + risk + contact_factor +
    month + (1 + month | ID)
   Data: longform.datax

     AIC      BIC   logLik deviance df.resid
    1408     1458     -694     1388     1190

Scaled residuals:
   Min     1Q Median     3Q    Max
-2.019 -0.791 -0.295  0.848  2.319

Random effects:
 Groups Name        Variance Std.Dev. Corr
 ID     (Intercept) 2.6907   1.64
        month       0.0529   0.23     -1.00
Number of obs: 1200, groups:  ID, 100

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)      1.875322   0.381700    4.91  9.0e-07 ***
age             -0.000331   0.004713   -0.07    0.944
medicineTx      -1.660902   0.212192   -7.83  5.0e-15 ***
genderM          0.028222   0.144485    0.20    0.845
risk            -0.055916   0.051128   -1.09    0.274
contact_factorN -0.309017   0.152663   -2.02    0.043 *
month           -0.210731   0.033079   -6.37  1.9e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) age    mdcnTx gendrM risk   cntc_N
age         -0.567
medicineTx  -0.405  0.125
genderM     -0.242  0.102 -0.053
risk        -0.328 -0.160  0.114 -0.074
cntct_fctrN -0.044 -0.187 -0.228  0.184  0.001
month       -0.606  0.000  0.167  0.010  0.029  0.013
convergence code: 0
boundary (singular) fit: see ?isSingular
```

Table 6B. Binary GEE model with AR1 correlation matrix

```
> summary(ar1.fitted.log.model<- geeglm(sideeffects ~ age + medicine + gender
+                                       +risk+contact_factor+ month ,
+                                        data=longform.datax, id=ID,
+                                       family=binomial(link="logit"), corstr="ar1"))

Call:
geeglm(formula = sideeffects ~ age + medicine + gender + risk +
    contact_factor + month, family = binomial(link = "logit"),
    data = longform.datax, id = ID, corstr = "ar1")

 Coefficients:
                 Estimate   Std.err  Wald Pr(>|W|)
(Intercept)      1.403151  0.299221 21.99  2.7e-06 ***
age              0.000774  0.004209  0.03   0.8542
medicineTx      -1.227180  0.125915 94.99  < 2e-16 ***
genderM          0.027827  0.132686  0.04   0.8339
risk            -0.040904  0.046234  0.78   0.3763
contact_factorN -0.383146  0.132860  8.32   0.0039 **
month           -0.163400  0.019103 73.16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.946  0.0125
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha        0       0
Number of clusters:   1200  Maximum cluster size: 1
> QIC(ar1.fitted.log.model) #1441
 QIC
1441
```

Table 6C. Binary GEE model with exchangeable working correlation matrix

```
> #fitting GEE model with exchangeable working correlation matrix
> summary(exch.fitted.log.model<- geeglm(sideeffects ~ age + medicine + gender
+                                        +risk+contact_factor+ month ,
+                                        data=longform.datax, id=ID,
+                                        family=binomial(link="logit"), corstr="exchange
able"))

Call:
geeglm(formula = sideeffects ~ age + medicine + gender + risk +
    contact_factor + month, family = binomial(link = "logit"),
    data = longform.datax, id = ID, corstr = "exchangeable")

 Coefficients:
                 Estimate   Std.err  Wald Pr(>|W|)
(Intercept)      1.403151  0.299221 21.99  2.7e-06 ***
age              0.000774  0.004209  0.03   0.8542
medicineTx      -1.227180  0.125915 94.99  < 2e-16 ***
genderM          0.027827  0.132686  0.04   0.8339
risk            -0.040904  0.046234  0.78   0.3763
contact_factorN -0.383146  0.132860  8.32   0.0039 **
month           -0.163400  0.019103 73.16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.946  0.0125
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha        0       0
Number of clusters:   1200  Maximum cluster size: 1
> QIC(exch.fitted.log.model) #1441
 QIC
1441
```

Table 6D. Binary GEE model with independent working correlation matrix

```
> #fitting GEE model with independent working correlation matrix
> summary(ind.fitted.log.model<- geeglm(sideeffects ~ age + medicine + gender
+                                       +risk+contact_factor+ month ,
+                                       data=longform.datax, id=ID,
+                                       family=binomial(link="logit"), corstr="independ
ence"))

Call:
geeglm(formula = sideeffects ~ age + medicine + gender + risk +
    contact_factor + month, family = binomial(link = "logit"),
    data = longform.datax, id = ID, corstr = "independence")

 Coefficients:
                 Estimate   Std.err  Wald Pr(>|W|)
(Intercept)      1.403151  0.299221 21.99  2.7e-06 ***
age              0.000774  0.004209  0.03   0.8542
medicineTx      -1.227180  0.125915 94.99  < 2e-16 ***
genderM          0.027827  0.132686  0.04   0.8339
risk            -0.040904  0.046234  0.78   0.3763
contact_factorN -0.383146  0.132860  8.32   0.0039 **
month           -0.163400  0.019103 73.16  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)    0.946  0.0125
Number of clusters:   1200  Maximum cluster size: 1
> QIC(ind.fitted.log.model) #1441
 QIC
1441
```

Table 7. Deviance testing for Poisson response

```
> #checking model fit
> null.model<- glm(N_CIGARETTES ~ 1,data=longform.data.cig,family = poisson(link = "log"))
> print(deviance<- -2*(logLik(null.model)-logLik(fitted.model)))
'log Lik.' 3822 (df=1)
> print(p.value<- pchisq(deviance, df=7, lower.tail=FALSE))
'log Lik.' 0 (df=1)
```

Table 8A. Random slope and intercept Poisson output

```
Fixed effects:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         0.969602   0.229453    4.23  2.4e-05 ***
sex_factorM         0.057573   0.075599    0.76    0.446
treatment_factorTx  0.246997   0.121154    2.04    0.041 *
AGE                -0.000914   0.002058   -0.44    0.657
WEIGHIN             0.000294   0.000674    0.44    0.663
Intention           0.090183   0.078172    1.15    0.249
Addiction.Status    0.508250   0.028599   17.77  < 2e-16 ***
time.factor        -0.230138   0.026989   -8.53  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 8B. Poisson GEE model with AR1

```
> summary(ar1.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor+treatment_factor+AGE
+                                        +WEIGHIN+Intention+Addiction.Status+time.factor,
+                                       data=longform.data.cig, id=ID, family=poisson(link="log"), corstr
="ar1"))

Call:
geeglm(formula = N_CIGARETTES ~ sex_factor + treatment_factor +
    AGE + WEIGHIN + Intention + Addiction.Status + time.factor,
    family = poisson(link = "log"), data = longform.data.cig,
    id = ID, corstr = "ar1")

 Coefficients:
                    Estimate   Std.err   Wald Pr(>|W|)
(Intercept)         1.351317  0.194269  48.38  3.5e-12 ***
sex_factorM         0.108062  0.062540   2.99    0.084 .
treatment_factorTx -0.419731  0.071240  34.71  3.8e-09 ***
AGE                -0.001014  0.001818   0.31    0.577
WEIGHIN            -0.000217  0.000543   0.16    0.689
Intention          -0.026741  0.067455   0.16    0.692
Addiction.Status    0.485724  0.027584 310.07  < 2e-16 ***
time.factor        -0.121304  0.021708  31.22  2.3e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = ar1
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     2.65   0.242
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.635  0.0409
Number of clusters:   100  Maximum cluster size: 6
> QIC(ar1.fitted.pois.model) #-25573.84 *** BEST FIT ***
      QIC      QICu Quasi Lik       CIC    params      QICC
-25573.84 -25577.27  12796.64      9.72      8.00 -25573.53
```

Table 8C. Poisson GEE model with exchangeable working correlation matrix

```
> #fitting GEE model with exchangeable working correlation matrix
> summary(exch.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor+treatment_factor+AGE
+                                         +WEIGHIN+Intention+Addiction.Status+time.factor,
+                                         data=longform.data.cig, id=ID, family=poisson(link="log"),
="exchangeable"))

Call:
geeglm(formula = N_CIGARETTES ~ sex_factor + treatment_factor +
    AGE + WEIGHIN + Intention + Addiction.Status + time.factor,
    family = poisson(link = "log"), data = longform.data.cig,
    id = ID, corstr = "exchangeable")

 Coefficients:
                   Estimate   Std.err   Wald Pr(>|W|)
(Intercept)        1.302276  0.198696  42.96  5.6e-11 ***
sex_factorM        0.081592  0.062390   1.71   0.1909
treatment_factorTx -0.218584  0.066806  10.71   0.0011 **
AGE                -0.000817  0.001927   0.18   0.6716
WEIGHIN            -0.000147  0.000567   0.07   0.7956
Intention           0.006349  0.065011   0.01   0.9222
Addiction.Status    0.472739  0.027369 298.34  < 2e-16 ***
time.factor        -0.121903  0.021956  30.83  2.8e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = exchangeable
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     2.77   0.238
  Link = identity

Estimated Correlation Parameters:
      Estimate Std.err
alpha    0.465  0.0372
Number of clusters:   100  Maximum cluster size: 6
> QIC(exch.fitted.pois.model) #-25437.9
      QIC      QICu Quasi Lik       CIC    params      QICC
 -25437.9  -25446.3   12731.2      12.2       8.0  -25437.6
> |
```

Table 8D. Poisson GEE model with independent working correlation matrix

```
> #fitting GEE model with independent working correlation matrix
> summary(ind.fitted.pois.model<- geeglm(N_CIGARETTES ~ sex_factor+treatment_factor+AGE
+                                        +WEIGHIN+Intention+Addiction.Status+time.factor,
+                                        data=longform.data.cig, id=ID, family=poisson(link="log"),
="independence"))

Call:
geeglm(formula = N_CIGARETTES ~ sex_factor + treatment_factor +
    AGE + WEIGHIN + Intention + Addiction.Status + time.factor,
    family = poisson(link = "log"), data = longform.data.cig,
    id = ID, corstr = "independence")

 Coefficients:
                   Estimate   Std.err   Wald Pr(>|W|)
(Intercept)        1.425804  0.213231  44.71  2.3e-11 ***
sex_factorM        0.115765  0.066592   3.02    0.082 .
treatment_factorTx -0.491531  0.078966  38.75  4.8e-10 ***
AGE                -0.001163  0.001936   0.36    0.548
WEIGHIN            -0.000271  0.000576   0.22    0.637
Intention          -0.045144  0.072408   0.39    0.533
Addiction.Status    0.480101  0.028917 275.66  < 2e-16 ***
time.factor        -0.121904  0.021961  30.81  2.8e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation structure = independence
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     2.68   0.252
Number of clusters:   100  Maximum cluster size: 6
> QIC(ind.fitted.pois.model) #-25552.6
      QIC      QICu Quasi Lik       CIC    params      QICC
 -25552.6 -25586.9   12801.5      25.2       8.0 -25552.3
```

Table 8D. Poisson GEE model with unstructured working correlation matrix

```
Call:
geeglm(formula = N_CIGARETTES ~ sex_factor + treatment_factor +
    AGE + WEIGHIN + Intention + Addiction.Status + time.factor,
    family = poisson(link = "log"), data = longform.data.cig,
    id = ID, corstr = "unstructured")

 Coefficients:
                    Estimate   Std.err   Wald Pr(>|W|)
(Intercept)         1.306816  0.180527  52.40  4.5e-13 ***
sex_factorM         0.088987  0.061630   2.08     0.15
treatment_factorTx -0.368115  0.067765  29.51  5.6e-08 ***
AGE                -0.001327  0.001764   0.57     0.45
WEIGHIN            -0.000223  0.000536   0.17     0.68
Intention          -0.012418  0.067023   0.03     0.85
Addiction.Status    0.495116  0.027653 320.57  < 2e-16 ***
time.factor        -0.119403  0.020837  32.83  1.0e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation structure = unstructured
Estimated Scale Parameters:

            Estimate Std.err
(Intercept)     2.67   0.238
  Link = identity

Estimated Correlation Parameters:
          Estimate Std.err
alpha.1:2    0.772  0.1037
alpha.1:3    0.432  0.0860
alpha.1:4    0.110  0.0689
alpha.1:5   -0.164  0.0837
alpha.1:6   -0.516  0.1164
alpha.2:3    0.537  0.0648
alpha.2:4    0.330  0.0555
alpha.2:5    0.153  0.0745
alpha.2:6   -0.116  0.1083
alpha.3:4    0.544  0.0544
alpha.3:5    0.512  0.0668
alpha.3:6    0.387  0.0932
alpha.4:5    0.850  0.0782
alpha.4:6    0.879  0.0918
alpha.5:6    1.322  0.1136
Number of clusters:   100  Maximum cluster size: 6
> QIC(uns.fitted.pois.model) #-25552.26
       QIC       QICu Quasi Lik       CIC    params      QICC
-25552.26 -25555.53  12785.76      9.63      8.00 -25550.34
```

APPENDIX E

R CODES FOR BAYESIAN ANALYSIS

```
#Interim Data Monitoring Clean Code
library(pracma)
library(extraDistr)

#############################################################################
############################### POISSON INFERENCE ###########################

######################### *Poisson-Gamma example*#####################
#Root find the parameters alpha, beta to use for the posterior model
p.prior<- 0.3
R0<- 0.024
beta<- function(alpha){R0/(alpha-1)}
eq<- function(alpha) {p.prior-pgamma(R0,alpha, 1/beta(alpha))}
alpha.sol<- uniroot(eq, c(2,10))$root
beta.sol<- beta(alpha.sol)
print(alpha.sol)
print(beta.sol)

#######################################
#Model for t1=400
t1=400
prob_poigam1<- function(r,n) {
  prob_poigam1<- r^(n+a-1)*exp(-r*(t1+1/b))
  return(unionvector=c(prob_poigam1))
}
area1<- function(n) integrate(prob_poigam1, lower=0, upper=Inf, n=n, abs.tol
= 0L )$value
normalizing.const.poigam<- Vectorize(area1)
n =c(seq(1,22,by=1))
normalizing.const.poigam(n)

prob_poigam2<- function(r,n) {
  prob_poigam2<- (r^(n+a-1)*exp(-r*(t1+1/b)))
  return(unionvector=c(prob_poigam2))
}

area2<-function(n) integrate(prob_poigam2, lower=0, upper=0.024,n=n )$value
v.areapoigam1<- Vectorize(area2)
n =c(seq(1,22,by=1))
v.areapoigam1(n)

post.prob.poi.gam.t400<- v.areapoigam1(n)/normalizing.const.poigam(n)


#Model for t2=500
t2=500
prob_poigam1<- function(r,n) {
  prob_poigam1<- r^(n+a-1)*exp(-r*(t2+1/b))
  return(unionvector=c(prob_poigam1))
```

```
}
area1<- function(n) integrate(prob_poigam1, lower=0, upper=Inf, n=n, abs.tol
= 0L )$value
normalizing.const.poigam<- Vectorize(area1)
n=c(seq(1,22,by=1))
normalizing.const.poigam(n)

prob_poigam2<- function(r,n) {
  prob_poigam2<- (r^(n+a-1)*exp(-r*(t2+1/b)))
  return(unionvector=c(prob_poigam2))
}

area2<- function(n) integrate(prob_poigam2, lower=0, upper=0.024, n=n)$value
v.areapoigam1<- Vectorize(area2)
n =c(seq(1,22,by=1))
v.areapoigam1(n)

post.prob.poi.gam.t500<- v.areapoigam1(n)/normalizing.const.poigam(n)


#Model for t3=500
t3=600
prob_poigam1<- function(r,n) {
  prob_poigam1<- r^(n+a-1)*exp(-r*(t3+1/b))
  return(unionvector=c(prob_poigam1))
}
area1<- function(n) integrate(prob_poigam1, lower=0, upper=Inf, n=n, abs.tol
= 0L )$value
normalizing.const.poigam<- Vectorize(area1)
n =c(seq(1,22,by=1))
normalizing.const.poigam(n)

prob_poigam2<- function(r,n) {
  prob_poigam2<- (r^(n+a-1)*exp(-r*(t3+1/b)))
  return(unionvector=c(prob_poigam2))
}

area2<-function(n) integrate(prob_poigam2, lower=0, upper=0.024,n=n )$value
v.areapoigam1<- Vectorize(area2)
n =c(seq(1,22,by=1))
v.areapoigam1(n)

post.prob.poi.gam.t600<- v.areapoigam1(n)/normalizing.const.poigam(n)


#These are the posterior probabilities at t=400,500,600 for this conjugate
model
print(post.prob.poi.gam.t400)
print(post.prob.poi.gam.t500)
```

```
print(post.prob.poi.gam.t600)



############## POISSON INVERSE GAMMA **************

#Poisson-inverse gamma example
library(invgamma)
p.prior<- 0.3
R0<- 0.024
beta<- function(alpha){R0*(alpha+1)}
eq<- function(alpha) {p.prior-pinvgamma(R0, alpha, beta(alpha))}
alpha.sol<- uniroot(eq, c(1,10))$root
beta.sol<- beta(alpha.sol)
print(alpha.sol)
print(beta.sol)

# Model for t1=400
t1=400
poi.invgam1<- function(r,n) {
  poi.invgam1<- r^(n-a-1)*exp(-(r*t1+b/r))
  return(unionvector=c(poi.invgam1))
}

area1<- function(n) integrate(poi.invgam1, lower=0, upper=Inf, n=n, abs.tol =
0L )$value
nonconj.normalizing.const<- Vectorize(area1)
n =c(seq(1,30,by=1))
nonconj.normalizing.const(n)

poi.invgam2<- function(r,n) {
  poi.invgam2<- r^(n-a-1)*exp(-(r*t1+b/r))
  return(unionvector=c(poi.invgam2))
}

area2<- function(n) integrate(poi.invgam2, lower=0, upper=0.024,n=n )$value
v.areapoigam1<- Vectorize(area2)
n =c(seq(1,30,by=1))
v.areapoigam1(n)

post.prob.poi.invgam.t400<- v.areapoigam1(n)/nonconj.normalizing.const(n)

# Model for t2=500
t2=500
poi.invgam1<- function(r,n) {
  poi.invgam1<- r^(n-a-1)*exp(-(r*t2+b/r))
  return(unionvector=c(poi.invgam1))
}
```

```
area1<- function(n) integrate(poi.invgam1, lower=0, upper=Inf, n=n, abs.tol =
0L )$value
nonconj.normalizing.const<- Vectorize(area1)
n =c(seq(1,30,by=1))
nonconj.normalizing.const(n)

poi.invgam2<- function(r,n) {
  poi.invgam2<- r^(n-a-1)*exp(-(r*t2+b/r))
  return(unionvector=c(poi.invgam2))
}

area2<- function(n) integrate(poi.invgam2, lower=0, upper=0.024,n=n )$value
v.areapoigam1<- Vectorize(area2)
n=c(seq(1,30,by=1))
v.areapoigam1(n)

post.prob.poi.invgam.t500<-v.areapoigam1(n)/nonconj.normalizing.const(n)

# Model for t3=500
t3=600
poi.invgam1<- function(r,n) {
  poi.invgam1<- r^(n-a-1)*exp(-(r*t3+b/r))
  return(unionvector=c(poi.invgam1))
}

area1<- function(n) integrate(poi.invgam1, lower=0, upper=Inf ,n=n, abs.tol =
0L )$value
nonconj.normalizing.const<- Vectorize(area1)
n =c(seq(1,30,by=1))
nonconj.normalizing.const(n)

poi.invgam2<- function(r,n) {
  poi.invgam2<- r^(n-a-1)*exp(-(r*t3+b/r))
  return(unionvector=c(poi.invgam2))
}

area2<- function(n) integrate(poi.invgam2, lower=0, upper=0.024,n=n )$value
v.areapoigam1<- Vectorize(area2)
n=c(seq(1,30,by=1))
v.areapoigam1(n)

post.prob.poi.invgam.t600<- v.areapoigam1(n)/nonconj.normalizing.const(n)

#These are the posterior probabilities at t=400,500,600 for this nonconj
model
print(post.prob.poi.invgam.t400)
print(post.prob.poi.invgam.t500)
print(post.prob.poi.invgam.t600)
```

```
#########################################################################
############################## NORMAL INFERENCE ##########################


#Normal-Normal example
delta=8
sigma=17.4
n=30

hypothesized_vals = c(seq(-10,7,by=0.10))
estimate_sigma<- function(delta,prior) {
  estimated_sigma <- (-delta)/qnorm((1-prior))
  return(estimated_sigma)
}
sigma0<- estimate_sigma(8,0.8)

prob_norm.norm<- function(n,mean_diff) {
  numerator=-((delta)/(sigma0^2)+(hypothesized_vals*n)/(2*sigma^2))
  denominator<- sqrt((1/sigma0^2)+n/(2*sigma^2))
  probability <- 1-pnorm(numerator/denominator)
  return(unionvector<- c(probability))
}
prob_norm.norm(n,hypothesized_vals)

#Normal-Normal non bayesian estimation of sample size
sample.size.est<- function(alpha,beta) {
  n <- 2*(sigma/delta)^2*((qnorm(1-alpha)-qnorm(beta))^2)
  return(n)
}
n<- sample.size.est(0.05,0.25)
ceiling(n)
#Computing the actual probability of type 2 error
alpha=0.05
beta=0.25
k=qnorm(1-alpha)
type2_beta<- pnorm(k-(delta/(sigma*sqrt(2/ceiling(n)))))
print(type2_beta)

##################  TRYING A NORMAL WITH CAUCHY PRIOR **********

sigma<- 17.4
p.prior<- 0.8
delta0<- 8
sigma0<- -delta0/tan(pi/2-pi*p.prior)
n<- 30
p.post<- c()

for (d in seq(-10.3, 5.3, by=0.1)){
  func<- function(x){exp(-(x-d)^2/(4*sigma^2/n))/(1+(x-delta0)^2/sigma0^2)}
```

```
  i<- round(10*(d+10.3)+1,0)
  p.post[i]<- integrate(func, 0, Inf)$value/integrate(func, -Inf, Inf)$value
}

round(p.post,5)



################################################################################
########################## BINOMIAL INFERENCE ##################################
#Calculating Numeric Binomial\Beta example

#Code to get alpha/beta parameter estimates
N<- 110
p.prior<- 0.4
mode<- 0.23
p0<- 0.25
beta<- function(alpha) { (alpha-1)/mode+2-alpha}
eq<- function(alpha) {p.prior-pbeta(p0,alpha, beta(alpha))}
alpha.sol<- uniroot(eq, c(1,7))$root
beta.sol<- beta(alpha.sol)

print(alpha.sol)
print(beta.sol)

#Specifying a normalizing constant(denominator) for bin-beta posterior
x = c(seq(1,100,by=1))
binbeta1 <- function(p,x) {
  binbeta1<- (p^(x+a-1)*(1-p)^(N-x+b-1))
  return(unionvector=c(binbeta1))
}
area<- function(x) integrate(binbeta1, lower=0, upper=1, x=x, abs.tol=
0L)$value
binbeta1.area<- Vectorize(area)
x =c(seq(1,100,by=1))
binbeta1.area(x)

binbeta2<- function(p,x) {
  binbeta2<- (p^(x+a-1)*(1-p)^(N-x+b-1))
  return(unionvector=c(binbeta2))
}
area<- function(x) integrate(binbeta2, lower=0, upper=0.25,x=x )$value
binbeta2.area<- Vectorize(area)
x =c(seq(1,100,by=1))
binbeta2.area(x)

post_prob_binbeta=binbeta2.area(x)/binbeta1.area(x)
print(post_prob_binbeta) #Looks good
```

```
### ########## CALCULATING BINOMIAL TRUNCATED NORMAL POSTERIOR
N=110
x = c(seq(1,100,by=1))

#The parameters are reused from the Bin/Beta example as it is impossible
#to estimate them with this prior
mu=0.5
sigma=1/4

#These constants basically disappear during integration. It's useless
#regardless of what values they are
alpha=(a-mu)/sigma
beta=(b-mu)/sigma

#Specifying the Normalizing constant(denominator)
TN1 <- function(p,x) {
  a1<- p^(x)*(1-p)^(N-x)
  b1<- exp(-(p-mu)^2/(2*sigma^2))
  c1<- sigma*sqrt(2*pi)*(pnorm(beta)-pnorm(alpha))
  TN1<- (a1*b1)/c1
  return(unionvector=c(TN1))
}

area<- function(x) integrate(TN1, lower=0, upper=1, x=x, abs.tol= 0L)$value
TN1.area<- Vectorize(area)
x=c(seq(1,100,by=1))
TN1.area(x)

#Specifying the Posterior function (numerator)
TN2<- function(p,x) {
  a2<- p^(x)*(1-p)^(N-x)
  b2<- exp(-(p-mu)^2/(2*sigma^2))
  c2<- sigma*sqrt(2*pi)*(pnorm(beta)-pnorm(alpha))
  TN2<- (a2*b2)/c2
  return(unionvector=c(TN2))
}
area<- function(x) integrate(TN2, lower=0, upper=0.25, x=x)$value
TN2.area<- Vectorize(area)
x=c(seq(1,100,by=1))
TN2.area(x)

post.TN.probs=TN2.area(x)/TN1.area(x)
print(post.TN.probs)


######################### SEQUENTIAL DOUBLE INTEGRAL EXAMPLE IN SECTION 4.1.3
#########
#Type 1 error (1-alpha)
h1<- function(x, y) exp(-0.5*(x^2+y^2))
```

```
F1<- function(x) {
  fun <- function(y) (1/(2*pi)) * h(x, y)
  integrate(fun, -Inf, k*sqrt(2)-x)$value
}
F1<- Vectorize(F1)  # requested when using integrate()

eq_type1<- function(k) {integrate(F1, -Inf, k)}


#The way I go about solving for the parameters is by testing numbers out

#Try k=3
k=k1=3
eq_type1(k1) #0.9975426

#Try k=2
k=k2=2
eq_type1(k2) #0.9620106

#Try k=1.75
k=k3=1.75
eq_type1(k3) #0.9350068

#The root is somewhere between k=1.75 and k=2

#Try k=1.9
k=k4=1.9
eq_type1(k4) #0.9525778

#Try k=1.8
k=k5=1.8
eq_type1(k5)#0.9413549

#Try k=1.85
k=k6=1.85
eq_type1(k6) #0.9472036

#Try k=1.88
k=k7=1.88
eq_type1(k7) #0.9499485, Close enough! So we set our k=1.88
k=1.88


####################################
#Type 2 error (beta)
h<- function(x, y) exp(-0.5*(x^2+y^2))
F2<- function(x) {
  fun<- function(y) (1/(2*pi)) * h(x, y)
  area<-integrate(fun, -Inf, k*sqrt(2)-2*sqrt(n_star)-x)$value
```

```
}
F2<- Vectorize(F2)  # requested when using integrate()
eq_type2<- function(n_star) {integrate(F2, -Inf, k-sqrt(n_star))}


#The way I go about solving for the parameters is by testing numbers out

#Try n_star=3
n_star=n_star1=3
eq_type2(n_star1) #0.2535443

#Try n_star=2
n_star=n_star2=3.1
eq_type2(n_star2)#0.2410883

#Try n_star=3.5
n_star=n_star3=3.5
eq_type2(n_star3)#0.1963617

#The root is somewhere in between 3 and 3.1

#Try n_star=3.05
n_star=n_star4=3.05
eq_type2(n_star4) #0.2472497

#Now, the root is somewhere between 3 and 3.05

#Try n_star=3.03
n_star=n_star5=3.03
eq_type2(n_star5) #0.2497515, Close enough! So let n_star=3.03
n_star=3.03


#########################################################
#DETERMINING THE SAMPLE SIZE (SEE SECTION 4.1.3)##########

sigma=17.4
alpha=0.05
beta=0.25
delta=8

n=(n_star*2*sigma*sigma)/(delta^2)
print(n)
print(ceiling(n))# From this data, 29 patients are needed

#########################################################################
```

REFERENCES

REFERENCES

Bartholomew M. 2002. “James Lind's Treatise of the Scurvy (1753).” *Postgraduate medical journal*, 78(925): 695–696.

Bhatt, Arun. 2010. “Evolution of Clinical Research: A History Before and Beyond James Lind.” *Perspectives in clinical research*, 1(1): 6–10.

Caruana, Edward. J., Marius Roman, Jose Hernández-Sánchez, & Piergiorgio Solli. 2015. “Longitudinal Studies.” *Journal of Thoracic Disease*, 7(11): E537–E540.

Cox, David R. 1972. “Regression Models and Life-Tables.” *Journal of the Royal Statistical Society*, 34: 187-220.

Davies, Madhu, and Faiz Kermani. 2008. *A Quick Guide to Clinical Trials: For People Who May Not Know It All*. BioPlan Associates.

Grunkemeier, G. L., Johnson, D. M., & Naftel, D. C. 1994. “Sample Size Requirements for Evaluating Heart Valves with Constant Risk Events.” *The Journal of heart valve disease*, 3(1): 53–58.

Gupta, Sandeep K. 2012. “Use of Bayesian Statistics in Drug Development: Advantages and Challenges.” *International Journal of Applied & Basic Medical Research*, 2(1): 3–6.

Kaplan, Edward L., and Paul Meier. 1958. “Nonparametric Estimation from Incomplete Observations.” *Journal of the American Statistical Association*, 53(282): 457-481.

Nelson, Wayne. 1972. “Theory and Applications of Hazard Plotting for Censored Failure Data.” *Technometrics*, 14(4): 945-966.

Shapiro, Arthur K. 1964. “A Historic and Heuristic Definition of the Placebo.” *Psychiatry*, 27(1): 52-58.