June 2021

Data-Driven Analytical Modeling of Multiple Myeloma Cancer, U.S. Crop Production and Monitoring Process

Lohuwa Mamudu
University of South Florida

Data-Driven Analytical Modeling of Multiple Myeloma Cancer, U.S. Crop Production and Monitoring Process

by

Lohuwa Mamudu

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
Department of Mathematics & Statistics
College of Arts and Sciences
University of South Florida

Major Professor: Chris P. Tsokos, Ph.D.
Kandethody M. Ramachandran, Ph.D.
Lu Lu, Ph.D.
Barbos Andrei, Ph.D.

Date of Approval:
June 2, 2021

Keywords: Statistical Modeling,Desirability Function, NHPP, Cancer Survivorship,
Production Evaluation

Dedication

To Almighty Allah

To my beloved family, especially

My parents, M.A. Mamudu and Zuwerah Mamudu

To my brother, Hadii Mamudu

To my wife and son, Adiza and Amir Lohuwa

My Advisor and Mentor, Chris P. Tsokos

To Rotary International.

Acknowledgments

To begin with, I send my sincere and profound thanks and gratitude to the Almighty Allah (SWT) for all His mercy and favor. I would not have gotten this far in my career without Allah's blessings and guidance.

Next, my deepest love and sincere appreciation to Professor Chris P. Tsokos and my brother, Professor Hadii Mamudu for their inexhaustible advice, unflinching support, assistance, and persistent guidance throughout my doctoral degree program. They are my mentors and source of inspiration. Especially, without Professor Chris P. Tsokos's guidance, this dissertation would not have come to fruition. Words are not enough to thank them.

Further, I am also grateful to the members of my dissertation committee: Dr. Kanadethody M. Ramachandran, Dr. Lu Lu, and Dr. Barbos Andrei for their support and advice throughout this process. I am thankful to Dr. Tu Yicheng for chairing the session.

Also, my deep thanks to Dr. Faustine Williams for her sincere support, through whom I am a Freelance Data Analyst for the National Institute of Health (NIH). I also thank Rotary International for the Skelton-Jones Scholarship global grant award, District 7570, in 2018. This scholarship brought immense changes to my life and my family.

I am also thankful to my colleagues Mahdi Goudarzi and his wife, Nawa Raj Pokhrel, and A.K.M Raquibul Bashar for their assistance and support; as well as Mohamed Abu Sheha, and all my fellow graduate friends for their constructive suggestions and critiques.

Finally, I express my heartfelt love and thanks to my parents for their priceless prayers and blessings, my beloved wife Adiza for her support and encouragement, my exciting son Amir Lohuwa, and my entire family. I share this grand achievement with you all.

Table of Contents

List of Tables

List of Figures

Abstract

Globally, cancer disease is a major health issue causing a lot of deaths. The duration of time an individual diagnosed with a particular type of cancer survives has become a major area of research concern. The Kaplan Meier and Cox Proportional Hazard (Cox-PH) model have been a traditionally used method for survival analysis of cancer data. These techniques of cancer survival analysis are developed from nonparametric and semi-parametric approaches, respectively, which are not as robust as a parametric approach. In this dissertation, we proposed a new method of cancer survival analysis based on a parametric approach using multiple myeloma cancer (MMC) data. Firstly, we performed a parametric analysis of only the survival times without taking into account the risk factors, obtaining the survival function, and comparing it with the Kaplan Meier survival function.

Next, we assessed the survival times taking into consideration the risk factors contributing to the survival times. We developed a high-quality and well-validated Cox-PH model, identifying the significant risk factors and estimating the survival function. Further, a parametric analysis was conducted, obtaining the survival function from highly accurately predicted survival times from a well-developed and validated nonlinear statistical model that identifies the significant risk factors and ranks them according to the percentage contribution to the survival times. We compared the quality of the two models and their robustness in estimating the survival function. Our parametric approaches of cancer survival analysis outperformed both the Kaplan Meier and Cox-PH model of cancer survival analysis, given better estimates of the survival times. The proposed statistical modeling and parametric approach for cancer survival analysis used in this dissertation for multiple myeloma cancer can be generalized and applied to the various cancer diseases in the world. This study offers a more improved, effective, and efficient therapeutic/treatment strategy for cancer diseases.

Another research study in this dissertation is corn production. Corn is globally known to be the most economically viable and versatile agricultural product. The United States (U.S.) is noticeable the world's leading producer of corn. In this study, we proposed a real data-driven analytical model for the returns of corn production in the U.S. utilizing data obtained from the U.S. Department of Agriculture (USDA) from 1975 to 2018. The developed model is of high quality, satisfies all necessary assumptions, well-validated, and predicts the returns with a high degree of accuracy. It identifies significant risk factors, including interaction, and ranks them according to percent contribution to the returns. We further performed an optimization analysis of the returns using the desirability function approach, obtaining the optimum return and the optimal values of the risk factors needed to maximize the returns of corn production in the U.S. We also obtained the confidence region of the optimum value of the return for the purpose of statistical inferences, as well as surface response plots to assess the combination of risk factors contribution to the returns. Finally, we proposed a time-dependent analytical model to evaluate and monitor the returns based on whether it is increasing, decreasing, or remaining unchanged. The evaluation and monitoring process utilizes $\beta - \textit{factor}$ obtained from the intensity function of the nonhomogeneous Poisson process (NHPP) / Power Law Process (PLP). This study offers a more robust and efficient approach for maximizing the returns of crop production.

The approach and methodology of evaluating and monitoring the returns of corn production used in this study can be extended to multidisciplinary fields studies, including different settings of production, finance in monitoring the stock returns, health science in monitoring the number of deaths and reported cases from disease, environmental science in monitoring the emission of carbon dioxide, cybersecurity in monitoring the vulnerability scores of software system, transportation in monitoring the number of accident, etc.

Chapter 1: Introduction

In the 21st century, cancer diseases have become one of the major leading causes of death in the world, with every one of six deaths caused by cancer. In 2017, an estimated 9.6 million people died from cancer across the globe from various types of cancer diseases [1]. This is a staggering and incredible number of deaths that requires tremendous attention. How long an individual survives or demise after being diagnosed with a particular kind of cancer is imperiously concern, given the fact that most cancer diseases remain incurable. Most research studies into cancer survivorship have focused on improving the therapeutic or treatment strategy to increase the survival times of individuals after being diagnosed with cancer. The methodology or approach adopted by most research studies on cancer survivorship uses nonparametric and semi-parametric techniques, which is less robust compared to using a parametric approach. To contribute immensely to improving the treatment strategy of cancer diseases, consequently the survival times of cancer patients, we have proposed a parametric statistical modeling approach to investigating cancer survivorship using multiple myeloma cancer data. We employed some statistical tools, including machine learning techniques in the parametric approach to obtain more robust results of the survival analysis.

Further, we proposed analytical modeling for crop production. Increasing or maximizing the returns from crop production is the inevitable goal of every crop production firm or industry in the agriculture sector. Not much research has been done in this subject area. Most crop production firms have relied on traditional economic concepts of profit or return maximization based on comparing the marginal revenue to marginal cost or average cost to the market price of the product. These concepts have several limitations and flaws, failing in their application in most cases. We developed a more comprehensive, effective, and efficient analytical modeling approach to maximize the returns of crop production in this dissertation. Our proposed approach predicts the returns with high accuracy, taking into consideration the variables or indicators contributing to returns. We optimized the returns and the contributing variables, and then evaluate and monitor the dynamics in the returns. Our study and investigation in the subject area of crop production focused on U.S. corn production. A similar study using our approach can be conducted for the returns of other crop production.

This dissertation is organized into nine chapters. The first four chapters after this chapter investigate cancer survivorship using multiple myeloma cancer data. The next three chapters investigate the returns of U.S. corn production. Finally, chapter nine is dedicated to some future research works in the subject areas. The methods or approaches we developed are not only applicable to the subject area in this dissertation but also applicable to research investigations of other cancer diseases and crop production, respectively.

Chapter 2: Parametric and Non-parametric Analysis of the Survival Times of Patient with Multiple Myeloma

Multiple myeloma cancer (MMC) is a type of cancer that remains incurable. In the last decade, most research into MM has focused on investigating the improvement in the therapeutic strategy. This chapter assesses or investigates the survival times of 48 patients diagnosed with MM based on parametric and non-parametric techniques. We performed parametric survival analysis and found a well-defined probability distribution of the survival time to follow three-parameter log-normal. We then estimated the survival probability and compared it with the commonly used non-parametric Kaplan-Meier survival analysis of the survival times. The comparison of the survival probability estimate of the two methods revealed a better survival probability estimate by the parametric method than the Kaplan-Meier. The study in this chapter offers therapeutic significance for further enhancement in the treatment strategy of multiple myeloma cancer. The finding in this chapter has been published [60].

The study in this chapter is organized as follows: Section 2.1 introduces and review some literature of studies on MMC; Section 2.2 presents the description of data used in this study; Section 2.3 presents the parametric analysis of the survival times of MMC patients; Section 2.4 assesses the Kaplan Meier analysis of the survival times of MMC patients; Section 2.5 presents the comparison of the parametric analysis of the survival times with that of the Kaplan Meier analysis; Section 2.6 discusses the findings in this study; and finally, the research contributions of this chapter is given in Section 2.7.

2.1 Introduction

Multiple Myeloma cancer (MMC), also known as Kahler disease, myelomatosis, and plasma cell myeloma is a type of cancer that starts from malignant plasma cell (Specifically the white blood cell) [2]. As part of the human immune system is antibodies produced by the plasma cell which fight against germs and other substances harmful to the human body. When the plasma cell becomes abnormal, called the myeloma cell, it causes myeloma [3]. When the myeloma cells increase, it accumulates in the bone marrow and overcrowds the active blood cells, and with time may destroy the solid part of the bone. Hence, the collection of several myeloma cells in the bones causes multiple myeloma cancer. [4, 5]. The development of the myeloma cells is shown in Figure 2.1, [3, 6].

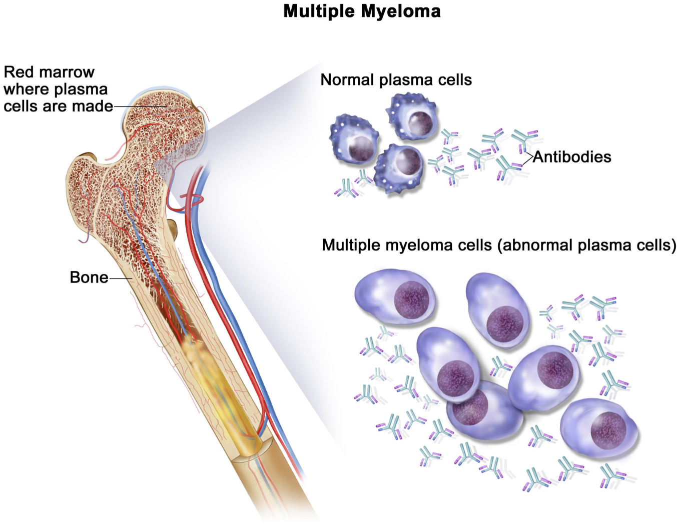
Multiple Myeloma
Red marrow where plasma cells are made
Normal plasma cells
Antibodies
Multiple myeloma cells (abnormal plasma cells)
Bone

Figure 2.1: Development of the Myeloma Cell

Abnormal antibodies are produced by the abnormal plasma cells causing kidney problems and highly thick blood [7]. MMC has no specific causes. However, some research has found obesity, radiation exposure, family history, and certain chemicals as associated with the cause of MMC [8, 9, 10, 11]. There have been some treatment recommendations for multiple myeloma focused on decreasing the clonal plasma cell population and consequently decrease the symptoms of disease [12]. A preferred treatment like high-dose chemotherapy, commonly with bortezomib-based regimens, and lenalidomide-dexamethasone followed by autologous hematopoietic stem-cell transplantation (ASCT), the transplantation of a person's stem cells have been recommended for MMC patients under 65 years[13]. In 2017, a meta-analysis performed has shown that post-ASCT maintenance therapy with lenalidomide has improved the progression-free survival and overall survival in persons at standard risk [14]. Whereas in 2012, it was found from a clinical trial that intermediate and high-risk disease patients benefit from a bortezomib-based maintenance regimen [15].

Statistically, approximately 30,000 new patients are diagnosed with MMC in the United States (U.S.) every year, making it the second most common hematologic malignancy in the U.S. [16]. In 2019, a report by the Surveillance, Epidemiology and End Results (SEER) Cancer Institute reported that of all new cancer cases in the U.S for MMC constitutes 1.8% and ranked among the top 14 list of cancer diseases [4]. A further projection by SEER indicates 32,110 estimated new cases of MMC and an estimated 12,960 MM patients are expected to die. Those figures are scary and intriguing and cannot be overlooked. There is a sufficient increase compared with the 24,050 estimated new MM cases reported in 2014 [17]. The identified risk factors of MM is reported to be common among black race, families with MMC history, and being a male [4, 18]. SEER cancer institute reported from 2012-2016 that 63.1% of all races and sexes of MMC cases are aged 65 or greater.

Though multiple myeloma cancer disease remains incurable, most researches into MM focused on how to improve the survival times of patients diagnosed with MMC. The Kaplan-Meier (KM) method has been popularly used for analyzing cancer survivorship data in recent times due to the simplicity of its usage. It is often used to compare the survival difference of observations/groups base on the log-rank test of the null hypothesis that there is no difference. KM is mostly used for longitudinal studies like a cohort study [19]; an example is the present study (i.e. the survival time of patients diagnosed with multiple myeloma). Brain et al [20] used Kaplan and Meier to test whether there was a significant difference in the overall survival duration between the categories of risk factors based on the generalized Wilcoxon test and the log-rank test. They found a significant difference in the survival duration between MMC patients with $LI\% < 1\%$ (i.e. low percentage labeling index) and $LI\% \geq 1\%$ (i.e. high percentage labeling index). Also, there was a significant difference in the survival duration for MMC patients with the number of DNA synthesizing (S) values $< 1.0 \times 10^{11}$ and S values $\geq 1.0 \times 10^{11}$. Shaji K. Kumar et al [21], used the Kaplan Meier to test for the significant difference of the overall survival from the time of post transplantation relapse between MMC group treated subsequently with one or more of the newer drugs (thalidomide, bortezomib, or lenalidomide) and those not exposed to the newer drug, and they found a significant difference between the two groups.

In the present study, we developed a parametric and non-parametric survival analysis of the survival time of patients diagnosed with multiple myeloma. We believe that being able to find the unique/correct probability distribution or probabilistic behavior that characterizes the survival time is a great step towards getting a good and accurate prediction of the survival duration. It is well known that parametric analysis is more powerful in decision analysis than its non-parametric counterpart. Feigl and Zelen ([1965] p. 835) and other authors have pointed out that the assuming exponential distribution works well for studying the survival of cancer-related cases [22]. However, almost every data given on any cancer survival problem may have an associated well-defined probability distribution. Hence, assuming an exponential distribution for a given cancer survival case without any further investigation is a serious mistake that will lead to making incorrect decisions. We compare the more powerful parametric analysis of the survival time to the commonly known non-parametric Kaplan-Meier analysis of cancer survivorship.

2.2 Data Description

The data used in this research is from West Virginia University Medical Center provided by Harley [23, 24]. Originally, the data constituted survival times of 72 multiple myeloma cancer (MMC) patients diagnosed and treated with alkylating agents [23]. 65 out of 72 patients provided complete data for 16 contributing variables (risk factor) whiles the remaining

7 were eliminated due to missing data in at least one of the 16 risk factors. Thus, the survival or death times of the MMC patients is driven by the 16 risk factors. Given that a patient is diagnosed with myeloma, the 16 risk factors were recorded, and the time up to which the patient survived the disease was noted (called the survival time from diagnosis to the nearest month), given by Table 2.1. Of the 65 patients, 48 and 17 were dead and alive, respectively. In the present research, we utilized the complete data of the 48 patients with the survival times for our analysis.

Table 2.1: Variables Recorded for Multiple Myeloma Patients

Symbol	**Variable Name**
t	Survival time from diagnosis to nearest month $+1$
X_1	Log blood urea nitrogen (BUN)/serum creatinine at diagnosis
X_2	Hemoglobin at diagnosis
X_3	Platelets at diagnosis 0 – abnormal, 1 – normal
X_4	Infections at diagnosis 0 – none, 1 – present
X_5	Age at diagnosis (complete years)
X_6	Gender 1 – male, 2 – female
X_7	Log white blood cell (WBC) at diagnosis
X_8	Fractures at diagnosis 0 – none, 1 – present
X_9	Log %BM at diagnosis (log % plasma cells in bone marrow)
X_{10}	% Lymphocytes in peripheral blood at diagnosis
X_{11}	% Myeloid cells in peripheral blood at diagnosis
X_{12}	Proteinuria at diagnosis
X_{13}	Bence Jone protein in urine at diagnosis 1 – present, 2 – none
X_{14}	Total serum protein at diagnosis
X_{15}	Serum globin (gm%) at diagnosis
X_{16}	Serum calcium (mgm%) at diagnosis

Before we proceeded to performed the parametric analysis of the survival times of patients with multiple myeloma, we wanted to know whether there is a difference in the survival times for gender, i.e. male and female. Given that we have a small data of only 48 patients, we used the log-rank test [50] to compare the difference in survival times of male and female . From Figure 2.2, the log-rank test resulted in a large $p - value = 0.45$, indicating a failure to reject the null hypothesis (i.e. $H_0 : \mu_M = \mu_F$) that there is no difference in the survival times of males and females. Given that we have a small sample size of only 48 patients and the fact that there is no difference in the survival times of males and females, we proceeded with parametric analysis without considering stratification based on gender.

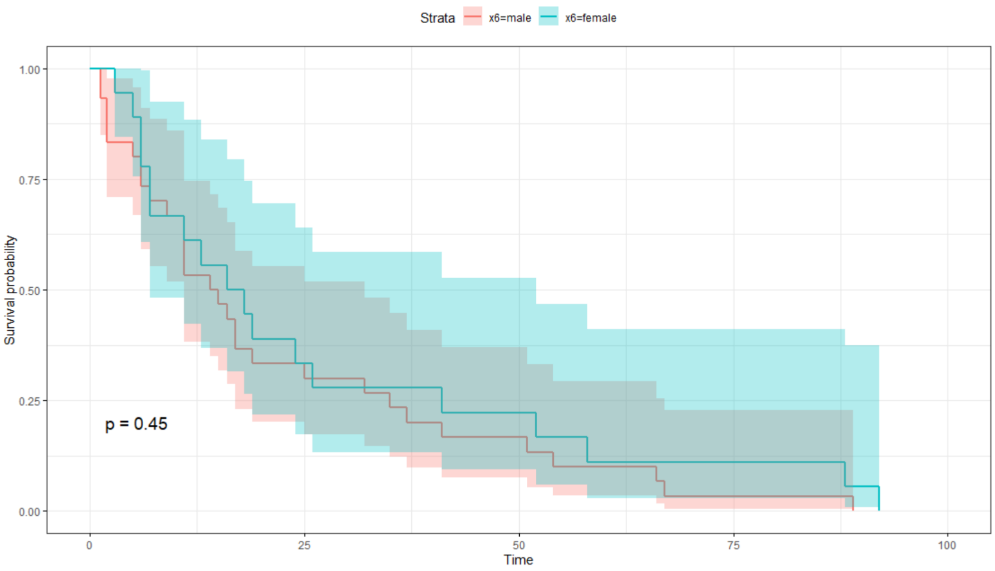

Figure 2.2: Log-Rank test for Difference in Survival Time of Gender

2.3 Parametric Analysis of the Survival Times of Multiple Myeloma

Multiple Myeloma cancer has been, and continue to be, the subject of many research studies. The main goals of these studies are to investigate the means of improving the therapeutic/treatment and survival of MMC patients. Nearly 1 to 5 per $100,000$ individuals are affected by MMC each year worldwide with a higher incidence in the West [26]. The continuous research into the survival rate of MMC has seen an improvement from 34.5% to 49.6% in the last 13 years, a statistic reported by the Multiple Myeloma Research Foundation (MMRF) [27]. The difficulties in the investigation of the survival of MMC has been hampered by the limitation of data. Most often data collected has a small sample size and duration. This makes it tedious to undertake a parametric analysis. As a result, most analysis of survival time of MMC has been based on nonparametric methods. Brian G.M. Durie et al [20] used the Kaplan-Meier method to calculate the actual survival curves of MMC stages and test their differences using the generalized Wilcoxon test, and then used the log-rank test to test the difference in the survival duration. Shaji K. Kumar et al [21] also used Kaplan-Meier to test for the overall survival improvement and tested for statistical significance using the 2-tailed log-rank test. Nonparametric tests are preferable if there is no unique probability distribution for the given data. However, their approach is statistically less robust or not as powerful compared to parametric analysis if a pdf can be found found. On the contrary, if parametric analyses are used prematurely (i.e. assumed a pdf), the result will be misleading.

In the last few decades, scientific transformations have made it possible to employ parametric analysis to data which initially lack a unique probability distribution, particularly for skewed data. For instance, ANOVA is a parametric analysis widely recommended for comparing statistical models and means of different data sets. ANOVA assumes normality, homoscedasticity, and random independent samples. Several suggestions on transformations have been proposed by various authors (Sokal and Rohlf 1995, Zar 1996, Hayek and Buzas 1997, Krebs 1999) to possibly achieve the required assumptions for parametric analysis. To use ANOVA, Rogers, Gilnack, and Fitz (1983) [28], and others used the arcsine-square root transformation. Brown et al. (2004) utilized the arcsine-square root transformation to use the paired t-test. Log transformation has often been applied to skewed data to achieve the normal distribution. Giampaolo Merlini et al [29] used log transformation to reduce asymmetry variables to a low level and avoid obvious outliers. Performing parametric analysis is an imperative approach to achieve good statistical analysis and inference by knowing pdf of the data.

2.3.1 Descriptive Statistics for the Survival Time of Multiple Myeloma

We plotted the histogram to investigate the distribution of the survival time of multiple myeloma of our data as shown in Figure 2.3. We can see that the distribution of the survival time of MMC is right-skewed. Table 2.2 displays the descriptive statistics of the survival time of MMC. If a skewed value is negative (less than zero) implies the data depict left or negative skewness, if a skewed value is positive implies right or positive skewness [54]. So, the positive skewed value of 1.41 in Table 2.2 is further evidence to support the right-skewed behavior as shown in Figure 2.3. Statistics like Kurtosis allow for the assessment of highly extreme values in the data. A positive kurtosis value demonstrates a leptokurtic behavior of the distribution, and a negative value displays a platykurtic behavior of the distribution. A kurtosis value of zero implies the distribution behavior is mesokurtic [31]. Thus, the kurtosis value of 0.78 in Table 2.2 attests to the leptokurtic behaviors in the survival time data of MMC.

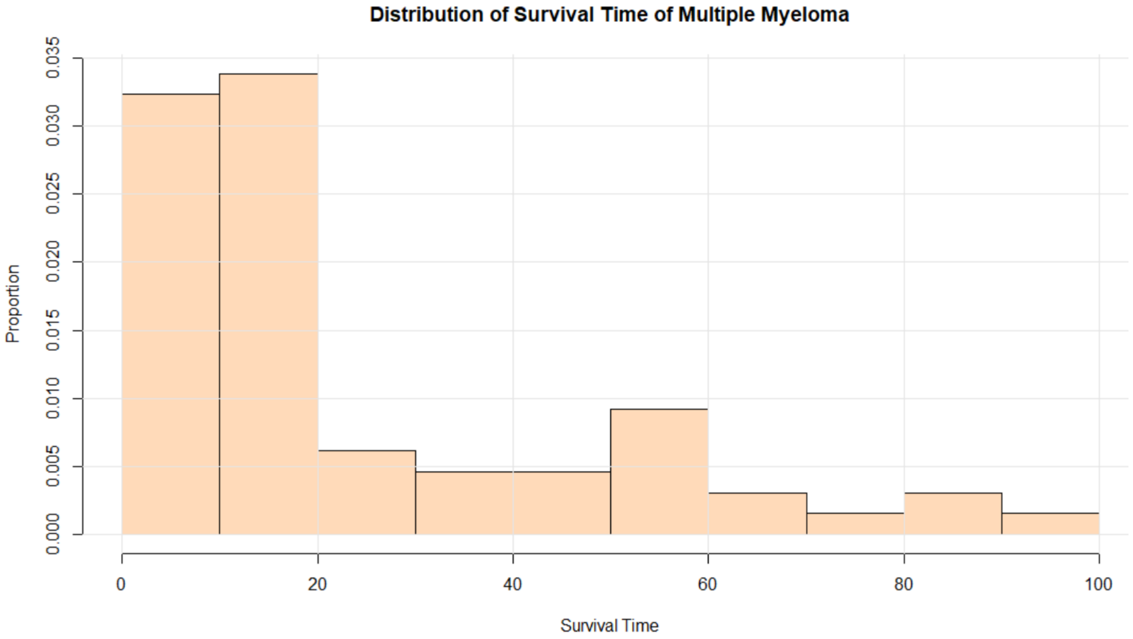

Figure 2.3: Histogram Showing the Distribution of Survival Time of Multiple Myeloma

Table 2.2: Descriptive Statistics of Survival Time of Multiple Myeloma

Mean	Median	Std Err	Std Dev	Kurtosis	Skewness
24.43	15.50	3.56	24.65	0.78	1.33

Table 2.3: Goodness-of-fit Test of the 3P-Lognormal Distribution of the Survival Time.

Type of Test	$p - value$
Kolmogorov-Smirnov	0.90171
Anderson-Darling	0.37878
Chi-Squared	0.69163

2.3.2 Three-Parameter (3P) Log-normal Survival Probability Estimation of the Survival Time of Patients with Multiple Myeloma

After the assessment of Figure 2.3, and the descriptive statistics in Table 2.2, we find that the probability distribution of the survival time of MMC follows the 3-parameter log-normal distribution. Table 2.3 shows three different goodness-of-fit tests of the 3p-lognormal probability distribution of the survival times of the MMC patients given by the Kolmogorov-Smirnov, Anderson-Darling and Chi-square test. Each of the tests resulted in a large $p - value$, indicating that we fail to reject the null hypothesis, H_0 : the probability distribution follows the 3p-lognormal. In this section, we defined the probability density function (pdf) of the 3p-lognormal distribution and estimate its parameters. Given the survival time of MMC, t as a random variable, then the pdf of the 3p-log-normal probability distribution is given by,

$$f(t; \gamma, \mu, \sigma^2) = \begin{cases} 0, & \text{if } t \leq \gamma \\ (2\pi\sigma^2)^{-1/2}(t_i - \gamma)^{-1} \exp\left(-\frac{1}{2}\left(\frac{\ln(t_i-\gamma)-\mu}{\sigma}\right)^2\right), & \text{if } t > \gamma \end{cases} \tag{2.1}$$

where $\sigma > 0$ denotes continuous shape parameter, $-\infty < \mu < \infty$ represents the continuous scale parameter and γ is the continuous location parameter. If $\gamma \equiv 0$ yields the 2p-log-normal distribution. To estimate the three parameters σ, μ and γ, we would apply the maximum likelihood estimation (MLE) procedure. The MLE estimates the parameters of the probability distribution by maximizing the likelihood function. Generally, MLE is the most used parameter estimation method for statistical inference due to the robustness compared to other traditional methods the method of moment estimation; the logic is intuitive and flexible [32]. It has some unique and important statistical properties like consistency, invariant, efficiency, sufficiency and asymptotic normality. The MLE for three-parameter log-normal PDF may not be asymptotically efficient, it is still regarded as a better parameter estimation technique than the method of moments or quantiles because the variance is much smaller [33]. Therefore, the MLE is considered the best parameter estimation method for the 3p-log-normal probability distribution. Calitz (1973) suggested that the MLE procedure by Cohen (1951) should be adopted because it works better than other procedures. To compute the MLE estimators γ, μ, and σ^2, we first find the likelihood function. The likelihood function of the random sample of n independent observations of survival time $t = (t_1, ..., t_n)$ for the 3p-log-normal pdf $f(t)$ can be expressed as follows:

$$\begin{aligned} L(\gamma, \mu, \sigma^2 | t_i) &= \prod_{i=1}^{n} f(t_i | \gamma, \mu, \sigma^2) \\ &= \prod_{i=1}^{n} \left[(2\pi\sigma^2)^{-1/2} (t_i - \gamma)^{-1} \exp\left(-\frac{1}{2} \left(\frac{\ln(t_i - \gamma) - \mu}{\sigma} \right)^2 \right) \right] \\ &= (2\pi\sigma^2)^{-n/2} \prod_{i=1}^{n} (t_i - \gamma)^{-1} \exp\left(-\frac{1}{2} \sum_{i=1}^{n} \left(\frac{\ln(t_i - \gamma) - \mu}{\sigma} \right)^2 \right), \quad \forall t_i > \gamma. \end{aligned} \tag{2.2}$$

To find the estimates $\hat{\gamma}$, $\hat{\mu}$, and $\hat{\sigma}$ of γ, μ, and σ, Cohen (1951) obtained local maximum likelihood estimators (LMLE) and equate the partial derivative of the logarithmic likelihood function to zero. Thus, we have

$$\begin{aligned} \ell(\gamma, \mu, \sigma^2; t_i) &= \ln L(\gamma, \mu, \sigma^2 | t_i) \\ &= -\frac{n}{2} \ln(2\pi) - n \ln \sigma - \sum_{i=1}^{n} \ln(t_i - \gamma) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (\ln(t_i - \gamma) - \mu)^2. \end{aligned} \tag{2.3}$$

We find $\hat{\mu}$ by equating the partial derivative of ℓ with respect to μ to zero and solving for μ, we have

$$\begin{aligned}\frac{\partial \ell}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(\ln(t_i-\gamma)-\mu)=0,\\ \hat{\mu} &= \frac{1}{n}\sum_{i=1}^{n}\ln(t_i-\hat{\gamma}).\end{aligned} \tag{2.4}$$

We find $\hat{\sigma}$ by equating the partial derivative of ℓ with respect to σ to zero and solving for σ.

$$\begin{aligned}\frac{\partial \ell}{\partial \sigma} &= -\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=1}^{n}(\ln(t_i-\gamma)-\mu)^2=0,\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^{n}(\ln(t_i-\hat{\gamma})-\hat{\mu})^2,\\ \hat{\sigma} &= \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\ln(t_i-\hat{\gamma})-\hat{\mu})^2}.\end{aligned} \tag{2.5}$$

Similarly, we equate the partial derivative of ℓ with respect to γ and set it equal to zero, that is,

$$\frac{\partial \ell}{\partial \gamma} = \sum_{i=1}^{n}(t_i-\gamma)^{-1}+\frac{1}{\sigma^2}\sum_{i=1}^{n}(t_i-\gamma)^{-1}(\ln(t_i-\gamma)-\mu)=0. \tag{2.6}$$

Hill (1963) demonstrated that arbitrarily large likelihoods can be obtained by allowing $\hat{\gamma}$ to converge on $t_{(1)} = min(t_{(1)}, ..., t_{(n)})$, and $t_{(1)}, ..., t_{(n)}$ are ordered samples. Thus, the global maximum likelihood result to the inadmissible estimates $\hat{\gamma} = t_{(1)}$, $\hat{\mu} = -\infty$ and $\hat{\sigma} = +\infty$, regardless of the sample. However, Cohen (1951) [34] found the localized estimate $\hat{\gamma}$ to be sufficient in the identification of the 3p-log-normal. The validation of the estimates from such procedure has since been investigated Calitz (1973), Cohen and Whitten (1980) [35], and Chen (2006) [36] and others. Cohen's (1951) procedure estimated the LMLE for γ by substituting $\hat{\mu}$ and $\hat{\sigma^2}$ from Equation (2.4) and (2.5) into Equation (2.6) to obtain $\hat{\gamma}$. Thus, we have

$$\begin{aligned}&\left[\sum_{i=1}^{n}(t_i-\hat{\gamma})^{-1}\right]\left[\sum_{i=1}^{n}\ln(t_i-\hat{\gamma})-\sum_{i=1}^{n}(\ln(t_i-\hat{\gamma}))^2+\frac{1}{n}\left(\sum_{i=1}^{n}\ln(t_i-\hat{\gamma})\right)^2\right]\\&-n\sum_{i=1}^{n}(t_i-\hat{\gamma})^{-1}\ln(t_i-\hat{\gamma})=0.\end{aligned} \tag{2.7}$$

To solve for $\hat{\gamma}$ in Equation (2.7), we can solve iteratively the local maximum likelihood estimate (LMLE) of γ to obtain $\hat{\gamma}$. Here, we take into consideration admissible roots for which $\hat{\gamma}$ is below $t_{(1)}$. Cohen and Whitten (1980) found only one of such roots; however, in the event of multiple admissible roots, choose the root that results in the closest agreement between the sample mean $\bar{t}$ and the expected value of the 3p-log-normal probability distribution, $E(t)=\hat{\mu}_t=\hat{\gamma}+\exp(\hat{\mu}+\hat{\sigma}^2/2)$.

Base on the procedure for the parameter estimation of the three-parameter log-normal probability distribution discussed above, the 3p-log-normal probability distribution of the survival time t of multiple myeloma ($\hat{\gamma}$, $\hat{\mu}$, $\hat{\sigma}$), the approximate estimates are given in Table 2.4 below.

Table 2.4: Parameter Estimates for the Three-Parameter Lognormal Probability Distribution of the Survival Time of Multiple Myeloma

Location ($\hat{\gamma}$)	Scale ($\hat{\mu}$)	Shape ($\hat{\sigma}$)
-0.17824	2.7015	1.0429

We substituted the estimates into Equation (2.1) to obtain the probability density function (pdf) of the 3p-log-normal distribution of the survival time of multiple myeloma given by,

$$f(t) = \begin{cases} 0, & \text{if } t \leq -0.17824 \\ 0.38253(t+0.17824)^{-1}\exp\left(-\frac{1}{2}\left(\frac{\ln(t+0.17824)-2.7015}{1.0429}\right)^2\right), & \text{if } t > -0.17824. \end{cases} \tag{2.8}$$

The above findings that the survival times of multiple myeloma patients data follows three-parameter log-normal distribution can ensure efficient and accurate analysis of the survival times of MMC patients. Given the pdf in Equation (2.1), we find the cumulative density function (cdf) by taken the integral of the pdf with respect to the random variable t, given by:

$$\begin{aligned} F_T(t;\gamma,\mu,\sigma^2) &= \int_0^t f_T(s|\gamma,\mu,\sigma^2)ds = P(0 \leq t \leq t)E(t|0 \leq t \leq t) \\ &= \int_0^t (2\pi\sigma^2)^{-1/2}(s_i-\gamma)^{-1}\exp\left(-\frac{1}{2}\left(\frac{\ln(s_i-\gamma)-\mu}{\sigma}\right)^2\right)ds \\ &= \int_0^t \frac{1}{\sigma(s_i-\gamma)\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln(s_i-\gamma)-\mu}{\sigma}\right)^2\right)ds. \end{aligned} \tag{2.9}$$

Given the tediousness in integrating the 3p-lo-normal pdf, we adopted the method of integration by substitution; substituting into Equation (2.9) the standard normal cdf given by

$$\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_0^t \exp\left(-\frac{1}{2}z^2\right)dz.$$

In Equation (2.9), we substitute $z = (\ln(t_i-\gamma)-\mu)/\sigma$ and the remaining part $1/\sigma(t_i-\gamma) \approx 1$. Hence, the 3p-log-normal cdf of the survival times can be expressed as

$$F_T(t;\gamma,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}}\int_0^t \exp\left(-\frac{1}{2}z^2\right)dz = \Phi\left(\frac{\ln(t_i-\gamma)-\mu}{\sigma}\right). \tag{2.10}$$

where $\Phi(.)$ represents the standard normal cdf at a given survival time t_i. Substituting the estimates of Table 2.4 into Equation (2.10), we have

$$F_T(t;\gamma,\mu,\sigma^2) = \Phi\left(\frac{\ln(t_i + 0.17824) - 2.7015}{1.0429}\right). \tag{2.11}$$

$\Phi(.)$ denotes the standardized normal cumulative probability distribution. The cdf can be useful in determining the probability of a given random observation (survival time t) would be less than or equal to some value T; thus, $\mathbf{P}(t \leq T)$. Figure 2.4, shows the cdf plot of the survival times of multiple myeloma patients data.

For example, we can estimate the probability that a patient with MMC survives up to time $t = 16$ months is approximately 0.53.

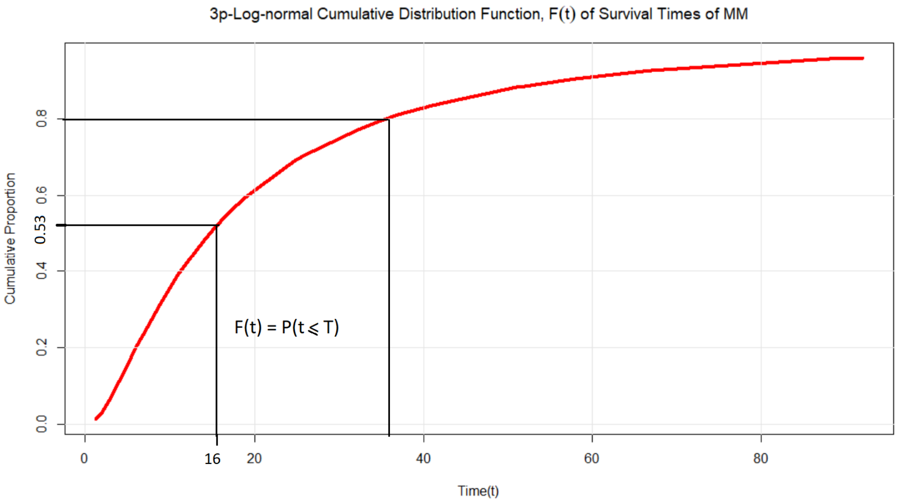

Figure 2.4: Cumulative Distribution Function Plot for the Survival Time of MM

Now, given the cdf of the survival times of MMC patients in Equation (2.12), we obtained the reliability of the survival time t of MMC patients, given by

$$\begin{aligned}\hat{S}(t_i;\gamma,\mu,\sigma^2) &= 1 - F_T(t;\gamma,\mu,\sigma^2)\\ &= 1 - \Phi\left(\frac{\ln(t + 0.17824) - 2.7015}{1.0429}\right).\end{aligned} \tag{2.12}$$

The survival function can be use to estimate the probability that a patient diagnosed with multiple myeloma would survive beyond time T; thus, $P(t > T)$. In Figure 2.5, we display the estimate of the survival function $\hat{S}(t)$ of the survival times of MMC patients for the 3p-log-normal. As expected, we can see that the survival function of the survival times is decreasing and approximately zero beyond time $t = 80$. For instance, the probability that a patient survives beyond 20 months is approximately 0.40.

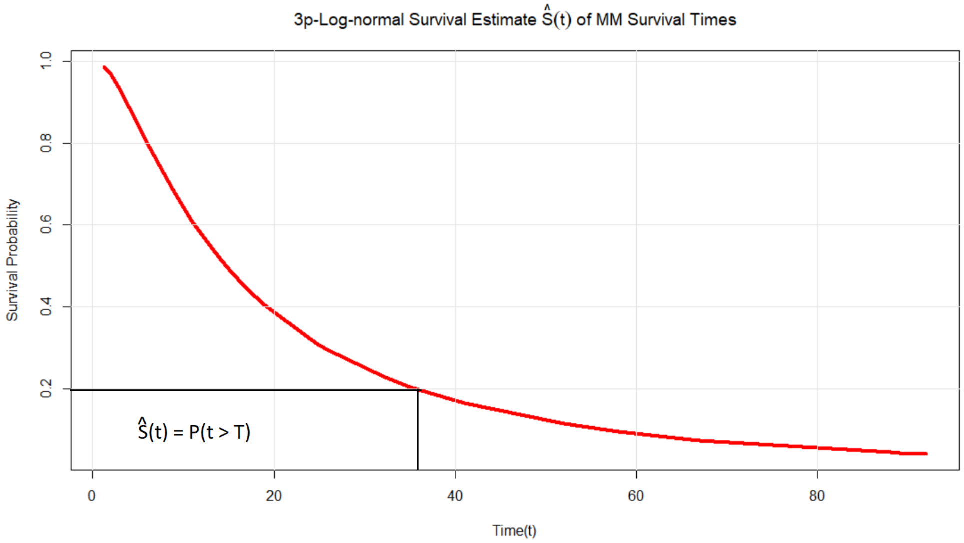

Figure 2.5: Survival Estimate for the Survival Time of MMC

2.4 Kaplan-Meier Estimation of Survival Probability of the Survival Time of Patients with Multiple Myeloma

Kaplan-Meier estimate (KM) was introduced by Edward L. Kaplan and Paul Meier (1958) [37], which is a non-parametric analytical tool. KM defined as the probability of survivorship at a given length of time called the survival time. Graphical representations of KM estimates are used to determine the events, censoring, and survival probability. Another name for KM estimator is the product-limit estimator which estimates the proportion of survival beyond a particular time t, given that some of the observed units may not die or fail. The KM survival estimate $S\hat{(t)}$ of MMC patients can be said to represent the reliability estimate $\hat{R}(t)$ of MMC patients obtained by taking the product of the conditional probability of surviving at time t_{i+1}, given that a patient survived until time t_i. We estimate the survival function $\hat{S}(t)/\hat{R}(t)$ or the probability that observation last longer than time t as

$$\hat{R}(t) = \hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{\tau_i}{n_i}\right),$$

where t_i is the time when at least one event happened, τ_i denotes the number of events (e.g. deaths or alive) at time t_i, n_i represents the individuals/observations at risk (not yet had an event or censored) up to time t_i, and τ_i/n_i denotes the probability of survival. In survival analysis, a nearly universal feature data is the *censoring data*, the commonest of which is the *right-censoring* where an individual expires or removed before the end of the study or clinical trial. The other cases of censoring which rarely happen are the *left-censoring* (the initial time at risk is unknown or individuals removed at the start of the study), and *interval-censoring* (a case of both right and left censored). A key advantage of the KM curve is that it can take into account censored data, particularly right-censoring, and is easy to estimate. The analysis of Kaplan-Meier survival curve takes into consideration the following *three assumptions*: (1) it assumes that censored observations have the same prospects of survival as those who continue in the study, (2) it assumes survival probabilities to be the same for individuals recruited early and late in the study, and (3) it assumes that the event happens at the specified time. We utilized KM to assess the overall survival of MMC. In Figure 2.6, we show the KM curve for the global estimates of the probability of survival times of patients diagnosed with multiple myeloma with a 95% confidence interval. We can see that all the MMC patients demised by time 92 (in months).

Though the KM estimator is the most commonly used technique of survival analysis, and is useful for examining recovery rates, the probability of death and effectiveness of treatment. However, as we mentioned before the KM method is not as powerful as the parametric survival analysis for decision making. We cannot obtain the hazard function to estimate the rate at which patients die with MMC using KM.

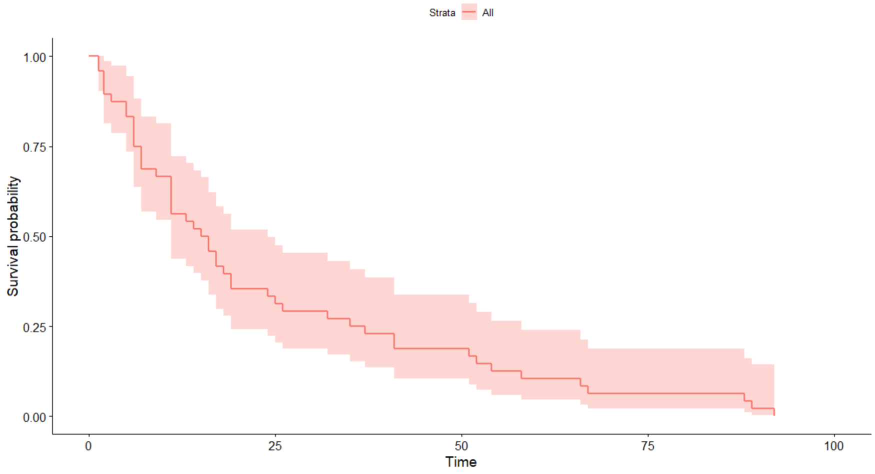

Figure 2.6: Kaplan Meier Global Estimates of the Survival Time with CI

2.5 Comparison of Three-Parameter Lognormal with the Kaplan Meier Estimation of the Survival Function.

In the parametric analysis, we found the survival times of patients with MMC follow the three-parameter lognormal probability distribution. In section 2.4, we performed a non-parametric analysis using the Kaplan-Meier to estimate the survival probability of the MMC patients. Now, we compare the survival estimate of the 3p-lognormal with the Kaplan-Meier survival estimate of the survival times of the MMC patients. The relevance of the survival function of the two methods is to estimate the survival probability of a patient diagnosed with MMC beyond a given time, after a given treatment. The survival times along with the survival probabilities of the two survival functions is given by Table 2.5. We see that the probability estimates by the 3p-lognormal survival function are mostly higher than that of Kaplan-Meier. Though there are times in which the KM estimates higher probabilities, the 3p-log-normal survival function from the parametric probability distribution estimates the survival proportion more accurately than the Kaplan-Meier. Thus, because parametric methods are more powerful, robust and efficient than non-parametric methods.

Table 2.5: Kaplan-Meier ($\hat{S}_{KM}(t)$) Vs Parametric (3P Lognormal, $\hat{S}_P(t)$) Survival Function Estimate of the Survival Times

t_i	$\hat{S}_{KM}(t)$	$\hat{S}_P(t)$	t_i	$\hat{S}_{KM}(t)$	$\hat{S}_P(t)$
1.25	0.958	0.988	25.00	0.313	0.308
2.00	0.896	0.967	26.00	0.292	0.295
3.00	0.875	0.931	32.00	0.271	0.230
5.00	0.833	0.845	35.00	0.250	0.205
6.00	0.750	0.801	37.00	0.229	0.190
7.00	0.688	0.758	41.00	0.188	0.164
9.00	0.667	0.679	51.00	0.167	0.118
11.00	0.563	0.609	52.00	0.146	0.115
13.00	0.542	0.547	54.00	0.125	0.108
14.00	0.521	0.519	58.00	0.104	0.098
15.00	0.500	0.493	66.00	0.083	0.076
16.00	0.458	0.469	67.00	0.063	0.074
17.00	0.417	0.446	88.00	0.042	0.044
18.00	0.396	0.424	89.00	0.021	0.043
19.00	0.354	0.404	92.00	0.000	0.040
24.00	0.333	0.321			

2.6 Discussion

Given the danger posed by multiple myeloma cancer in recent years, it is imperative to investigate the prognosis and how to enhance the therapeutic/treatment strategy of MMC. While MMC currently remains incurable, there has been some improvement in the treatment process. The treatment progress has been largely due to the introduction of therapeutic agents such as thalidomide, lenalidomide, and bortezomib, as well as the introduction of high-dose chemotherapy and stem-cell rescue (ASCT). The quest to improve the survival time (or increase the death time) of patients with MMC has resulted in adopting several different research approaches and methodologies.

In the present chapter, (1) we identified a well-defined probability distribution that characterizes the survival times of the 48 patients diagnosed with MMC and used it to estimate the survival function. (2) we estimated the proportion of survival time utilizing the commonly used Kaplan-Meier (KM) technique of analysis of cancer survivorship. (3) we compare and contrast the relevance and efficiency of the two different methods of analysis of survival probability estimation of patients diagnosed with MMC beyond a given survival time. Firstly, we investigated utilizing the log-rank test to test the difference between the survival times of the males and females diagnosed with MMC. We found that the survival times of both males and females diagnosed with MMC are not different, so we performed the analysis using the combined data of the males and females. We then found a well-defined probability distribution that characterizes the survival time of the 48 patients diagnosed with MMC follows the 3p-log-normal probability distribution. We believe that being able to find the best probability distribution that characterizes the probabilistic behavior for a given cancer patient survival time can lead to estimating the survival probability with much more accuracy and efficiency. The fact that we found a unique probability distribution for our study of the survival times of patients diagnosed with MMC refute the suggestion of assuming exponential distribution (as suggested by Feigl and Zelen ([1965] p. 835) and other authors) or using the non-parametric Kaplan-Meier for most cancer survivorship studies. We found both the 3p-log-normal survival most often estimates higher survival probability than the KM survival function, given by Table 2.5.

We know that KM is the most commonly used technique for analyzing cancer survivorship data. Statistically, the parametric technique is said to be more robust and efficient than the non-parametric counterpart. Therefore, our finding of the parametric 3p-log-normal probability distribution is better and of a higher degree of accuracy in estimating the survival probability of the patients diagnosed with MMC than the Kaplan-Meier. The KM technique is more often used to compare the difference between the survival probability of the survival times of two or more entities or groups usually based on the log-rank test. However, by obtaining the best parametric probability distribution that characterizes the survival times, we can find the survival function and estimate the survival rate and compare the results of two or more entities with a high degree of accuracy. The only situation where non-parametric becomes highly applicable or recommended is when there is no parametric form (i.e. pdf) for the given data. One key demerit of KM is that it cannot be used to estimate the rate at which patients die or survive with MMC (i.e. the hazard function). The hazard function can be easily calculated after finding the parametric probability distribution by dividing the probability density function, *pdf* by the survival function.

2.7 Contribution

We estimated the survival probability of patients diagnosed with multiple myeloma cancer using two different methods; the parametric three-parameter log-normal and the non-parametric Kaplan-Meier. We found the parametric method to often give a higher estimate of the survival probability than the non-parametric method. Even though there are times when the non-parametric survival probability estimates are higher, all-important arguments favor the parametric method. The difficulty of the parametric survival analysis is the fundamental intrinsic assumption that the survival times of the population under study follow a specific probability distribution. But if such a limitation can be overcome, then we can obtain a more robust or powerful result from the parametric analysis. We can also estimate the hazard function to assess the rate at which patients die with MMC after finding the right parametric distribution.

Base on the two different methods utilized for estimating the probability of survival of patients diagnosed with MMC, we offer the following essential recommendations. (1) If the only information about patients is the survival (death) time, then estimating the survival probability using the parametric technique will yield more accurate, robust, and efficient results than the commonly used Kaplan-Meier. (2) However, if no unique or well-defined parametric probability distribution is found, then we still recommend the use of Kaplan-Meier technique for the estimation of the survival probability. Though the use of non-parametric Kaplan-Meier survival analysis may, in some cases result in a similar or higher estimate of the survival rate (such as in our case), the parametric analysis remains more powerful, robust and efficient, and hence must be considered first in the analysis of any given cancer survivorship data.

Chapter 3: Survival Analysis of Multiple Myeloma Cancer (MMC) Using the Cox-Proportional Hazard Model.

Though multiple myeloma cancer (MMC) remains incurable, research into improving the therapeutic strategy has increased dramatically in recent years. In this chapter, the semi-parametric Cox proportional hazard model was employed to examine the survival probability taking into consideration the sixteen risk factors in Table 2.1 in Chapter 2 presumed to be contributing to the survival times of MMC patients. The chapter provides a much-improved method for investigating the survival times of MMC patients. A careful and rigorous assessment of the risk factors resulted in finding seven risk factors and one interaction term statistically significantly contribute to the proportion of survival times of MMC. Four of the seven risk factors and the interaction term are new significant contributing risk factors to the survival time of MMC identified by our proposed model, namely; platelets, gender, white blood cells, and fractures; and the interaction term, infections and serum calcium. The final Cox-PH model was well-validated and satisfied the key assumptions. The identified risk factors are rank according to the prognostic effect on the survival time based on the hazard ratio. Blood urea nitrogen (BUN)/serum creatinine was the greatest prognostic factor (most contributing factor, and highly negatively related to the MMC deaths or survival times), followed by white blood cells (WBC), and normal platelet was found to be the minimum prognostic factor (least contributing factor to MMC death or survival times). This study offers prognostic and therapeutic significance for further enhancement in the treatment strategy of multiple myeloma cancer disease. The research findings of this chapter have been published [61].

The organization of this chapter is as follows: Section 3.1 introduces and review some literature of studies on MMC; Section 3.2 reviews the Cox-PH model; Section 3.3 presents the proposed Cox-PH model; Section 3.4 discusses the findings in this study; and finally, the research contributions of this chapter is given in Section 3.5.

3.1 Introduction

There are no major findings of what specifically causes multiple myeloma cancer (MMC), given that the disease remains incurable.Most risk factors of MMC are reported to be common among the age, males, black race, and families with MMC history [4, 18]. However, research has discovered several risk factors presumed to have some relation with the duration of the survival of patients with MMC [20, 39]. Most of these factors were identified at the time a patient was diagnosed with the MMC disease. Some common risk factors identified through clinical trials and research studies included hemoglobin, immunoglobulin type, extent and type of lesions, serum calcium, age, sex, white blood cells, blood urea nitrogen, serum calcium, serum albumin, infections, platelets, hemoglobin, presence of Bence Jones protein, and performance status, all at the time patients were diagnostic with MMC, are believed to have contributed to the survival of patients with MMC [39, 40, 41].

A struggle to find a lasting solution or treatment to the incurable MMC has resulted in research into some statistical analysis on the survival of patients with MMC given the event that a patient died or survived. Kaplan-Meier technique has been commonly used for analyzing cancer survivorship data in recent times due to the simplicity of its usage. It is often used to compare the survival difference of observations/groups base on the log-rank test. KM is mostly used for longitudinal studies like a cohort study [19]; an example to the present study (i.e. the survival time of patients diagnosed with multiple myeloma). The disadvantage of using KM is that it does not take into consideration the risk factors (covariate) contributing to the length of patients' survival duration of the MMC disease, hence, nullifying the relevance of KM if risk factors are contributing in the given survival data.

Brain et al [20] used Kaplan-Meier to test whether there was a significant difference in the survival duration between the categories of risk factors based on the generalized Wilcoxon test and the log-rank test. They further used a non-linear Cox regression to ascertained the combination of patients' characteristics relative to survival duration. They identified a significant difference in the survival duration among patients based on performance status, cell mass and percentage labeling index, Nephrotic status, Hemoglobin, age, and k/λ subtype. John M. Krall et al [23] developed a set-up procedure for selecting variables associated with the survival times of patients with MMC utilizing the data used in the present study. They found blood urea nitrogen, hemoglobin, percent plasma cell in bone marrow, and Serum calcium to be associated with the survival of patients with multiple myeloma. Shaji K. Kumar et al found continued improvement in survival in MMC with changes in early mortality and outcomes in older patients. Giampaolo Merlini, Jan G. Waldenstrom, and Suresh D. Jayakar [29] proposed a new improved clinical staging system for the survival of MMC based on analysis of 123 treated patients. In their findings, serum calcium, % bone myeloma plasma cell (% BMPC) and serum creatinine/BUN were significantly related to the survival of IgG myeloma stage; hemoglobin, serum calcium, and M-component related significantly with the survival of IgA myeloma stage; and creatinine/BUN, % BMPC and serum calcium to be related significantly with the survival of BJ myeloma stage; but no significant relation to survival with age or sex. Our study found five new significant attributable out of the sixteen risk factors presumed to be contributing to the survival times of MMC. They are platelets, gender, white blood cells, fractures, and an interaction term between infections and serum calcium. In most of the research studies, either one or two of the five newly identified risk factors were analyzed, but not found significant or not part of the data analyses. We studied the semi-parametric Cox-PH survival analysis of the survival times to estimate the survival rate of patients diagnosed with multiple myeloma. We utilized the Cox-PH model to analyze the proportion of survival time, taking into account the 16 risk factors, Table 2.1 of Chapter 2, considered to be contributing to the survival time of the patients diagnosed with MMC. Thus, we assessed the relationship between the proportion of survival time as a function of 16 attributable risk factors and two-way interactions based on the Cox proportional hazard (PH) model. The significant attributable risk factors identified were carefully investigated and selected based on the stepwise model selection method, with the final model representing the model with the least AIC. The final Cox-PH model was validated to satisfy all the key assumptions, and no presence of multicollinearity measured based on the variance inflation factor (VIF).

3.2 Review of the Cox Proportional Hazard Model.

In survival analysis, two things are of utmost importance; time and event. Thus, survival analysis models the time an event occurred called the *survival time.* For example, the time a patient died of MMC. The survival time can be associated/influenced by one or several attributable factors/risks, often termed as *covariates* by most survival analysis literature. Cox proportional hazard model is also known as the Cox model, introduced by Cox (1972) has been widely recommended for semi-parametric modeling of the relationship of the survival time as a function of the covariates in survival analysis. A good basic review of the introduction and methodology is given by Kleinbaum [46], and more extensive discussions have been provided by Kalbfleisch and Prentice [47]. We are given a brief review of the Cox proportional hazards model in this section. An important aspect of the Cox model is the hazard function. The hazard function measures the rate of death at time t. We define the hazard function as follows; Let random variable T denote the survival time with cumulative density function $F_T(t)$, given by

$$F_T(t) = P(T \leq t) = \int_0^t f_T(t)dt.$$

Thus, $F_T(t)$ is the probability of failure by time t and $f_T(t) = dF_T(t)/dt$ is the probability density function. The survival function is defined as

$$S_T(t) = P(T > t) = 1 - P(T \leq t) = 1 - F_T(t).$$

Therefore, the hazard function which examines the risk of instantaneous death at time t, is conditional on the survival function defined by

$$\begin{aligned} h(t) &= \lim_{\partial t \to 0} \frac{F_T(t + \partial t) + F_T(t)}{\partial t . S_T(t)} \\ &= \lim_{\partial t \to 0} \frac{P(t < T \leq t \ + \partial t)}{\partial t . S_T(t)} \\ &= \lim_{\partial t \to 0} \frac{P(t < T \leq t \ + \partial t | T > t)}{\partial t} \\ &= \frac{f_T(t)}{S_T(t)}. \end{aligned} \tag{3.1}$$

From the hazard function given by Equation (3.1), we can obtain the cumulative hazard function, expressed as

$$H(t) = \int_0^t h(s) ds.$$

The integral can be expressed in close form as $H(t) = -\ln S(t) = -\ln R(t)$.

The Cox model which includes interacting covariates is expressed by the hazard function, estimated as follow:

$$h_i(t) = \alpha_i(t) \exp\left(\sum_{i=1}^{k} \beta_i X_i + \sum_{i \neq j = 1}^{k} \rho_{ij} X_i X_j \right),$$

and

$$\ln\left(\frac{h_i(t)}{\alpha_i(t)} \right) = \sum_{i=1}^{k} \beta_i X_i + \sum_{i \neq j = 1}^{k} \rho_{ij} X_i X_j \tag{3.2}$$

where t is the survival time, $h_i(t)$ is hazard function obtained by the set of k covariates, β_i is the coefficients measuring the impact of the covariates X_i on $h_i(t)$, ρ_{ij} is the coefficient measuring the impact of interacting covariates X_iX_j on $h_i(t)$, $\alpha(t)$ is the baseline value of $h_i(t)$ if all X_i and X_iX_j equals zero. The Cox model is a multiple linear regression of the logarithmic form of the hazard on X_i's and X_iX_j's, with $\alpha(t)$ as an intercept that varies with time t. A major assumption of the Cox model is the proportional hazard assumption, which explains that the hazard function of observations (or patients) should be proportional and independent of time t [45]. Consider the case of two patients i and i' with varying values of covariates; the corresponding hazard functions for i^{th} patient is

$$\eta_i(t) = \alpha(t)\exp\left(\sum_{i=1}^{k}\beta_iX_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right),$$

and the corresponding hazard functions for i'^{th} is

$$\eta_i'(t) = \alpha(t)\exp\left(\sum_{i'=1}^{k}\beta_{i'}X_{i'} + \sum_{i'\neq j'=1}^{k}\rho_{i'j'}X_{i'}X_{j'}\right).$$

The hazard ratio of the two patients is

$$\begin{aligned}\frac{\eta_i(t)}{\eta_i'(t)} &= \frac{\alpha(t)\exp\left(\sum_{i=1}^{k}\beta_iX_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)}{\alpha(t)\exp\left(\sum_{i'=1}^{k}\beta_{i'}X_{i'} + \sum_{i'\neq j'=1}^{k}\rho_{i'j'}X_{i'}X_{j'}\right)}\\ &= \frac{\exp\left(\sum_{i=1}^{k}\beta_iX_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)}{\exp\left(\sum_{i'=1}^{k}\beta_{i'}X_{i'} + \sum_{i'\neq j'=1}^{k}\rho_{i'j'}X_{i'}X_{j'}\right)} = \exp(coef),\end{aligned} \tag{3.3}$$

which is independent of time t. The consequence of the above hazard ratio implies that the Cox model is a proportional-hazards model. The hazard ratio is a relative measure of the hazards between observations/groups [44]. We interpret the hazard ratio (HR) in the following three ways: (1) $HR = 1$, implies that there is no hazard effect. Thus, the covariates have no relationship with the event probability, hence, no influence on the length of survival. (2) $HR > 1$ (i.e. equivalently $\hat{\beta}_i > 0$), implies an increase in hazard. That is, the covariates have a positive association with the event probability, hence, a negative association with the length of survival (bad prognostic factor). (3) $HR < 1$ (i.e. equivalently $\hat{\beta}_i < 0$), implies a decrease in hazard. That is, the covariates are negatively associated with the probability of the event, hence, positively associated with the length of survival (good prognostic factor). A comprehensive review of the hazard ratio have been provided by L. Douglas Case et al [48].

To compute the baseline hazard function, we performed the following computation:

$$\hat{\alpha}(t) = \sum_{t_i \leq t} \hat{h}(t_i),$$

with

$$\hat{h}(t_i) = \frac{d_i}{\sum_{i \in R(t_i)} \exp(X_i' \hat{\beta})},$$

where $t_1 < t_2 < ... < t_n$ denote the distinct event times, d_i is the number of events at t_i, and $R(t_i)$ is the risk set at t_i containing all individuals still susceptible to the event at t_i. The base line hazard function can assume any functional form of the covariates. In section 3.2.1, we discussed in detail the major assumptions of the Cox-PH model. We will show that the assumptions are satisfied once we have developed the Cox-PH model for the given data.

3.2.1 Cox-Proportional Hazards (PH) Model Assumptions.

A good Cox proportional hazard model should satisfy the following three key assumptions, prior to its implementation. Failure to satisfy the assumptions will lead to wrong decision about the subject matter.

1. Proportional hazard (PH) assumption.

   The PH assumption of the Cox model can be assessed based on formal statistical tests. A non-statistical significance of the covariates and the global test is an indication that the PH assumption is valid. Another method to check for the PH assumption is by investigating the plot of scaled Schoenfeld residuals against the transformed time. The Schoenfeld residuals are independent of time, a non-random pattern against time is evidence of a violation of the PH assumption. We calculate the Schoenfeld residuals with one per observation per covariate. This can be expressed as

$$r_{ik} = X_{ik} - \hat{\hat{X}}_{w_i k}(\beta, t_i),$$

where X_{ik} denotes the value of the k^{th} covariate for i^{th} observation. $\hat{\hat{X}}_{w_i k}(\beta, t_i)$ represents the weighted mean values of covariates at risk at the given event time, t_i, denoted by $R(t_i)$, and given by

$$\hat{\hat{X}}_{w_i k}(\beta, t_i) = \sum_{j \in R(t_i)} X_{ik} w_i(\beta, t_i).$$

The weight function, $w_i(\beta, t_i)$ for i^{th} observation at risk, $R(t_i)$ is the probability that observation i fails at time t_i, defined by

$$w_i(\beta, t_i) = \frac{\exp(\beta^T X_i)}{\sum_{l \in R(t_j)} \exp(\beta^T X_l)}.$$

A positive value of r_{ik} depicts an X value higher than expected at that death time. For a binary (0,1) variable, Schoenfeld residuals will be between -1 and 1. In that situation,

$$r_{ik} = \begin{cases} 0 - \hat{\hat{X}}_{w_i k}, & \text{for } X = 0 \\ 1 - \hat{\hat{X}}_{w_i k}, & \text{for X =1.} \end{cases}$$

2. Linear functional form of continuous covariates

We assume that the functional form of the covariates are linear. T. Therneau and P. Grambsch suggested this assumption can be checked by visualizing the plot of Martingale residuals against the continuous covariates with fitted lowess (locally weighted smoothing) line function. A trend or pattern in the plot is evidence of a violation of the linear functional form of the covariates. Martingale residual is defined by

$$\hat{M}_i = \delta_i - \hat{\Gamma}_0(t_i)\exp(\hat{\beta}_1 X_{1i} + ... + \hat{\beta}_k X_{ki}),$$

where δ_i denotes the event indicator for i^{th} observation, $\hat{\Gamma}_0(t_i)$ is the estimated cumulative hazard at the final follow-up time for i^{th} observation, and $\exp(\hat{\beta}_1 X_{1i} + ...)$ is the estimated coefficients applied to the observed covariate for the i^{th} observation. Martingale residuals, $\hat{M}_i$, have a skewed distribution. The $\hat{M}_i$ values are

$$\hat{M}_i = \begin{cases} 1, & \text{for maximum possible values} \\ -\infty, & \text{for minimum possible values.} \end{cases}$$

A positive Martingale residual value implies individuals demised too soon, negative value implies individuals lived too long. A transformation of $\hat{M}_i$ to obtain approximate symmetric distribution can be essential. Such a transformation is motivated by deviance residuals defined below.

3. Examining influential observations (or outliers)

In examining influential observations, we visualized the *dfbeta* values. The *dfbeta* values estimates the influence of $i^{th} - case$ (or observation) on the regression coefficients. A very high value of *dfbeta* should be closely investigated. Another technique for checking influential observations is by assessing the deviance residuals (symmetric/normalized transformation of the Martingale residuals) plot. The deviance residual is defined by

$$d_i = sin(\hat{M}_i)\sqrt{2}\sqrt{-\hat{M}_i - \delta_i log(\delta_i - \hat{M}_i)}.$$

Note that $d_i = 0$ is only when $\hat{M}_i = 0$. The square root shrinks the large negative martingale residuals, while the logarithm transformation expands those residuals that are close to zero (i.e as $n \to \infty$). The distribution of the residuals should be roughly symmetrical about zero mean and standard deviation of one. A very large/small/distant deviance residual values indicate influential observations or outliers. The values of the deviance residual values can be compared with the expected value of survival time. A positive value implies individuals demised too soon, negative value implies individuals lived too long.

Now, we proceed to develop the Cox-PH model for the survival times of multiple myeloma patients. After we develop the model, we will verify that the above three assumptions are met to validate the applicability and quality of the proposed model.

3.3 Proposed Cox-PH Model for the Survival Times of Patients with MMC

We started by fitting the Cox-PH model to the survival times t as a function of all 16 covariates X_i together with their two-way interactions. A stepwise model selection method was adopted to select the final model with the least Akaike information criterion ($AIC = 2ln(L) + 2k$, where L is the value of the maximum likelihood function of the model and k represents the estimated model parameters) [42]. AIC gives an estimation of the relative amount of information missing in the model; hence, the smaller the AIC value the better the quality of the model. It deals with the danger posed by overfitting or under-fitting the model.

The stepwise model variable selection procedure is one of the best ways used for determining significant covariate for Cox-PH models. It is based on iterations between forwarding and backward steps. All covariates and their interactions are included to be part of the "variable list" for selection. The significance levels for entry (α_{entry}) and stay (α_{stay}) are suggested to be set at 0.15 or larger for being conservative. Then, the best Cox-PH model is obtained by manually removing the covariates with $p - value > 0.05$ one at a time until all model coefficients are statistically significant at the chosen level of significant, $\alpha = 0.05$. The final model with all significant covariates and possible interactions is the model with the least AIC value. Hence, based on the stepwise model selection procedure criteria, our final proposed model that significantly contributes to the probabilistic survival time of patients diagnosed with multiple myeloma includes seven significant contributable covariates (risk factors) and one interaction; given by

$$\begin{aligned} \ln\left(\frac{h_i(t)}{\alpha_i(t)}\right) &= 2.008X_1 - 1.608X_3 normal - 0.815X_6 female \\ &+ 1.878X_7 + 0.854X_8 present + 0.108X_{12} \\ &+ 1.576X_{13} none + 0.114X_4 present.X_{16}). \end{aligned} \tag{3.4}$$

The Table 3.1, below displays the estimates of the model coefficients/parameters, their hazard ratios (HR) ($\exp(\hat{\beta})$), standard error of coefficients, statistical significance, and 95% confidence interval. The significant contributing coefficients or risk factors have been ranked based on the prognostic effect to the survival times of patients diagnosed with MMC using the hazard ratio (HR). Thus, we ranked from the most contributing factor to the least contributing factor to the death or survival times of MMC patients. The covariates, $X_i's$ are defined in Table 2.1 in Chapter 2.

Table 3.1: Ranking of the Significant Contributing Covariates (Risk Factors) Base on Prognostic Effect to the Survival Time Using the Hazard Ratios

Rank	Covariates	Coeff ($\hat{\beta}$)	HR ($\exp(\hat{\beta})$)	S.E.($\hat{\beta}$)	Pr(> $\|z\|$)	lower .95	upper .95
1	X$_1$	2.008	7.454	0.619	1.165e^{-3**}	2.217	25.056
2	X$_7$	1.878	6.543	0.773	1.505e^{-2*}	1.439	29.745
3	X$_{13}$	1.576	4.835	0.418	1.63e^{-4***}	2.131	10.972
4	X$_8$	0.854	2.349	0.409	3.693e^{-2*}	1.053	5.243
5	X$_4$: X_{16}	0.113	1.121	0.040	4.873e^{-3**}	1.035	1.213
6	X$_{12}$	0.108	1.114	0.030	3.84e^{-4***}	1.049	1.183
7	X$_6$	-0.815	0.443	0.391	3.711e^{-2*}	0.206	0.952
8	X$_3$	-1.608	0.200	0.502	1.355e^{-3**}	0.075	0.536

In Table 3.1, we tested the statistical significance of each of the chosen risk factors and interaction (coefficients) in the Equation (3.4) based on the $p - value$ from Wald statistic

Table 3.2: Global statistical significance of the model

Type of test	Test statistic	df	$P - value$
Likelihood ratio test	32.6	8	$7e^{-05}$
Wald test	30.38	8	$2e^{-04}$
Score (log-rank) test	32.49	8	$8e{-}05$

Table 3.3: First Ten Baseline Hazard Estimates

Obs.	Baseline Hazard	Time
1	$2.498e^{-06}$	1.25
2	$1.463e^{-05}$	2.00
3	$1.951e^{-05}$	3.00
4	$3.162e^{-05}$	5.00
5	$6.104e^{-05}$	6.00
6	$8.818e^{-05}$	7.00
7	$9.806e^{-05}$	9.00
8	$1.549e^{-04}$	11.00
9	$1.676e^{-04}$	13.00
10	$1.818e^{-04}$	14.00

value. All the selected risk factors are tested significant, with "three stars ***" indicating a very highly statistically significant risk factor. A positive coefficient ($\hat{\beta} > 0$) means a higher hazard rate, and thus a bad prognostic factor. By contrast, a negative coefficient ($\hat{\beta} < 0$) means a lower hazard rate, and thus a good prognostic factor. For instance, $\hat{\beta}_6 = -0.815$ representing gender implies that females are good prognostic of the survival time of MMC; thus, females have a lower risk of death (higher survival rates) of MMC than males. The $\exp(\hat{\beta})$ is the hazard ratio measures the size of the effect of the risk factor. Thus, $\exp(-0.815) = 0.443 < 1$ for gender means being a female has a reduced risk of dying with MMC than being a male. The ranking of the significant attributable risk factors based on the HR shows that blood urea nitrogen (BUN)/serum creatinine (X_1) is the greatest prognostic factor to the survival of MMC, followed by white blood cells (WBC) (X_7), and platelets (X_3) is the least prognostic factor. Table 3.2 displays three different tests for the overall significance of the proposed Cox model; the likelihood-ratio test, the Wald test, and score log-rank statistics. The three tests are asymptotically equivalent and give similar results for large samples. However, for small samples like in our case, the likelihood ratio test is robust and generally preferred. The global statistical significance test demonstrates that the proposed Cox-PH model in Equation (3.4) is highly statistically significant. In Table 3.3, we displayed the baseline hazard function $\hat{\alpha}(t)$ for the first ten observations. Figure 3.1, below is a graphical display of the results given in Table 3.3 according to the order of prognostic effect based on the hazard rate. Clearly, we can see that blood urea nitrogen (X_1) is the greatest prognostic factor to the survival time of MMC patient, followed by white blood cells (X_7), and platelets (X_3) is the least prognostic factor.

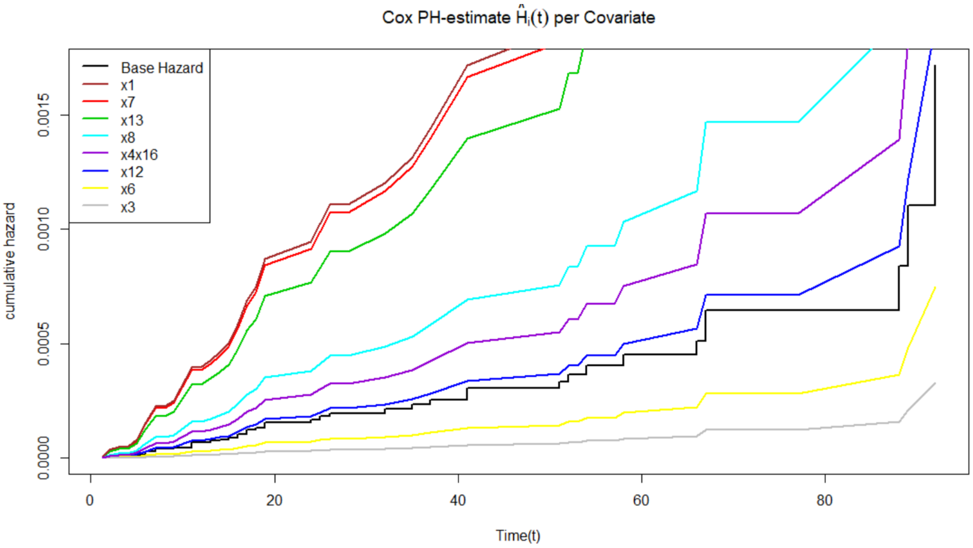

Figure 3.1: Ranking of Prognostic Effect of Risk Factors

3.3.1 Validation of the Proposed Cox-PH Model

We validated the goodness-of-fit of the proposed Cox-PH model by satisfying the three major Cox-PH model assumptions outlined in section 3.2.1. Firstly, we verified that the proportional hazard assumption is satisfied. Figure 3.2, shows the plot of the scaled Shoenfeld residual against time. It shows that there is no pattern as a function of time. Thus, the residuals are randomly scattered with no systematic departures from the horizontal fitted smoothing spline deep line (i.e. the residuals are independent of time). A formal test for the PH assumption is given in Table 3.4. The covariates and the global test are non-statistically significant given by the large $p - values$. This is a further justification of the validity of the PH assumption for our proposed model.

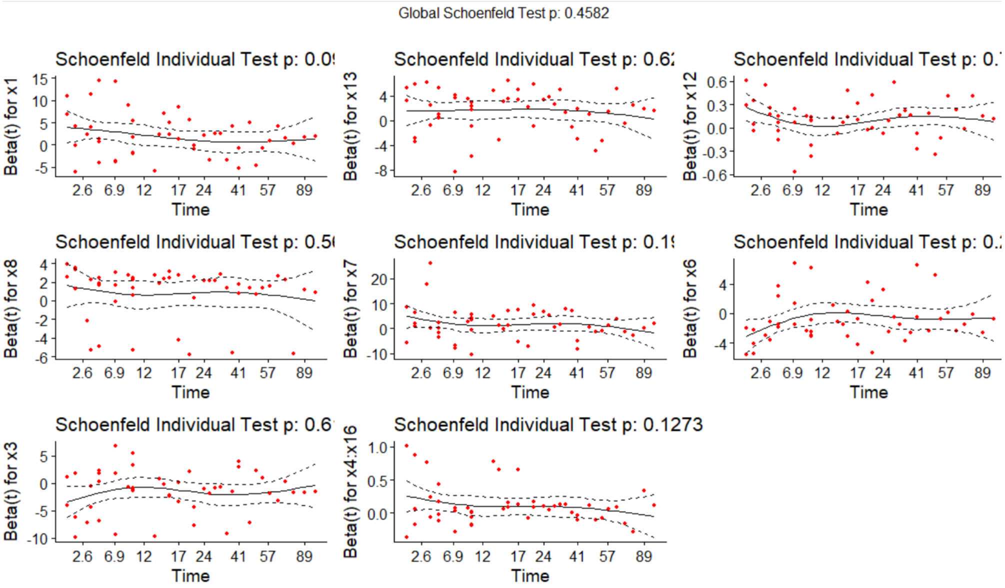

Figure 3.2: Testing Proportional Hazard Assumption

Table 3.4: Formal Test of Proportional Hazard Assumption

Covariate	ρ	χ^2	$P - value$
X_1	-0.2136	2.8454	0.0916
X_3	0.0636	0.2491	0.6177
X_6	0.1391	1.1587	0.2817
X_7	-0.1569	1.6895	0.1937
X_8	-0.0861	0.3291	0.5662
X_{12}	-0.0323	0.0651	0.7986
X_{13}	-0.0607	0.2348	0.6280
$X_4 : X_{16}$	-0.2067	2.3255	0.1273
GLOBAL	NA	7.7505	0.4582

Secondly, we assessed the functional form of continuous covariates. The continuous covariates are expected to have a linear form. However, categorical covariates do not have any issue of nonlinearity. Figure 3.3 is a plot of Martingale residuals against continuous covariates with fitted lowess (locally weighted smoothing) function. The plot demonstrates no major trend or pattern. Thus, the linear functional form of continuous covariates is reasonable. Therefore, continuous covariates have a linear functional form. We further investigated the presence of influential observations (or outliers). In Figure 3.4, we plot the magnitude of *dfbeta* against the model coefficients. We can see that there are no major influential observations, given that all the residuals are within one standard deviation of the residuals. Multicollinearity can negatively impact the precision of the estimated model coefficients and prediction. In Table 3.5, we employed the variance inflation factor (VIF) to assess multicollinearity. A $VIF > 2.5$ and $VIF > 5$ for categorical and continuous covariates, respectively, are evidence of the presence of multicollinearity. Given the VIFs in Table 3.5, implies that there is no multicollinearity in our proposed Cox PH model.

Table 3.5: Variance Inflation Factor of Model Coefficient

Covariates	X_1	X_{13}	X_3	X_{12}	X_8	X_7	X_6	$X_4 * X_{16}$
VIF	1.564	1.602	1.283	1.612	1.656	1.574	1.385	1.243

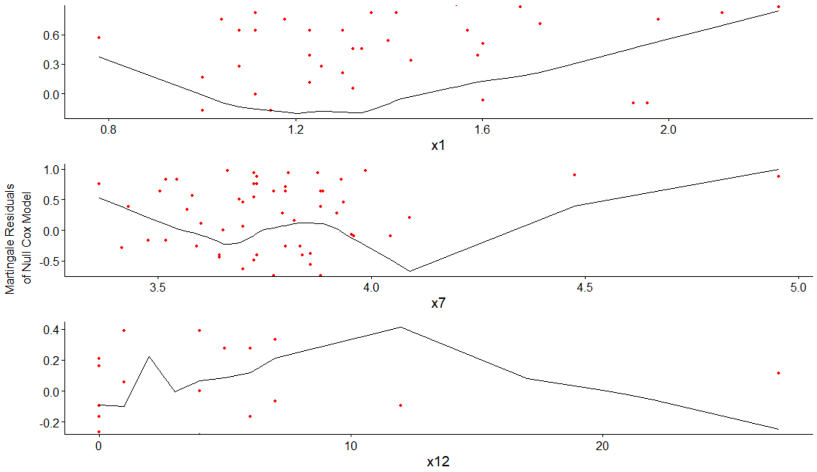

Figure 3.3: Assessing the Functional Form of the Continuous Covariates

3.3.2 The Proposed Cox-PH Model Survival Function

The survival function of the Cox-PH model is a reverse process of the hazard function in Equation (3.1). In Equation (3.1), we are given $f(t)$ and $\hat{S}(t)$, then we proceed to find $h(t)$. In Cox-PH, we find $\hat{S}(t)$ given $f(t)$ and $h(t)$. In addition to the relationship between $h(t)$ and $\hat{S}(t)$ in Equation (3.1), another alternate relation is given by

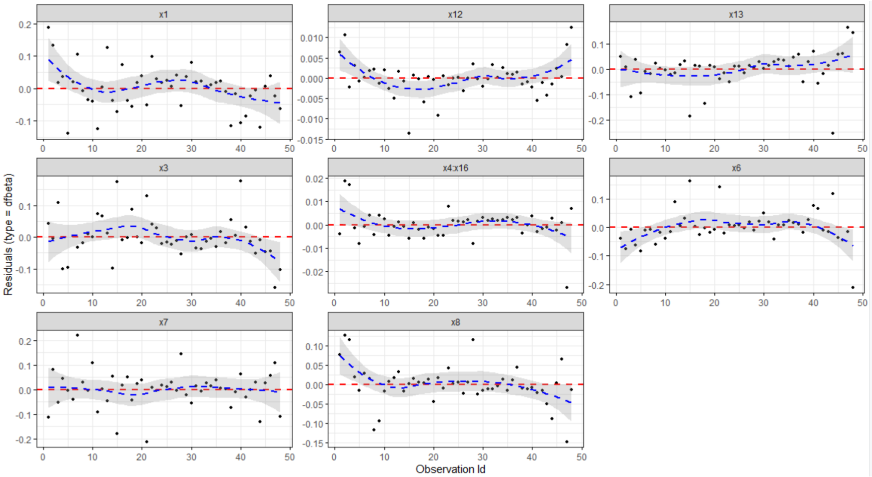

Figure 3.4: Assessing Influential Observations (or Outliers)

$$\hat{S}(t) = \exp\left(-\int_0^t h(t)dt\right) = \exp\left(-\int_0^t H(t)\right). \tag{3.5}$$

The Cox-PH model given by Equation (3.2) can be re-written in the form

$$h(t_i; X_i, X_iX_j) = h_0(t)\exp\left(\sum_{i=1}^{k}\beta_i X_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right). \tag{3.6}$$

We can modify the Cox-PH model for the survival function by employing Equation (3.5) above. Therefore, the survival function of the Cox-PH model can be expressed as

$$\begin{aligned}\hat{S}(t_i; X_i, X_iX_j) &= \exp\left(-\int_0^t h(t; X_i, X_iX_j)dt\right)\\ &= \exp\left(-\int_0^t h_0(t)\exp\left(\sum_{i=1}^{k}\beta_i X_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)dt\right)\\ &= \exp\left(-\exp\left(\sum_{i=1}^{k}\beta_i X_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)\int_0^t h_0(t)dt\right)\\ &= \exp\left(-\int_0^t h_0(t)dt\right)^{\exp\left(\sum_{i=1}^{k}\beta_i X_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)}\\ &= S_0(t)^{\exp\left(\sum_{i=1}^{k}\beta_i X_i + \sum_{i\neq j=1}^{k}\rho_{ij}X_iX_j\right)},\end{aligned} \tag{3.7}$$

where $h_0(t)$ is baseline hazard function, which assumes any functional form of the covariates. The coefficient parameters of the covariates has been estimated, given by Table 3.1.

In section 3.3.1, we validated the quality of the proposed Cox-PH model in Equation (3.4) by showing that it satisfied all the three model assumptions outlined in section 3.2.1. Given that the model is of high quality, we estimated the Cox-PH survival function (i.e. the proportions of survival beyond time t) of patients diagnosed with multiple myeloma cancer (MMC) as a function of the seven covariates (risk factors) and the interaction term, given by Equation (3.7). Figure 3.5 shows the proposed Cox proportional hazard model survival function (i.e. the also known as the survival function) of the survival time. This plot demonstrates the predicted survival proportion at any given point in time for the mean values of the risk factors. The cox model is very useful in predicting the probability of the survival time for an individual patient based on the significant attributable risk factors that we have identified. Thus, given that a patient is diagnosed with MMC, we put into the model the seven contributing risk factors and the interacting factor to estimate the probability of survival beyond a given survival time (death time).

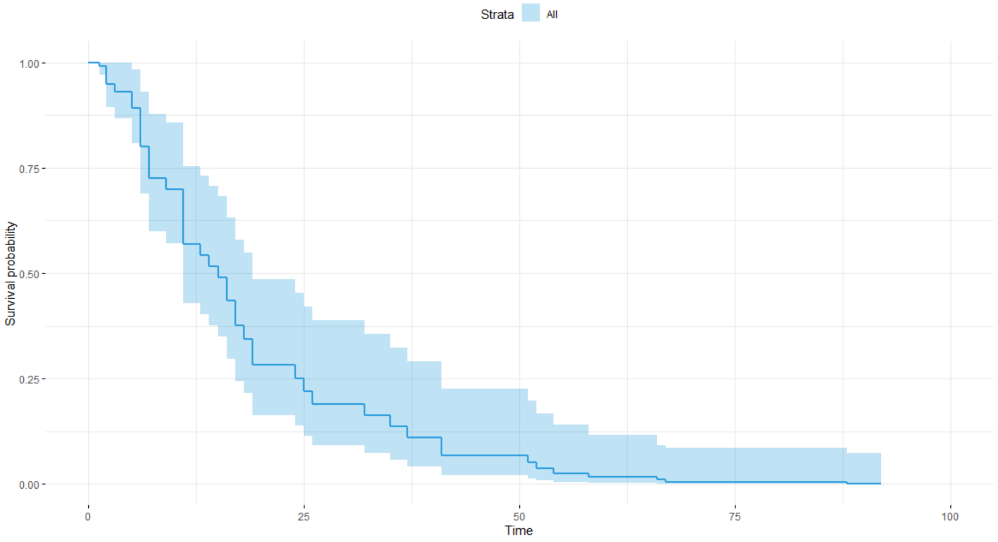

Figure 3.5: Survival Estimate $\hat{S}(t)$ from the Proposed Cox-PH Model

3.4 Discussion

The multiple myeloma cancer (MMC) cancer diseases may be incurable. However, the introduction of therapeutic agents such as thalidomide, lenalidomide, bortezomib, and high-dose chemotherapy and stem-cell rescue (ASCT) has improved the treatment progress, hence the survival time of the patient. Also, many research techniques and approaches have been adopted to enhance the patients' survival time after been diagnosed with MMC.

In the present study, we performed the Cox-PH model analysis of the survival times without considering stratification of the data, given that the survival times between males and females were not different (See Figure 2.2 in Chapter 2). We then estimated the proportion of the survival time as a function of covariates utilizing the Cox proportional hazard regression model.

Our data on the survival times of patients diagnosed with MMC included 16 risk factors presumed to be contributing to survival times. We used the Cox proportional hazard model to estimate the proportion of survival time because it takes into consideration the risk factors (covariates) contributing to the patients' survival time. Therefore, we developed the Cox-PH model for the survival time of patients diagnosed with MMC base on the sixteen risk factors. Our final proposed Cox-PH model given by Equation (3.4) identified seven significant risk factors and one interaction term as contributing to the survival probability. They are **blood urea nitrogen (BUN)/serum creatinine**, **white blood cells (WBC)**, **Bence Jone protein in the urine (BJPU)**, **fractures**, **proteinuria**, **gender**, **platelets**, and the interaction **infections and serum calcium**. It is interesting and highly important to point out that the two interacting risk factors do not individually significantly contribute to the survival probability, a highly useful finding. Seldom do we see interaction terms in Cox-PH models because they are difficult to find. However, not including interaction(s) in the Cox-PH model given that they exist and affect the survival time of the patient can result to the wrong estimation of the proportion of the survival time, hence diminishing the true relevance and quality of the Cox-PH model, and consequently endangering the treatment process of the MM cancer disease.

The proposed Cox-PH model satisfied all the key assumptions of the Cox-PH model discussed in section (3.2.1). Most research uses the Cox-PH models without discussing or validating the assumptions that allowed for its usage. Therefore, we cannot justify the quality of the Cox-PH model they proposed in their findings. Our proposed Cox-PH model is of high quality because the underlying key assumptions are satisfied and well-validated. The proposed Cox-PH model has the least AIC based on the stepwise model selection procedure with uncorrelated covariates given by the very small VIF values in Table 3.5. We rank the identified significant attributable risk factors or covariates based on the prognostic effect (i.e. the highest contributing factor to MMC deaths to the least contributing factor to MMC deaths) on the survival time utilizing the hazard ratio. We found all the identified risk factors, except the female gender and normal platelet to be highly prognostic factors (negatively associated with the survival time).

We found five new risk factors contributing to the survival of patients with MMC. They include the individual risk factors platelets, gender, white blood cells and fractures; and the interaction term, infections and serum calcium. These newly identified risk factors are not found in the findings by [20, 23, 40, 29], who developed statistical models to determine the association of some risk factors to the survival time of MMC. The risk factor, serum calcium was individually found to significantly contribute to the survival time by [29]. Whereas we found it to be significant as it interacts with infections, and it negatively relates to the survival time, given by the hazard ratio ($HR > 1$). However, we believe our model is more genuine and accurate, given the fact that the identified significant contributing risk factors were carefully assessed and selected based on the AIC of stepwise model selection technique, and validated to satisfy all the key model assumptions.

3.5 Contribution

We estimated the survival probability of patients diagnosed with multiple myeloma cancer (MMC) using the semi-parametric Cox proportional hazard model. We believe the proposed Cox-PH model given by Equation (3.4) gives an accurate estimate of the survival probability of patients diagnosed with MMC. The Cox-PH model was used to estimate the probability of the survival time because it incorporates into the model the additional information about risk factors contributing to the survival time. We identified seven risk factors and one interaction term as contributing to the survival probability of patients diagnosed with MMC. They are **blood urea nitrogen (BUN)/serum creatinine**, **white blood cells (WBC)**, **Bence Jone protein in the urine (BJPU)**, **fractures**, **proteinuria**, **gender**, **platelets**, and the interaction of **infections and serum calcium**. The interacting risk factors (infections and serum calcium), but individually did not significantly contribute to the survival probability. However, both together should be considered as significant interaction when identified at the same time a patient is diagnosed with MMC. Our final proposed Cox-PH model is of very high quality, robust, and efficient given by the fact that it satisfies all the key required assumptions in section (3.2.1). The stepwise model selection procedure was utilized to carefully assess and select the risk factors and the interaction term based on their statistical significance to the survival probability.

The final proposed Cox-PH model is the model with the least AIC. The identified significant contributing risk factors and the interaction have been rank according to the prognostic effect on the survival time using the hazard ratio. The interaction term between infections and serum calcium has been ranked 5, and has negative association with the length of survival time of MMC. The relevance of the proposed Cox-PH model is that we can estimate the survival probability of a patient given the values of the seven identified attributable risk factors and the interaction term. Of the seven risk factors, four of them, and the interaction term are new significant contributing risk factors to the survival time of MMC identified by our proposed model, namely; platelets, gender, white blood cells and fractures; and the interaction term, infections and serum calcium. Our findings offer further prognostic and therapeutic importance for decision making for the treatment of multiple myeloma cancer.

Base on the Cox-PH analysis of the survival times of the MMC patients, we recommend the following. (1) If additional information about what is causing the survival time (risk factors) is known, then we recommend the use of the Cox proportional hazard model to estimate the survival probability. Thus, the Cox-PH model takes into consideration the additional information given by the risk factors, hence the resulting survival probability can be more accurate, robust, and efficient. (2) The investigation of the significance of interaction terms in Cox-PH models should not be overlooked or underestimated because they can greatly affect the accurate prediction of the patients' survival rate of the multiple myeloma cancer disease, leading to dangerous medical and therapeutic/treatment issues.

Chapter 4: Data Driven Statistical Modeling and Analysis of the Survival Times of Multiple Myeloma Cancer (MMC)

To further improve the therapeutic/treatment strategy of multiple myeloma cancer (MMC), this chapter focuses on developing a highly accurate predictive model of real values of the survival times of patients diagnosed with MMC taking into consideration the 16 risk factors presumed to influence the survival times, given by Table 2.1 in Chapter 2. That is, we proposed a data-driven statistical model for the survival time of 48 patients diagnosed with MMC as a function of 16 attributable risk factors. We identified nine attributable risk factors out of sixteen and one interaction term to significantly contribute to the survival time. The bootstrap resampling technique was utilized to increase the sample size to 300, resulting in asymptotic significance of the coefficients of the risk factors and coefficient of determination (R^2) of 91%. The proposed model satisfied all the model assumptions, passes the residual analysis test, and has a very high prediction accuracy of 93%. Thus, it passes the goodness-of-fit test and the qualities of a good model. The identified significant attributable risk factors and the interaction has been ranked based on the percent contribution to the survival time. The proposed model was evaluated and compared with other existing models of survival of multiple myeloma. Our model is very accurate and also identifies some new significant risk factors. The research findings of this chapter have been published [62].

The organization of this chapter is as follows: Section 4.1 introduces and review some literature of studies on MMC; Section 4.2 presents the proposed statistical modeling and analysis; Section 4.3 discusses the findings in this study; and finally, the research contributions of this chapter is given in Section 4.4.

4.1 Introduction

Evidence about the risk factors or what typically causes multiple myeloma cancer (MMC) remains scant. The existence of the myeloma plasma cell has not been quantified to be able to assess the contributing risk factors of MMC. However, several risk factors have been identified to have some relation with the survival of patients with MMC [20, 39]. Most of these factors were identified at the time a patient was diagnosed with MMC. [39, 40, 41], all stated in their findings that hemoglobin, immunoglobulin type, extent and type of lesions, serum calcium, serum albumin, presence of Bence Jones protein, and performance status, at the diagnostic of MMC are known to be essential in association with survival of patients with MMC.

Some statistical analysis has been done on the survival of patients with MMC given the event that a patient died or survived. Most of the research works done on MM has focused on how to improve the therapeutic strategy of MMC. Brain et al [20] used Kaplan and Meier to test whether there was a significant difference in the survival duration between the categories of risk factors based on the generalized Wilcoxon test and the log-rank test. They further used a non-linear Cox regression analysis to determined the combination of patients characteristics relative to survival duration. They identified a significant difference in the survival duration among patients based on performance status, cell mass and percentage labeling index, Nephrotic status, and Hemoglobin but no significant difference regarding patients age. Another statistical analysis by John M. Krall et al [23] developed a set-up procedure for selecting variables associated with the survival times of patient with MMC utilizing the data that we are using in the present study. They identified log blood urea nitrogen (BUN), hemoglobin, log percent plasma cells in bone marrow (BM) and Serum calcium to be associated with the survival of patients with multiple myeloma.

Durie BGM and Salmon SE (1975) [40], developed a clinical staging system for MMC based on the correlation of measured myeloma cell mass of 71 patients determined from the measurement of monoclonal immunoglobulin (M-component) synthesis and metabolism.

They found a significant correlation of the measured myeloma cell burden with the extent of the bone lesion, hemoglobin level, serum calcium level, and M-component levels in serum and urine. However, serum creatinine/BUN had a strong correlation with the survival, and not the myeloma cell mass. Their findings produced a clinical staging system based on 3 tumor cell mass indices, namely, low (0.6×10^{12} cells/sq m), intermediate ($0.6 - 1.2 \times 10^{12}$ cells/sq m) and high ($> 1.2 \times 10^{12}$ cells/sq m). Giampaolo Merlini, Jan G. Waldenstrom, and Suresh D. Jayakar [29], proposed a new improved clinical staging system for the survival of MMC based on the analysis of 123 treated patients. They found serum calcium, % bone myeloma plasma cell (% BMPC) and serum creatinine/BUN to be strongly associated with the survival of IgG myeloma stage; hemoglobin, serum calcium, and M-component to be strongly associated with the survival of IgA myeloma stage; and creatinine/BUN, % BMPC and serum calcium to be strongly associated with the survival of BJ myeloma stage. Brian G.M. Durie, Sydney E. Salmon, and Thomas E. Moon [20] proposed a pretreatment tumor mass, cell kinetics, and prognosis in MMC of 150 patients base on the % labeling index (LI%) and DNA synthesizing cells (S). The findings of LI% <1% was associated with long survival, LI% > 3% in high cell mass patients with high S had a very poor prognosis.

In the present chapter 4, we developed a real data-driven statistical model of the significant attributable risk factors of survival time of patients diagnosed with multiple myeloma cancer. The objective is find a model that predicts the real value survival times with high accuracy. The previous chapters focused on finding the proportion of the survival times of MMC. However, being able to provide patients with how long they can survive given the identified risk factors and how much each risk factor contributes to their survival provides a highly useful information towards enhancing treatment of the cancer. The clinical trial that was conducted consisted of 65 patients who were diagnosed with MMC, given by Table 2.1 in Chapter 2. However, our study concentrated on 48 of the patients that we have death times (survival times) from diagnosis. The remaining 17 patients, we did not have information about their death times, so they were excluded from our analysis and modeling. Because of the low amount of the data, we did the modeling utilizing the 48 pieces of information. The data was filtered to fulfill all the modeling assumptions. After the development of the statistical model, we used the bootstrapping, resampling method to increase the amount of information, and then improved the accuracy of our statistical model. We identified the significant risk factors, and interaction contributing to the survival time of MMC. The significant risk factors including the interaction identified were ranked based on the percentage of contribution to the death of MMC patients, using the coefficient of determination (R^2) of the survival times. The quality and accuracy of the proposed model was assessed based on the R^2 and $R^2_{adjusted}$ statistic, the Akaike information criterion (AIC) of model selection, the prediction error sum of squares (PRESS), the root mean square error (RMSE), the variance inflation factor (VIF), the residual analysis, and the prediction accuracy (the correlation of the actual and predicted survival times based on 80% training set and 20% testing set).

4.2 Statistical Modeling

We develop a statistical model for the survival times (death times) of the 48 patients diagnosed and died of multiple myeloma. In the building of the statistical model for multivariate linear regression, the following assumptions must be satisfied:

1. Linearity: there should be a linear relationship between the response variable t (survival time) and the risk factors including interactions. This is expressed as

$$t_i = \alpha + \sum_{i=1}^{k} \beta_i X_i + \sum_{i \neq j=1}^{k} \rho_{ij} X_i X_j + \epsilon_i, \tag{4.1}$$

where the response variable $t_i = (t_1, ..., t_n)^T$, $\alpha = (1, ..., 1)^T$ is the model intercept parameter, $\beta_i = (\beta_1, ..., \beta_k)^T$ is the coefficient parameter of the attributable risk factors X_i's, ρ_{ij} is the coefficient parameter of interaction between i^{th} and $j^t h$ attributable risk factors, $\epsilon_i = (\epsilon_1, ..., \epsilon_n)^T$ represents the model residual error term, and $k = 16$ and $n = 48$ is the number of attributable risk factors and the sample size, respectively.

Linearity was assessed using the matrix of scatter plots and correlation between the response and the continuous risk factors.

2. Multivariate normality: the errors should follow Gaussian normal probability distribution with zero mean and standard deviation of one, $\epsilon \sim N(0, 1)$ as $n \to \infty$. This was tested using the normal probability $Q - Q$ plot and was verified using a formal test of normality i.e. the Shapiro Wilk's test with the null hypothesis H_0, that the residuals errors follow the normal probability distribution.

3. Homoscedasticity: the residual errors should have constant variance, $var(\epsilon_i) = \sigma^2$. We verify this by observing the plot of residuals versus fitted values; no pattern implies errors have constant variance. We then supported it with a formal test of non-constant variance with the null hypothesis H_0, that the variance of the errors is constant.

4. None or very minimum multicollinearity: the risk factors should not be highly correlated. Usually, a correlation coefficient of $r \geq 0.9$ indicates a very high correlation. A formal test for multicollinearity is using the variance inflation factor $VIF = \frac{1}{1-R_j^2}$; $VIF > 10$ implies the presence of multicollinearity.

5. No auto-correlation: residual errors are independent and uncorrelated, $\epsilon_i \sim i.i.d/N(0, \sigma^2)$. We checked this using a formal test of autocorrelation, i.e. Durbin Watson test with null hypothesis H_0, that there is no autocorrelation.

We started by visually inspecting the matrix of scatter plot to assess the linear relationship between the response variable t and the continuous risk factor X_i. As shown in Figure 4.1, there is a weak linear relationship between the response variable t, and all the continuous risk factors given that the highest correlation coefficient is $r = 0.31$, which is with X_1. The distribution of the survival times t is right-skewed as it follows the three-parameter log-normal probability distribution (from the parametric analysis). We can see that some of the risk factors have skewed shaped distributions (a possible influence of outliers or extreme values). However, we continued to fit a model of the response variable as a function of the 16 attributable risk factors resulting in a coefficient of determination (R^2) of .48 (48%); this cannot be considered a good model given that there are discrepancies associated with the data such as skewness or kurtosis [54]. However, fitting the first model to the original data allow us to check for other model assumptions.

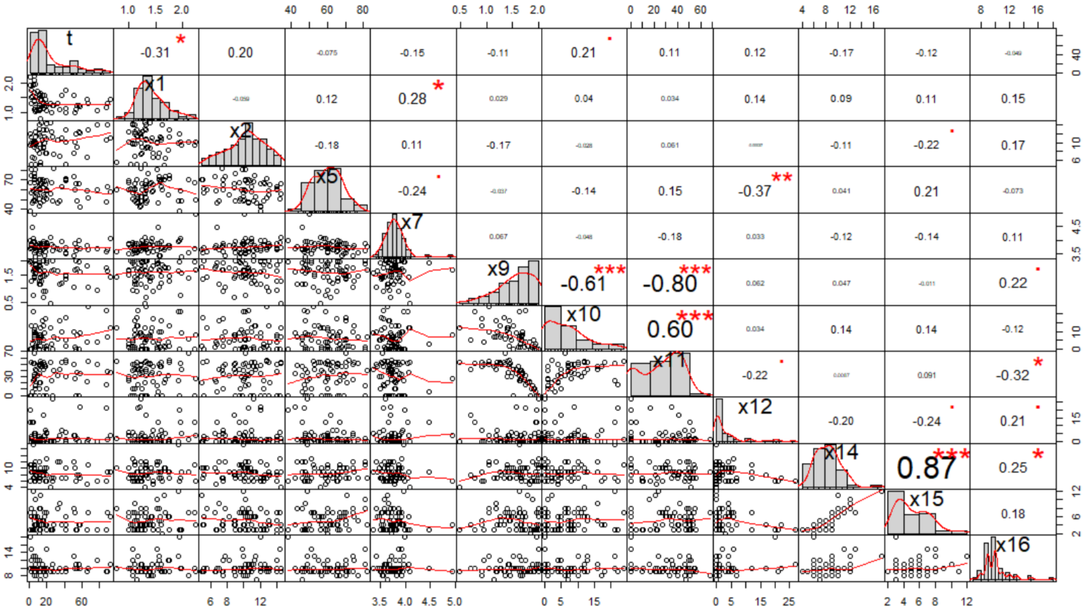

Figure 4.1: Correlation Matrix Scatter Plots of t and the Continuous Risk Factors

In Figure 4.2, we plotted the $Q-Q$ plot of residuals of the model built from the original data to assess the multivariate normal probability distribution. There is evidence of violation from normality as shown by the skewed ends of the $Q-Q$ plot. A formal test for normal distribution using the Shapiro Wilk's normality test resulted in a $p - value = 8.632e^{-03}$, which is an indication of lack of the normal probability distribution. This implies that the survival time t of patients with multiple myeloma does not follow the Gaussian probability distribution.

Given that there is weak linear relationship and no multivariate normality, we applied log transformation to the response variable t and the skewed risk factors X_{10}, X_{12}, X_{14} and X_{16}.

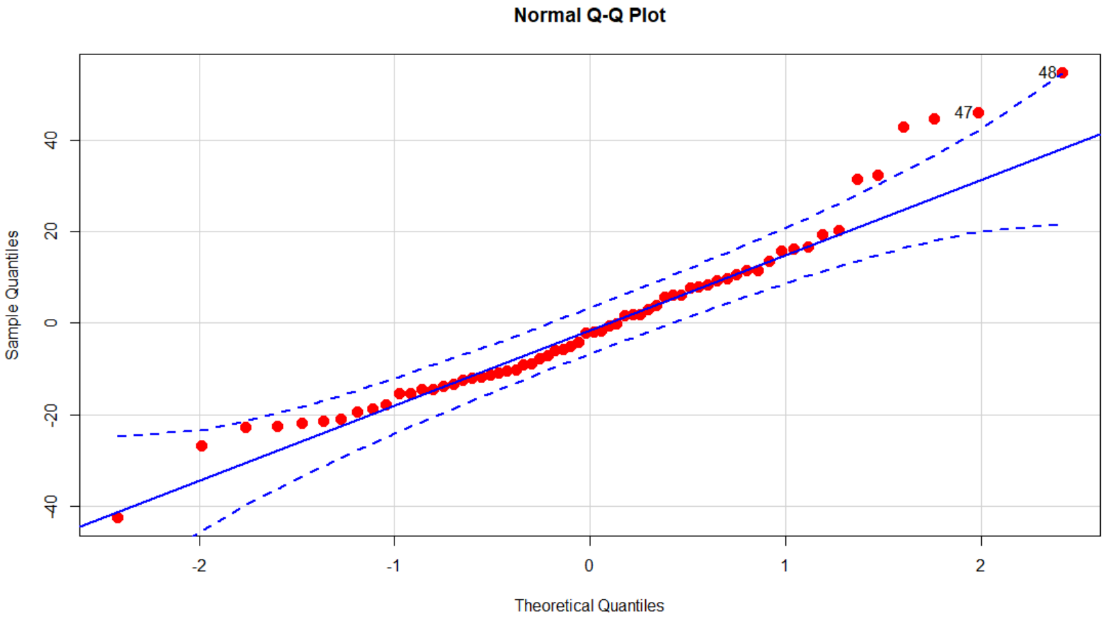

Figure 4.2: Testing Normal for the Distribution of the Model Residuals From Original Data

Log transformation stabilizes the variance and suppresses the impact of outliers or extreme values in the data [51]. The transformations are giving by the expressions below:

$$t' = log(t) \tag{4.2}$$

and

$$X_i' = \begin{cases} -log(-X_i + 1), & \text{if } x < 0. \\ log(X_i + 1), & \text{otherwise,} \end{cases}$$

where X_i' denotes the transformed risk factor of X_i. After the variable transformations, we proceeded to fit the full model for the survival times t as a function of the 16 risk factors and all two-way interactions between them. We then utilized the backward elimination procedure for model selection to find the attributable risk factors and the interaction(s) that significantly contributes to the survival time t. The backward elimination model selection technique is often used because it provides less bias mean square error (MSE) values and turns to prevents overfitting of the model, which is essential for the prediction performance of the model. Using this method of model selection, we selected the best model with the least Akaike information criterion ($AIC = 2ln(L) + 2k$, where L is the value of the maximum likelihood function of the model and k represents the estimated model parameters) [42]. AIC gives an estimation of the relative amount of information missing in the model; hence, the smaller the AIC value the better the quality of the model. Therefore, given the model selection method and criterion of choice of a good model, the best-proposed model with $R^2 = 0.8741$ which includes all the attributable risk factors and interaction that significantly contributes to the survival time of patients with multiple myeloma is given by

$$\begin{aligned} log(\hat{t}) = & -4.037 - 1.167X_1 + 0.267X_3 normal - 0.977X_4 present \\ & + 0.016X_5 + 0.504X_6 female - 0.581X_8 present + 0.020X_{11} \\ & - 1.209X_{13} none + 4.011X'_{16} - 0.228X_7.X'_{14}. \end{aligned} \tag{4.3}$$

Thus, there are nine attributable risk factors, namely, **Bence Jone protein in urine**, **blood urea nitrogen (BUN)/serum creatinine**, **infections**, **% myeloid cells in peripheral blood**, **fractures**, **serum calcium**, **gender**, **platelets** and **age**, and one interaction term, namely,**white blood cells and total serum protein** that significantly contribute to the survival of MMC patients. The following remaining five risk factors do not contribute to the survival time of MMC patients at diagnostic: **hemoglobin**, **plasma cells in bone marrow**, **lymphocytes in peripheral blood**, **proteinuria** and **serum-globin (gm%)**. Because the estimated survival time t' and the attributable risk factors X'_i from equation (4.3) are based on the log-transform data from equation (4.2), we utilized the anti-logarithmic to transform back to the original values. The backward transformation of the attributable risk factors X_{14} and X_{16} can be expressed as

$$X_i = \begin{cases} 1 - e^{-X'_i}, & \text{if } x < 0,. \\ -1 + e^{X'_i}, & \text{otherwise, for i = 14, 16.} \end{cases} \tag{4.4}$$

To use the above proposed model given by equation (4.3), we first take the anti-logarithmic of the log transform attributable risk factors into the original values, given in equation (4.4). We then take the anti-logarithmic of the entire model in equation (4.3) to arrive at the actual estimate of the survival time $\hat{t}$ of an MMC patient.

Now, given the above-proposed model of survival time t of patients diagnosed with multiple myeloma, one may ask how useful can this model be?. If a new patient is diagnosed with multiple myeloma, then given the values of the significant attributable risk factors identified in equation (4.3), we can use our proposed statistical model to accurately estimate the survival time $\hat{t}$ of that patient.

How accurate are the results/usefulness that we obtain in using the proposed nonlinear statistical model? We answer this question using the coefficient of determination statistic, R^2 and $R^2_{adjusted}$. The R^2 is generally used to measure the goodness-of-fit of a statistical model. It estimates the proportion of variation in the response variable explained by the model attributable risk factors [52, 53]. The higher the R^2 statistic the better the goodness-of-fit of a statistical model. In general, the R^2 is defined by

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}, \tag{4.5}$$

where $SS_{tot} = \sum_i (t_i - \bar{t})^2$, $SS_{reg} = \sum_i (\hat{t}_i - \bar{t})^2$ and $SS_{res} = \sum_i (t_i - \hat{t}_i)^2 = \sum_i e_i^2$; and t_i are the survival times, $\bar{t} = \frac{1}{n} \sum_i^n t_i$, $\hat{t}_i$ is the estimated survival time in equation (4.4). SS_{reg} is the regression sum of squares representing the variation explained by the proposed model, SS_{res} is the residual sum of squares representing the variation in the proposed model left unexplained and SS_{tot} is called the total sum of squares is the proportional to the sample variance, and equals to the sum of SS_{reg} and SS_{res}. Generally, the R^2 has the problem of increasing by increasing the number of parameters or predictors in the model. Therefore, it is recommended that we estimate the R^2 along with the $R^2_{adjusted}$ to adjust for the degrees of freedom of the model, and is given by

$$R^2_{adjusted} = 1 - \frac{SS_{res}/(n-p)}{SS_{tot}/(n-1)} = 1 - \frac{SS_{res}/df_{res}}{SS_{tot}/df_{tot}}. \tag{4.6}$$

Our proposed statistical model given in equation (4.3) resulted in an R^2 of 87.41% and $R^2_{adjusted}$ of 84.01%. This means the proposed model explains 87.41% variation in the response variable (i.e. the survival time of MMC patients), a very good quality model.

4.2.1 Bootstrapping with the Proposed Statistical Regression Model

To further improve the efficiency of the proposed statistical model, we utilized the bootstrapping resampling method that due to Efron (1979). Bootstrapping is a general approach to statistical inference that allows assigning measures of accuracy (defined in terms of bias, variance, confidence intervals, prediction error or some other such measure) to sample estimates based on building a sampling distribution for a statistic by resampling from the actual data that we analyzed in the present study [55]. We applied the bootstrap sampling to resampled with replacement the data used to build the proposed analytical model given by equation (4.3); increasing the sample size by 300. Thus, the bootstrap modeling involves the following algorithm;

- We bootstrap sample $r = 300$ drawn from $n = 48$ observations with replacement.
- In bootstrapping statistical model:
    - We estimate the model coefficients α, β_i, and ρ_{ij} for the original sample $n = 48$, and calculate the fitted value and residual for each observation.
    - Select bootstrap samples of the residuals, ϵ^*_b, and from these, calculate bootstrapped t values, $t^*_b = [t^*_{b1}, t^*_{b2}, ..., t^*_{bn}]$, where $t^*_{bi} = \hat{t}_i + \epsilon^*_{bi}$.
    - The bootstrap estimates are calculated by least-squares regression, then $b^*_b = (X'X)^{-1}X't^*_b$ for $b = 1, ..., r$

The bootstrap modeling asymptotically increased the level of significance of the coefficient estimates, making them equally highly significant, and increased both the R^2 and $R^2_{adjusted}$ to 91.16% and 90.85%, respectively. The modified version of the model in equation (4.3) based on the bootstrapping resampling method is given by

$$\begin{aligned} log(\hat{t}_{Bootstrap}) = & -4.377 - 1.097X_1 + 0.332X_3 normal - 0.949X_4 present \\ & + 0.016X_5 + 0.562X_6 female - 0.586X_8 present + 0.022X_{11} \\ & - 1.268X_{13} none + 4.151X'_{16} - 0.252X_7.X'_{14}. \end{aligned} \tag{4.7}$$

4.2.2 Validation of the Proposed Statistical Model

Before validating the proposed model, we need to be sure that all assumptions that underline our proposed model are satisfied. We tested for linearity by showing the linearity plot (sometimes referred to the partial residual plot) of the response variable and the significant attributable risk factors as shown in Figure 4.3, below. We can see that there is a well-established linear relationship between the response variable and the continuous attributable risk factors (shown by the blue and pink lines). Therefore, the linearity assumption which was initially a problem we encountered has been rectified in our final proposed statistical model.

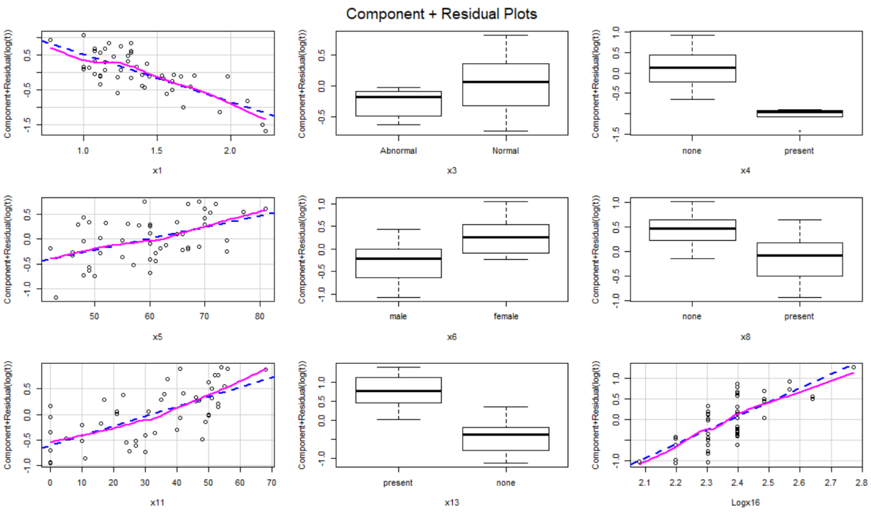

Figure 4.3: Evaluation of Linearity of the Proposed Statistical Model

To verify that the proposed statistical model satisfies multivariate normal probability distribution assumption, we used the normal $Q - Q$ plot shown by Figure 4.4. We see that the residuals are normally distributed with no major outlier and all the points in the plot fall within the 95% confidence bound. The evidence of normality is supported by the Shapiro Wilk's test of the normal probability distribution (a formal test), given by a high $p - value$ of 0.818. The plot of the distribution of studentized residuals in the second panel of Figure 4.4, is further evidence that the proposed model's normality assumption is valid.

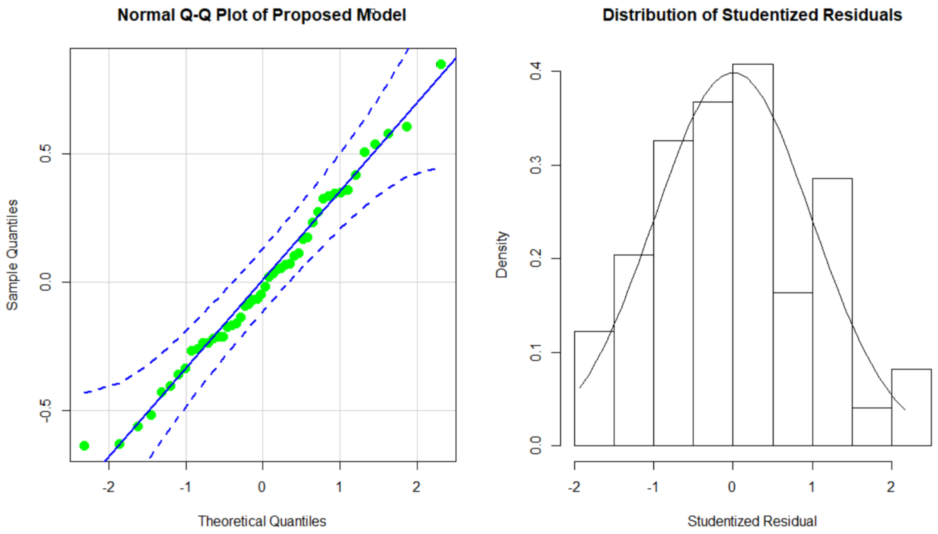

Figure 4.4: Test for Multivariate Normal Probability Distribution

We performed residual analysis to assess the model residuals and constant variance. Figure 4.5, depicts the residual plot of the proposed model. Thus, we can conclude that there is no problem of homoscedasticity.

Our proposed statistical model perfectly satisfies the assumption of constant variance, indicated by the randomly scattered points about the zero line with no major outliers. A formal test for homoscedasticity revealed a $p - value$ of 0.506, strongly suppots that homoscedasticity of our proposed model is valid. The mean absolute value of the residuals, $|\bar{r}| = \sum_i^n e_i$ is $4.779 \times e^{-2}$, close to zero and the variance $var(r) = 1/(n-1)\sum_i^n (r_i - \bar{r})^2$ is

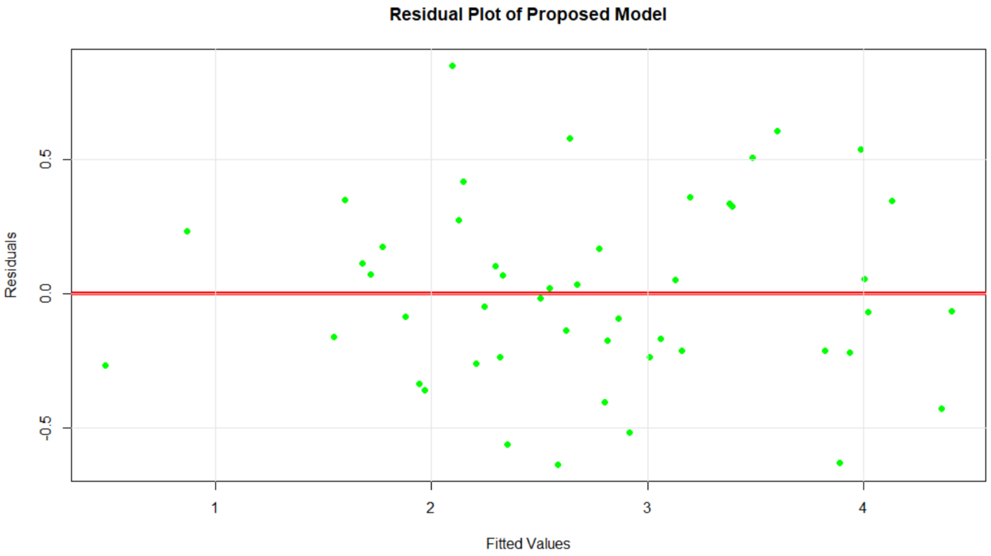

Figure 4.5: Residual Plot of the Proposed Statistical Model

0.636. The proposed statistical model has a very small root mean square error ($RMSE = \sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)/n}$) of 0.384.

Multicollinearity is a major problem in statistical modeling which must be addressed. It can distort the precision of the estimated coefficients leading to overfitting and misinterpretation on the results of the model. All the estimates of the parameters in our proposed model have a very small variance inflation factor, $VIF < 3$, indicating that there is no problem of multicollinearity. Also, we expect the model residuals to be independent and uncorrelated. We tested for the presence of auto-correlation among errors in the proposed model using the Durbin Watson of testing the null hypothesis H_0, no autocorrelation is present. Accepting the hypothesis with a large $p - value$ of 0.624 indicated that there is no autocorrelation among residuals in our proposed model.

To validate the prediction accuracy of our proposed statistical model, we trained 80% of the data to build our model and tested on the remaining 20% test data. The prediction of the original model and the trained model using the test data is given in Table 4.1, below.

We checked the accuracy of the predictions by finding the correlation coefficient r, and the corresponding R^2 (square of r) between the actual and the predicted values. This resulted in R^2 of 0.943, a very high prediction accuracy. The comparison of the logarithmic survival

Table 4.1: Comparison of Prediction of the Survival Time of Multiple Myeloma

Log(t)	Original Model	Trained Model
0.2231	0.4917	0.6379
1.0986	0.8676	1.3378
1.6094	1.9708	2.3358
1.9459	1.5979	1.1624
2.3979	2.1245	2.0809
2.7726	3.0083	2.9017
3.1781	3.1269	2.6158
3.7136	3.3900	3.1649
3.9889	3.4851	3.2839
1.3863	1.5493	2.1449

times with the two models (i.e. model developed using all the 48 patients and the 80% trained model) prediction on the test data resulted in R^2 of 0.943 and 0.930, respectively, attesting to the high prediction accuracy of our proposed model.

4.2.3 Ranking of the Contribution of Attributes/Risk Factors of the Survival Times of Multiple Myeloma

In this section, we rank the individual significant risk factors and the interaction based on their contribution to the survival time of MMC patients using the percentage of R^2. Table 4.2, shows the rank of each of the identified significant risk factors and the interaction term. Bence Jone protein in urine is ranked first, followed by blood urea nitrogen (BUN), the interaction term is ranked eighth, and age has the least contribution to the survival time of patients diagnosed with multiple myeloma among the significant attributable risk factors. A detailed discussion of the rankings will continue in the next section.

4.3 Discussion

The evaluation of the survival time of patients diagnosed with MMC is an essential prerequisite for improving the prognosis and therapeutic/treatment strategy of multiple myeloma. The present study was designed to find a real data driven statistical model that accurately

Table 4.2: Rank of Contribution of Attributing Risk Factors to Survival Time

Rank	variable	Description	R^2	%Contribution
1	X_{13}	Bence Jone protein in urine at diagnosis 1-present, 2-none	0.2672	30.57
2	X_1	Log BUN at diagnosis	0.2052	23.48
3	X_4	Infections at diagnosis 0 none, 1 present	0.0949	10.86
4	X_{11}	% Myeloid cells in peripheral blood at diagnosis	0.089	10.18
5	X'_{16}	Serum calcium (mgm%) at diagnosis	0.0661	7.56
6	X_8	Fractures at diagnosis 0 none, 1 present	0.0613	7.01
7	$X_7\&X'_{14}$	Log WBC at diagnosis and Total serum protein at diagnosis	0.0379	4.34
8	X_6	Gender 1 male, 2 female	0.0329	3.76
9	X_3	Platelets at diagnosis 0 abnormal, 1 normal	0.011	1.26
10	X_5	Age at diagnosis (complete years)	0.0086	0.98
Total			0.8741	100

predicts the survival time from diagnosis to the nearest month of multiple myeloma patients deaths. In the present study we accomplished the following: (1) we identified the significant attributable risk factors. (2) we identified the significant interactions among the risk factors. (3) we determined the percentage of contributions of each identified risk factor and interaction that causes the death of the patients. It was important to assess whether there is a difference in the survival times with gender in which we found no difference, a good characterization for our data analysis of the development of our model. We started building the statistical model with 16 predictors (risk factors) reported to be contributing to the survival of MMC but we only found nine (9) individually contributing factors along with a single interaction. Most of the risk factors in our data have been reported to be important by several researchers [20, 29, 39, 40, 57, 58, 59], however, we did not find all of them to be important. The final proposed model that accurately predicts the survival time is given by equation (4.7), in a transformed form. We proceed to take the anti-logarithm of the transformed model to get the original values of the survival time utilizing equation (4.4). The goodness-of-fit of the model was very carefully evaluated as follows: (1) the model satisfies all the (1-5) assumptions of a good statistical regression model as we described it in section 4.2, (2) it passes the residual test of a good model, i.e. $\epsilon_i \sim N(0, 1)$, (3) it has a very good R^2 of 87.41%; the R^2 of the model was further increased to 91.16% using the bootstrapping methods of resampling with replacement, and (4) it has a very high prediction accuracy of about 94% base on 80% training data and 20% test data.

The justification of the usefulness/relevance of the proposed statistical model compared to other existing models or findings was assessed and evaluated. Our proposed model identified the 9 risk factors and one interaction term to be significantly contributing to the survival time of patients with MMC, given in Table 4.2. Given any set of values of the significant risk factors that we have identified, we can predict the survival time of a patient with multiple myeloma with at least 94% accuracy. Serum calcium, blood urea nitrogen (BUN)/serum creatinine, and Bence Jone protein in urine (BJPU) were identified to be significantly contributing to the survival time, a finding consistent to that reported by others [40, 29]. BJPU was ranked as the highest contributor to the survival time, followed by BUN, and serum calcium was ranked sixth; see Table 4.2. Both BUN and serum calcium were identified to be a significant contributor to the survival time in the IgG myeloma group and BJ myeloma group, a finding reported by Giampaolo Merlini et al [29]. We expected the percentage of bone marrow plasma cells (%BMPC) to significantly contribute to the survival time, but that was not the case in our findings; an observation difficult to explain. Giampaolo Merlini et al found %BMPC not correlated with survival in the IgA myeloma group, parallel to our finding. We found age (ranked 10), and gender (ranked 8) to be significant contributors to the survival time, a finding mostly ignored by some researchers. Giampaolo Merlini et al reported age and gender to have no major correlation with the survival of MMC, a contrast to our findings. Our findings are consistent with that reported by the national cancer institute for Surveillance, Epidemiology, and Ends Results (SEER cancer) [4], as they reported age and sex as important risk factors to multiple myeloma. This suggests that age and gender are important attributable risk factors to survival of MMC, as indicated by our findings.

Other risk factors we identified to be significantly contributing to the survival time of MMC, and are not found in other studies, for example infections (ranked 3), percentage myeloid cells in peripheral blood (ranked 4), fractures (ranked 5), platelets (ranked 9), and an interaction between white blood cells (WBC) and total serum protein (ranked 7), all at diagnosis. With our proposed model, we can tell the influence that a given risk factor has on the survival time holding the other risk factors constant. For instance, assume that the values of all the other risk factors remain unchanged in a patient diagnosed with MMC, then we can tell that an increase in Bence Jone protein in urine would decrease the survival time (death time) of a patient, and vise Versa. This observation can be very important in aiding and improving the therapeutic/treatment process of MMC. Also, the fact that Bence Jone protein in urine was ranked to be the highest contributor to MMC survival, means that an MMC patient with an increased Bence Jone protein in urine can be a life-threatening situation, and would require immediate and critical treatment attention. WBC and total serum protein were not individually found to be significantly contributing to survival time. However, having both risk factors present at the same time at diagnosis was found to be a significant contributor to survival time. This finding can be very important and useful as a therapeutic means and treatment process of multiple myeloma, this is not found in other research publications.

4.4 Contribution

We have developed and propose a data-driven statistical model that identifies nine significant risk factors and one interaction term, namely **Bence Jone protein in urine**, **blood urea nitrogen (BUN)/serum creatinine**, **infections**, **% myeloid cells in peripheral blood**, **fractures**, **serum calcium**, **gender**, **platelets** and **age**, and **white blood cells and total serum protein** that contribute to the survival time of patients diagnosed with multiple myeloma. The proposed model has been evaluated using the statistical model assumptions, coefficient of determination (R^2 along with $R^2_{adjusted}$) statistic, the Akaike information criterion (AIC) of model selection, the prediction error sum of squares (PRESS), the root mean square error (RMSE), the variance inflation factor (VIF), the residual analysis, and the prediction accuracy (the correlation of the actual and predicted survival times based on 80% training set and 20% testing set) to be of high quality. Our proposed statistical model offers five important and useful findings in the multiple myeloma patients. (1) Given any set of values of the identified significant risk factors, we can obtain a good estimate/prediction of the survival time of patients diagnosed with MMC. (2) Identifies the individual risk factors and interaction that are significantly contributing to the survival time of MMC patients. (3) We can obtain the ranks of the attributable risk factors based on the percentage of contribution to the survival time of MMC patients. (4) We can perform surface response analysis to assess the contribution by each risk factor as a way to maximize the survival time of multiple myeloma patients. (5) We can compute confidence limits with a desirable degree of confidence that will be essential in controlling the survival time; for instance, when the survival time of a patient fall below the confidence limit he/she can be said to be in a critical condition, and hence requires immediate attention and treatment. The above statistical findings are with a high degree of accuracy.

Chapter 5: A New Statistical Modeling Approach for Survival Analysis of Cancer Patients - Multiple Myeloma Cancer

The Cox Proportional Hazard (Cox-PH) model has been a popularly used method for survival analysis of cancer data given the survival times as a function of covariates or risk factors. However, it is very seldom to see the assumptions for the application of the Cox-PH model satisfied in most of the research studies, raising questions about the effectiveness, robustness, and accuracy of the model predicting the proportion of survival times. This is because the necessary assumptions in most cases are difficult to satisfy, as well as the assessment of interaction among covariates. In the present chapter, we proposed a new approach to survival analysis using multiple myeloma (MM) cancer data. We utilized the nonlinear statistical model developed in Chapter 3 to predict 300 survival times. We then performed a parametric analysis on the predicted survival times to obtain the survival function which is used in estimating the proportion of survival times. The new proposed approach for survival analysis has proved to be more robust and gives better estimates of the proportion of survival than the developed Cox-PH model in Chapter 2. It gave higher R^2 and lower AIC. Also, Satisfying the proposed model assumptions and finding interactions among risk factors is less difficult compared to the Cox-PH model. The new proposed nonlinear statistical model approach for survival analysis of cancer diseases is very efficient and provides an improved and innovative strategy for cancer therapeutic/treatment. The research findings of this chapter have been published [63].

We organized the chapter as follows: Section 5.1 introduces and review some literature of studies leading to the new proposed statistical modeling approach for survival analysis in this chapter; Section 5.2 presents the development of the survival function of the nonlinear statistical model; Section 5.3 presents the comparison of the survival function of the proposed approach with that of proposed Cox-PH model's survival function in Chapter 2; Section 5.4 outlines the algorithm for the new proposed cancer survival analysis; Section 5.5 discusses the findings in this study; and finally, the research contributions of this chapter is given in Section 5.6.

5.1 Introduction

In our previous Chapters 1 and 2, we obtained the parametric, non-parametric and semi-parametric analysis of the survival times of 48 patients diagnosed with multiple myeloma. In the parametric analysis, we found the survival times to follow the three-parameter log-normal distribution and then we proceeded to obtain the survival function. In the case of the non-parametric analysis, we used the commonly used Kaplan-Meier model to obtain the survival function and then estimate the probability of survival times. For the semi-parametric analysis, we adopted the Cox proportional hazard to obtained the survival function of the death times. In comparing the parametric and non-parametric analysis of the survival time, we justified that the parametric method is more robust and efficient. However, none of the two methods take into consideration the risk factors contributing to the survival time. Therefore, we believe the Cox-PH model is more relevant for estimating the proportion of patients' survival beyond a given time than the other two because it takes into account the additional useful information given by the risk factors (covariates) contributing to the survival times. The Cox-PH on the other hand has some flaws. The necessary assumptions for applying the Cox-PH model are often difficult to satisfy, so as the finding of interaction among the risk factors. As a result, most research studies use the Cox-PH model without satisfying the underlying assumptions and also finding the interaction among covariates. This makes it difficult to justify the genuineness of conclusions made from using the Cox-PH model and the accuracy of predicting the proportion of survival.

Also in Chapter 3, we developed a data-driven nonlinear statistical model of the 48 patients diagnosed with MM and obtained a very accurate and high coefficient of determination, $R^2 = 0.8741$ along with $R^2_{adj} = 0.8401$. We further utilized the bootstrapping resampling technique to increase the sample size of the survival times from 48 to 300, thereby increasing $R^2 = 0.9116$ along with $R^2_{adj} = 0.9085$. Similar to the Cox-PH model, the nonlinear statistical model takes into consideration the additional information given by the risk factors contributing to the survival times. Most often the Cox-PH model has been employed in analyzing survival data with given covariates or risk factors. Whilst the Cox-PH model is used to estimate the proportion of patients surviving beyond a given time for a given number of risk factors, the nonlinear statistical model is used to estimate or predict the real value of the survival times of patients for a given number of risk factors. In the present study, we developed the survival function from the nonlinear statistical model and use it to estimate the proportion of patients survival of MM beyond a given survival time, and compared with the survival function of the commonly used Cox-PH model as a means of survival data analysis of the survival time as a function of covariates or risk factors. Please, see Section 2.2 of Chapter 2 for the detail description of the data used in this chapter analysis.

In Table 5.1, we show the significant attributable risk factors identified for the Cox-PH model and those of the nonlinear statistical model in ranking order from most significant to the least significant. Interestingly, we can recognize that almost all the risk factors (covariates) that were identified to be significantly contributing to the survival times of MM by the Cox-PH model were as well found significant in the nonlinear statistical model. The only exception is that the ranking positions are different. This is an extremely important feature to support the high quality and accuracy of our research findings. Thus, the fundamental justification for the comparison of the two models is given below by Table 5.1. We can obtain more and detailed information about the risk factors causing multiple myeloma from [4, 5, 8].

Table 5.1: Significant Attributable Risk Factors of the Cox-PH Model and the Nonlinear Statistical Model.

Rank	**Cox-PH**	**Nonlinear Statistical Model**
1	blood urea nitrogen (BUN)	Bence Jone protein in urine, X_{13}
2	White Blood Cells (WBC)	Blood Urea Nitrogen (BUN), X_1
3	Bence Jone protein in the urine	Infections, X_4
4	Fractures	% Myeloid cells in peripheral blood, X_{11}
5	Infections & serum calcium	serum calcium, $X_{16}{'}$
6	Proteinuria	Fractures, X_8
7	Gender	WBC & total Serum protein, $X_7 : X_{14}{'}$
8	Platelets	Gender, X_6
9		Platelets, X_3
10		Age, X_5

The rankings of the risk factors in the Cox-PH model are based on the hazard ratio, HR, and those of the nonlinear statistical model are based on the coefficient of determination, R^2. The HR measures the relative risk or the prognostic effect of the covariates to the length of survival time. The higher the HR, the more the impact or contribution of a covariate to the survival time of MMC. Generally, $HR > 1$ implies that the covariate has an increased risk of association with the length of survival time, $HR < 1$ implies that the covariate has a decreased risk of association with the length of survival time, and $HR = 1$ means that the covariate has no risk of association with the length of survival time. R^2, on the other hand, measures the variability in the survival time explained by the covariates or risk factors. The higher the percentage of R^2 of a given covariate, the more contribution it makes towards explaining the variation in the survival time. Therefore, both HR and R^2 can be said to play a similar role in determining the prognostic effect of covariates on the survival time. However, R^2 is most recommended and efficient given that it measures the overall contribution of the risk factors in the model. Hence, R^2 gives more accurate information about the impact or prognostic effect of risk factors to the survival times than the HR.

The Cox-PH model identified eight attributable risk factors, including one interaction. The nonlinear statistical model identified ten attributable risk factors, including one interaction. Blood urea nitrogen is ranked first as the highest prognostic factor in the Cox-PH model but ranked second in explaining the variability in the survival times of the nonlinear statistical model. Bence Jone protein in urine was ranked first as the highest contributor in explaining the variability in the survival times but ranked as the third prognostic factor in the Cox-PH model. Interestingly, both models identified only one significant interaction. However, the interacting risk factors are different in the two models. The Cox-PH model identified infections and serum calcium as interaction, and the nonlinear statistical model identified white blood cells (WBC) and total serum protein as interaction factors. Another interesting information we can derive from Table 5.1 is that the risk factors making-up the interaction in the Cox-PH model was identified to be individually significantly contributing to the survival times in the nonlinear statistical model. Furthermore, white blood cells (WBC) individually significantly contributed to the proportion of survival in the Cox-PH model, but it was part of the interaction term identified by the nonlinear statistical model. The risk factor, proteinuria, was identified as significant in contributing to the proportion of Survival by the Cox-PH model, but not identified significant by the nonlinear statistical model. Whereas, age and myeloid cells in peripheral blood were significantly identified to be contributing to the survival time by the nonlinear statistical model, but not identified by the Cox-PH model. For the nonlinear statistical model, a positive coefficient or parameter means that a unit increase in the risk factor increases the survival time by the size of the coefficient, and a negative coefficient means that the survival time decreases by the size of the coefficient whenever there is a unit increase in the risk factor given that the other risk factors remain unchanged. In the Cox-PH model, a unit increase in a covariate with a positive coefficient leads to a decrease in the proportion of survival beyond a given time by the size of the coefficient. Whereas increasing a covariate with a negative coefficient by a unit will increase the proportion of survival beyond a given time by the size of the coefficient, given that the remaining other covariates remain constant. It is important to recognize that the criterion of model selection of the Cox-PH model was based on choosing the model with the least Akaike information criterion (AIC) [42], whereas the nonlinear statistical model was based on choosing the largest R^2 along with the R^2_{adj} and the least AIC. Though the two models have some differences, in general, each model identified significantly most of the risk factors found by the other. We recommend that the nonlinear statistical model is more powerful in identifying the risk factors and their percentage contribution to the response, the survival time.

5.2 Development of the Survival Function of the Nonlinear Statistical Statistical Model.

In the present chapter, we find the survival function of the death times t^* predicted from the final proposed nonlinear model with 300 failure or survival times, given by equation (4.7) in chapter 4. The proposed statistical model we developed, is given by

$$\begin{aligned} t_i{}^* = \; & \exp\left(-4.377 - 1.097X_1 + 0.332X_3 normal - 0.949X_4 present\right. \\ & + 0.016X_5 + 0.562X_6 female - 0.586X_8 present + 0.022X_{11} \\ & \left. - 1.268X_{13} none + 4.151X'_{16} - 0.252X_7.X'_{14}\right), \end{aligned} \tag{5.1}$$

where $i = 1, 2, ..., 300$ and

$$X_j = \begin{cases} 1 - e^{-X'_j}, & \text{if } x < 0,. \\ -1 + e^{X'_j}, & \text{otherwise, for j = 14, 16.} \end{cases} \tag{5.2}$$

Using the above model, we generated $t_1{}^*, t_2{}^*, ..., t_{300}{}^*$ survival times that is based on the risk factors that have been identified for each patient of MM. That is, $t_1{}^*$ is the survival time of patient 1 given the influence of each risk factor, ${t_n}^*$ is the survival time of the n^{th} patient based on the influence of each of its risk factors.

To investigate the distribution of the predicted survival times t^* of the MM patients, we first displayed the descriptive statistics of the survival times t^*. A detailed explanation of the implication of the value of the statistic (especially kurtosis and skewness) is given in Chapter 1. The values of the descriptive statistics show that t^* is skewed, given by the higher value of skewness and kurtosis in Table 5.2.

Table 5.2: Descriptive Statistics of Survival Times t^* of Multiple Myeloma.

Survival time	Mean	Median	Std Err	Std Dev	Kurtosis	Skewness
t^*	22.07	15.68	1.14	19.59	1.46	1.46

We found the pdf distribution of the t^* data to follow the three parameter-log-normal probability distribution, which is the same distribution we found for the base sample of 48 patient's survival times t. This was expected because the 300 bootstrap samples come from the 48 samples with the 3p-log-normal probability distribution, justifying the high quality of our proposed statistical nonlinear model in Chapter 3. To further support the fact that t and t^* follow the same distribution, we performed a non-parametric Kruskal-Wallis test [64, 65] to compare the difference in two survival times t and t^*. From Table 5.3, the Kruskal-Wallis rank-sum test resulted in a very large $p - value = 0.9066 \approx 1$, hence failing to reject the null hypothesis (i.e. $H_0 : \eta_t = \eta_{t^*}$), indicating no difference in the survival times t and t^*. We obtained an approximate estimate of the parameters of the 3p-log-normal parameter distribution utilizing the maximum likelihood estimation (MLE) method as we presented in [60, 33, 34]. In Table 5.4 below, we are given the estimates of the parameters of the 3p-log-normal pdf of the t^*.

Table 5.3: Kruskal-Wallis rank sum test of the Difference Between t and t^*.

Type of Test	**Survival time**	**Data: list$(t,\ t^*)$**
Kruskal-Wallis	$chi - squared(\tilde{\chi}^2) = 0.013776$	$p - value = 0.9066$

Table 5.4: Parameter Estimates for the 3p-Lognormal pdf for t^*.

Survival time	Location ($\hat{\gamma}$)	Scale ($\hat{\mu}$)	Shape ($\hat{\sigma}$)
t^*	4.4832	2.5603	1.0303

Thus, the 3p-log-normal pdf, $f(t^*)$, of the survival times of 300 MM patients is given by

$$f(t^*|\gamma, \mu, \sigma^{*2}) = \begin{cases} 0, & \text{if } t^* \leq \gamma^* \\ (2\pi\sigma^{*2})^{-1/2}(t_i{}^* - \gamma^*)^{-1} \exp\left(-\frac{1}{2}\left(\frac{\ln(t_i^* - \gamma^*) - \mu^*}{\sigma^*}\right)^2\right), & \text{if } t^* > \gamma^*. \end{cases} \tag{5.3}$$

By substituting the parameter estimates given in Table 5.4, we have

$$f(t^*) = \begin{cases} 0, & \text{if } t^* \leq 4.4832 \\ 0.38253(t^* - 4.4832)^{-1} \exp\left(-\frac{1}{2}\left(\frac{\ln(t^* - 4.4832) - 2.5603}{1.0303}\right)^2\right), & \text{if } t^* > 4.4832. \end{cases}$$

The plot of the pdf, $f(t^*)$, of the survival times of the 300 MM patients is given by Figure 5.1. With the pdf plot, we can compute the probability that the survival time of a patient diagnosed with multiple myeloma will fall between a given time t_k^* and t_{k+1}^*. For example, we can compute the probability that an MMC patients will survive between 20 months and 40 months, given by $P(20 \leq t^* \leq 40) = 0.025 - 0.008 \approx 0.017$, as shown in Figure 5.2. We interpret this as there is approximately a 1.7% probability that a patient will survive between 20 months and 40 months. On thee other hand, for survival times t, $P(20 \leq t \leq 40) = 0.019 - 0.007 \approx 0.012$.

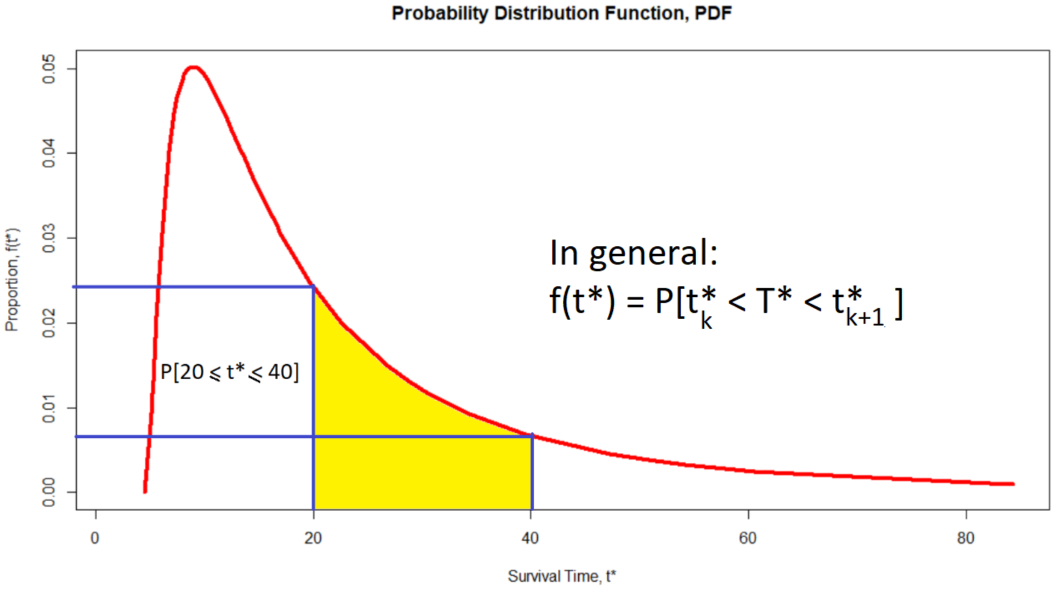

Figure 5.1: Probability Distribution Function of Survival Time, t^* of Multiple Myeloma,

The cumulative distribution of the 3p-log-normal, $F(t^*)$, of the survival times of t^* is given by

$$F_{T^*}(t^*|\gamma^*, \mu^*, \sigma^{*2}) = \frac{1}{\sqrt{2\pi}} \int_0^{t^*} \exp\left(-\frac{1}{2}z^2\right) dz = \Phi\left(\frac{\ln(t_i{}^* - \gamma^*) - \mu^*}{\sigma^*}\right). \tag{5.4}$$

Substituting the parameter estimates obtained we have,

$$F_{T^*}(t^*|\gamma^*, \mu^*, \sigma^{*2}) = P[t^* \leq T^*] = \Phi\left(\frac{\ln(t_i{}^* - 4.4832) - 2.5603}{1.0303}\right),$$

where $\Phi(.)$ is the standardized normal CDF. Figure 5.3, is a graph of the CDF of the survival times t^* of multiple myeloma patients. That is, we can estimate the probability that a patient with MMC survives up to a given time t^* from Figure 5.3. For example, the probability that an MM patient will survive up to time $t^* = 40$ months can be computed as; $F(t^* = 40) = P(t^* \leq 40) \approx 0.82$, as shown in Figure 5.2. Thus, there is about 82% chance that an MM patient will survive up to 40 months. On the other hand, we can find the probability that the patient will survive beyond 40 months to be $P(t^* > 40) = 1 - F(t^* = 40) = 0.18$. For t, $F(t = 40) = P(t \leq 40) \approx 0.83$, and $P(t > 40) = 1 - F(t = 40) = 0.17$.

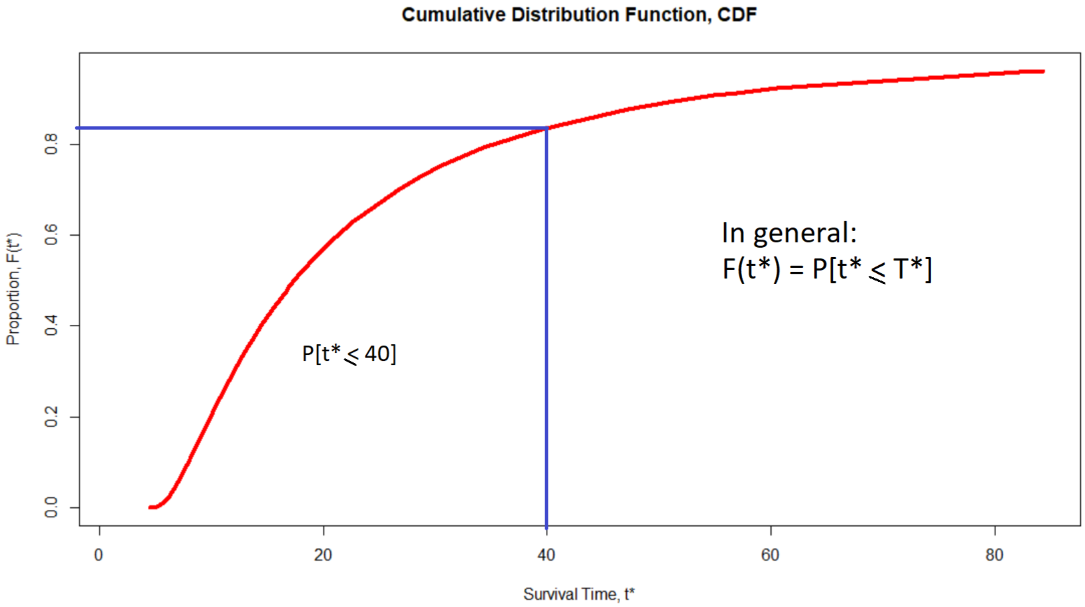

Figure 5.2: Cumulative Distribution Function of Survival Time, t^* of Multiple Myeloma

The survival function $\hat{S}(t^*)$ of the survival times t^* is given by

$$\begin{aligned}\hat{S}(t_i{}^*|\gamma^*, \mu^*, \sigma^{*2}) &= 1 - F_{T^*}(t_i{}^*|\gamma^*, \mu^*, \sigma^{*2}) \\ &= 1 - \Phi\left(\frac{\ln(t^*{}_i - \gamma^*) - \mu^*}{\sigma^*}\right).\end{aligned} \tag{5.5}$$

We substitute the estimates of the parameter given in Table 5.4,we have

$$\begin{aligned}\hat{S}(t_i{}^*|\gamma^*, \mu^*, \sigma^{*2}) &= 1 - F_{T^*}(t_i{}^*|\gamma^*, \mu^*, \sigma^{*2}) \\ &= 1 - \Phi\left(\frac{\ln(t^*{}_i - 4.4832) - 2.5603}{1.0303}\right),\end{aligned}$$

where $\Phi(.)$ is the standardized normal CDF of the survival time t^* and $\hat{S}(t^*)$ estimates the probability that a patient with multiple myeloma survive beyond a given time t^*. From Figure 5.3, we can compute the probability that an MMC patient will survive beyond 40 months; that is, $\hat{S}(t^* = 40) = P(t^* > 40) \approx 0.18$. For t, $\hat{S}(t = 40) = P(t > 40) \approx 0.17$

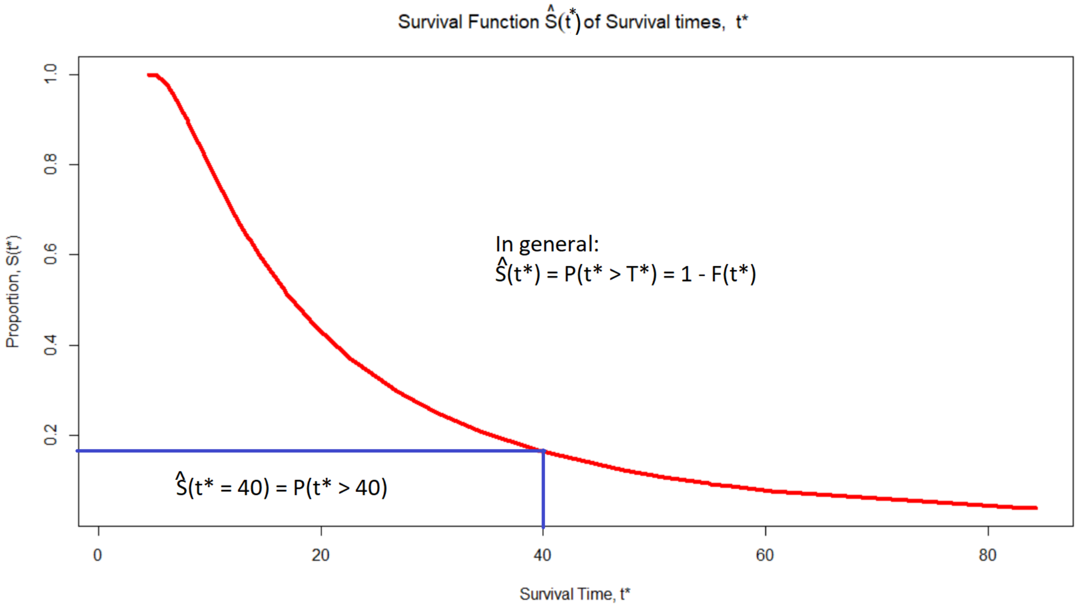

Figure 5.3: Survival Function of the Survival times, t^* of MM Patients.

5.3 Comparing the Survival Function of the Cox-PH Model with that of the Non-linear Statistical Model of Survival Times of Multiple Myeloma.

Table 5.5: Comparison of Survival Functions of NLSM and Cox-PH Model.

Rank	**Model**	R^2	**AIC**
1	Proposed model	0.9116	-65.720
2	Cox-PH	0.853	28.523

From Chapters 2 and 3, we found both the Cox-PH model and the nonlinear statistical model to be of high quality since they satisfy all the respective required model assumptions and pass all the criteria for measuring the robustness and efficiency of a high profile model. Whiles the Cox-PH model predicts the proportion of survival at a given time given the values of the significant attributable risk factors or covariates, the nonlinear statistical model predicts the real value of the survival time given values of the attributable risk factors. In Chapter 2, we found the Cox-PH model to be of utmost importance and hence recommended it as more relevant compared to the parametric and nonparametric Kaplan Meier in Chapter

1, since it takes into account the additional useful information given by the attributable risk factors of the survival time, at the time a patient is diagnosed with MMC. Now, we draw the comparison of the survival estimates of the survival times t by the Cox-PH with the survival estimates of the survival times t^* by the nonlinear statistical model of the MMC patients, given by Figure 5.4. We can see that the survival function of the nonlinear statistical model lies above that of the Cox-PH model. That is, the nonlinear statistical model consistently gives a higher prediction of the probability of survival of patients diagnosed with MM beyond a given time t^* than the Cox-PH model, making it a better choice. This is because the underlying distribution of the nonlinear statistical model survival function of the survival times t^* is based on a well-defined parametric probability distribution, which is more powerful and sophisticated than the survival function of the survival times t of the semi-parametric Cox-PH model. Therefore, it is not surprising to see that the nonlinear statistical model performs better estimating the proportion of the survival time than the Cox-PH model. In addition, the statistic measurement of model performance show that the nonlinear statistical model has better performance than the Cox-PH, given by higher R^2 and lower AIC from Table 5.5.

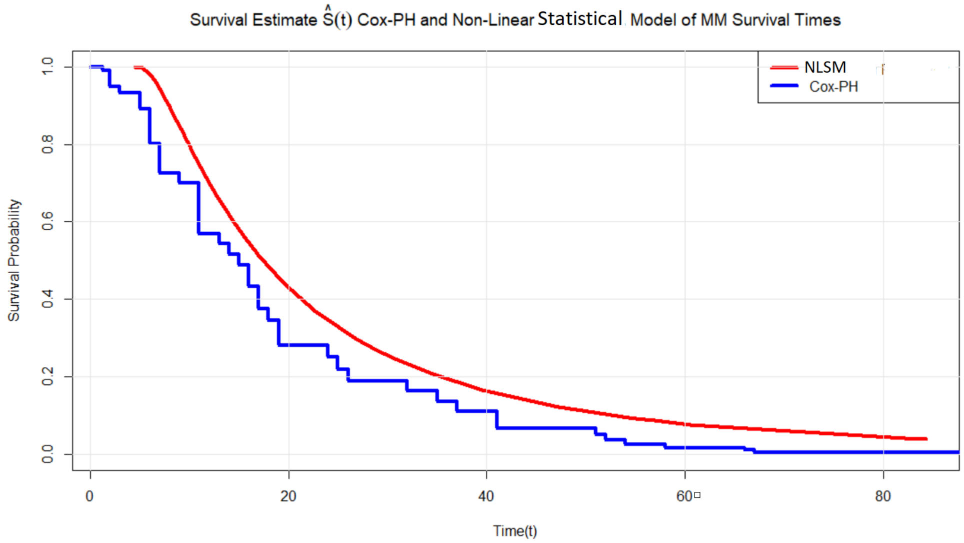

Figure 5.4: Comparison of the Survival Function of Cox-PH and the Non-linear Statistical Model of Multiple Myeloma.

5.4 Algorithm for the Nonlinear Statistical Modeling to Survival Analysis

The flowchart in Figure 5.5 shows the algorithmic process of performing the survival analysis described in this study. The process involves developing a high-quality statistical model that gives high prediction accuracy, followed by making a prediction, and finally performing parametric analysis on the predicted values and finding the survival function for estimating the proportion of survival time.

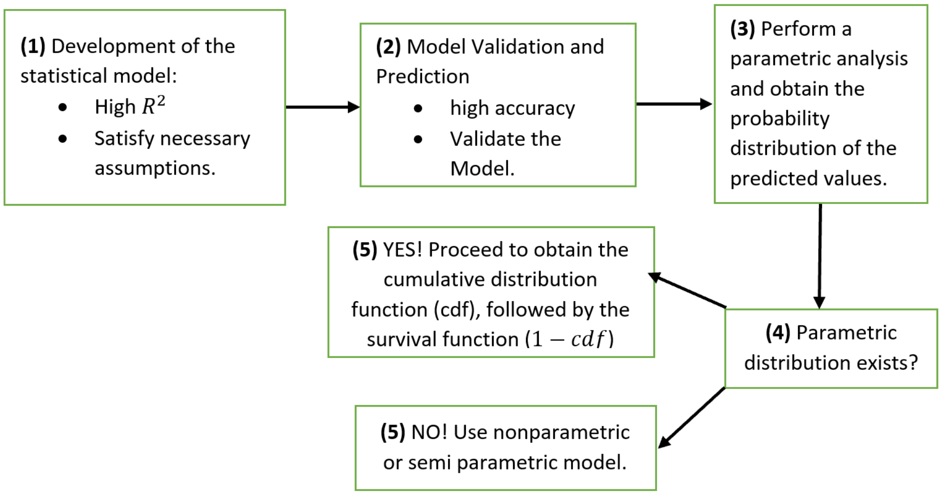
(1) Development of the statistical model:
• High R^2
• Satisfy necessary assumptions.
(2) Model Validation and Prediction
• high accuracy
• Validate the Model.
(3) Perform a parametric analysis and obtain the probability distribution of the predicted values.
(4) Parametric distribution exists?
(5) YES! Proceed to obtain the cumulative distribution function (cdf), followed by the survival function $(1 - cdf)$
(5) NO! Use nonparametric or semi parametric model.

Figure 5.5: Flow Chart of the Development of the Survival Function of the Nonlinear Statistical Model.

5.5 Discussion

In the present study of the survival times t of 48 patients diagnosed with MM, we predicted 300 survival times t^* from the final proposed statistical model in Equation (5.1) based on the bootstrap resampling method and investigated the probability distribution. We found the pdf probability distribution of the t^* follows the 3p-log-normal (same as the probability distribution of the original sample of 48 MM patients). Using the method of maximum likelihood parameter estimation as described in [60], we obtained the parameter estimates of the 3p-log-normal pdf $f_{T^*}(t^*)$ as given by Table 5.4. We found the CDF $F_{T^*}(t^*)$

by integrating the $f_{T^*}(t^*)$ with respect to t^*, and then find the survival function $\hat{S}(t^*)$ (i.e. $1 - F_{T^*}(t^*)$). The CDF estimates the survival proportion up to a given time t^*, and the survival function estimates the proportion of survival beyond a given time t^*. We then compare the $\hat{S}(t^*)$ of the nonlinear statistical model given by Equation (5.5) with the $\hat{S}(t)$ of the Cox-PH model given by equation (3.7) in Chapter 3, as shown by Figure 5.4. The comparison shows that the $\hat{S}(t^*)$ of the nonlinear statistical model provided a better estimate of the proportion of survival times than the $\hat{S}(t)$ of the Cox-PH model, given that the $\hat{S}(t^*)$ of the nonlinear statistical model is developed from the originally identified well-defined parametric probability distribution of the patients diagnosed with MM.

In Table 5.1 of the ranking of the significant attributable risk factors identified by the two models, the Cox-PH model identified infections and serum calcium as an interaction term, ranked fifth, as significantly contributing to the proportion of the survival times of MM patients. However, those two risk factors were individually identified to be significantly contributing to the survival times by the nonlinear statistical model and ranked third and fourth respectively. The ranking process of the Cox-PH model is based on the prognostic effect of the risk factor on the survival time using the hazard ratio, and the nonlinear statistical model ranking of the risk factors is based on the percentage of contribution to the variability in the survival time explained by the significantly identified attributable risk factors (i.e. the coefficient of determination, R^2). Both the hazard function and the R^2 can play a similar role in measuring the prognostic effect of a given risk factor on the survival time. However, we recommend the use of the R^2 along with the R^2_{adj} because it measures the entire variability of the survival times of MM explained by the risk factors with a high degree of accuracy.

It is very important to recognize that both models are high profile considering the quality involved in the model building process. However, in reality, medical personnel and patients would be more concerned about the real value of the survival time rather than the probability of surviving. Also, we would be more concerned with the percentage of contribution of a risk factor to the survival time than whether it is a good or bad prognostic factor to the survival time. Thus, making the ranking of the risk factors by the nonlinear statistical model more relevant. Moreover, developing the Cox-PH model is more difficult satisfying the assumptions and finding the interaction between covariates. Douglas G. Altman and Bianca L. De Stavola (1994) [66] presented the practical problems in fitting a proportional hazards model. Ian Ford, John Norrie, and Susan Ahmadi (1995) [67] also assessed model inconsistency illustrated by the Cox-PH model. These studies allow us to strongly support the robustness of using the statistical model approach to survival analysis, which provides more flexibility than the Cox-PH model. The present finding shows that we can obtain a better and accurate prediction of the proportion of survival time as long as we can find a well-defined parametric probability distribution that characterizes a given cancer survival data. Given that our objective is to maximize the survival times, the nonlinear statistical model is better for improving the therapeutic/treatment strategy of maximizing the survival times of multiple myeloma patients than the Cox-PH model.

5.6 Contribution

In the present study, we have demonstrated that both the Cox-PH model and the nonlinear statistical model are of high quality and useful. The two models predict the proportion of survival time with a high degree of accuracy. However, we recommend the nonlinear statistical model over the Cox-PH model because it offers a better model performance and prediction of the survival probability of the MM patients, as shown in Figure 5.4. The nonlinear statistical model does not only provide a better prediction estimate of the survival probability, but it also provides us with several useful outcomes. (1) We can predict the real values of the survival time of a patient given the significant attributable risk factors and the interaction term. (2) We can rank the attributable risk factors and the interaction term according to the percentage of contribution to the survival time. (3) We can perform surface response analysis for the maximization of the survival time, given the values of the risk factors and the interaction. (4) We can generate confidence intervals for the survival time. (5) We can also perform parametric analysis of the predicted survival values and obtain better survival estimates (i.e. the probability that a patient survival beyond a given time) than the survival estimates from the popularly known traditional survival models. On the other hand, we can only obtain the outcome (5) from using the Cox-PH model. The present study provides therapeutic/treatment significance for further improvement in the survival times of patients diagnosed with multiple myeloma.

Chapter 6: Real Data-Driven Nonlinear Analytical Model for Corn Production in the U.S.

The production of corn plays a major role in the economics of the United States. The U.S. is noticeably the world's leading producer of corn, with corn serving many purposes in the economics of ethanol production, beverage alcohol production, livestock feeds, cereals, sweeteners, among others. Planning rationally and judiciously in distributing economic resources effectively and efficiently can result in maximizing the returns from corn production, hence investor motivation and sustaining the U.S. as the world's leading producer of corn. In this chapter, we developed a data-driven multivariate nonlinear statistical model that identified seven significant individual contributable factors and six significant contributable interaction terms that accurately predict the returns from corn production in the U.S. from 1975-2018. The proposed statistical model is of high quality and accurately predicts the returns, satisfying all assumptions, residual analysis, and goodness-of-fit tests. The identified contributable factors are ranked according to individual factor percentage of contribution to the returns in descending order of magnitude. The opportunity cost of land was ranked first; followed by fuel, lube and electricity; custom services; the market value of the grain; fertilizer; etc; and the interaction repair & operating capital ranked last, thirteenth contributable factor. The proposed model performs better compared to other least square models. The present study would offer corn farmers or industries strategy to maximizing the returns from corn production, and further, stimulate investor confidence and sustaining the U.S. as the world's producer of corn. This study has been considered for patent[105].

We organized this chapter as follows: Section 6.1 introduces and review some literature of studies on the returns of corn production; Section 6.2 presents the description of data used in this study; Section 6.3 presents the parametric analysis of the US corn production returns; Section 6.4 presents the statistical/analytical modeling and analysis for the returns of corn production; Section 6.5 discusses the findings in this study; and finally, the research contributions of this chapter is captured in Section 6.6.

6.1 Introduction

Corn production (also known as “maize”) plays a significant role in the United States economy. The U.S. is the largest corn producer in the world [68], utilizing 96,000,000 acres (39,000,000 ha) of land reserved for corn production. Corn is the most widely produced crop and feed grain in the U.S., accounting for over 95% of total production and use. Corn has a wide range of usefulness to both humans and animals (especially livestock); among these are food and industrial products including cereal, alcohol, sweeteners, and byproduct feeds, and energy ingredient in livestock feed [69]. In 2017, the U.S. grew 15.1 billion bushels of corn production, and Iowa State is the largest producer of corn, producing 2.7 billion of those bushels. The U.S. corn growth is dominated by west/north central Iowa and east central Illinois with approximately 13% of its annual yield is exported [71]. It is reported for the year 2013-2014 that the total production of corn in the US was 13.016 billion bushels, of which the major use is for manufacture of ethanol and its co-product (Distillers’ Dried Grains with Solubles) accounting for 37% (27% + 10%), or 4,845 million bushels (3,552 + 1,293) [70]. Even the maize cobs which in mostly serve as a by-product are used as a biomass fuel source such as specialized corn stove (similar to the wood stove) [80]. For the year 1950-1959, the final estimated production was 3 billion bushels, and the recent years production is 9 billion bushels per year [72].

The wild difference in corn production in the U.S. over the rest of the world is that farmers obtain 20% more corn per acre than any part of the world [73]. Most U.S. corn production farming practice is based on irrigation and implementing soil conservation measures which have reduced soil erosion. Experts believe that Iowa has become the world’s largest producer

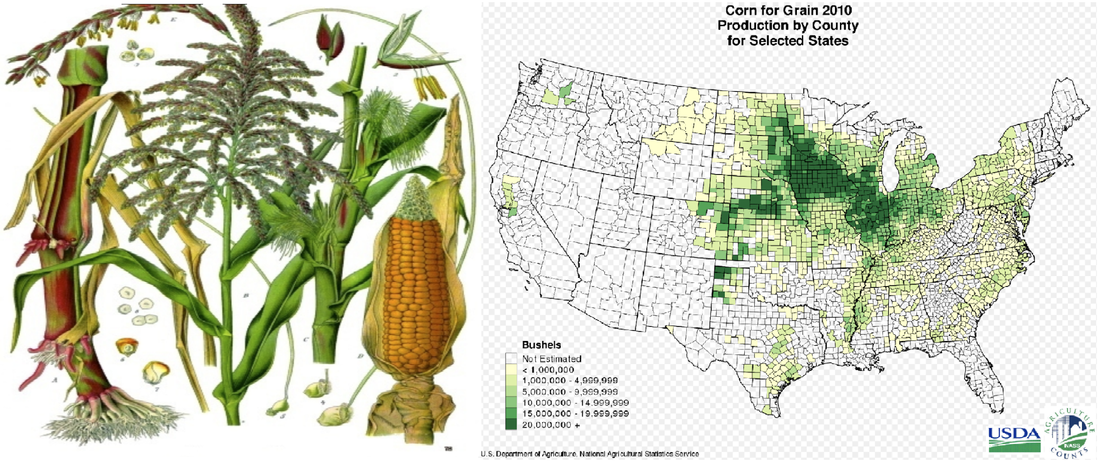

Figure 6.1: Corn Production in United States

of corn and the home of most of the world's finest corn production farmers mainly because it houses the most fertile topsoil on the planet [70]. One vital reasons for the increasingly high production of corn in the U.S. is due high subsidies of corn [74].

There is a high acceleration of corn demand in the U.S. The average American on average spends US$267 annually on purchasing corn [75]. The overwhelming demand for maize is partly due to the use of maize for biofuel production. In the U.S., large portion food prices (80%) are affected by the cost of transportation, production, and marketing. As a result, the use of maize as a biofuel has shifted farmers from the production of other food crops to maize production so as to meet the growing demand for maize and increase their profitability. This has resulted in a decrease in the supply of other food crops and increases corn prices [81]. The value of the corn produce depends on the number of bushels, the quality of the corn, and varies from one location to the other [76]. The value of corn in the U.S. is continuously increasing, largely due to the higher demand and reliance on corn [77]. In general, factors such as weather and economic predicaments/crisis may influence the value or price of corn produced at a particular period [78, 79], which in turn influences the returns/profit made from corn production. Maize is usually bought and sold by investors and price speculators as a tradable commodity using corn futures contracts, which directly/indirectly affects the returns earned on maize production.

The returns on corn production may be negative, positive, or zero. Negative returns, also known as net loss, occur when the cost of the corn production exceeds the revenue/income earned. Positive returns occur when the revenue earned exceeds the cost of production of the corn. Zero returns, in other words, known as "break-even", occurs when the cost equals the revenue earned. Thus, the return of production is mainly influenced by cost and revenue. The cost of production, often called the "Total Cost (TC)" plays a major role in the returns from corn production. TC consists of the fixed cost (FC) (the cost incurred on fixed factors/inputs such as capital, equipment, farmland, etc.) and the variable cost (VC) (the cost incurred on variable factors such as cost of labor, farm inputs, etc.). In production, Fixed factors remain unchanged as output changes, but variable inputs or factors change with varying units of output. TC is influenced by marginal cost (MC) (the addition to TC from producing one more unit of product) of production.

The revenue referred to as "Total Revenue (TR)" is the earnings/income often from the sale of the maize. The TR of production is also influenced by the marginal revenue (MR) (the addition to TR from the sale of one more unit of product). TR largely depends on the supply, demand, and market value or price of the maize at the time of sales. In 2017/18, U.S. domestic demand for corn increased boosted by the production of ethanol and feed used, according to USDA [82]. In general, if the supply of corn remains unchanged, an increase in demand creates a natural shortage causing the value of corn to increase in the short run. On the other hand, if the supply of corn remains the same, a decrease in demand creates excess supply or natural surplus, leading to a fall in the price of corn in the short-run period. Therefore, this explains the fundamental principle of demand and supply in influencing prices of corn.

Other fundamental factors affecting and impacting prices of corn are the weather, ethanol production, and the amount of bushels/acreage planted. Wescott and Jewison (2012), researched and compiled the results of the impact of the 2012 drought on the yield and price of corn. There was a significant reduction in yield compared to the previous year, causing the price of corn to rise. Gardebroek and Hernandez (2012) [84], researched the volatility transmission in ethanol, oil, and corn prices between 1997 and 2011. In their study, they found significant volatility spillovers from corn to ethanol, but not from ethanol to corn. According to Wallander, Claassan, and Nickerson (2011) [85], planted corn acreage and expected corn yield are some major factors affecting corn prices, which increased along with the prices of corn.

The profitability of a firm into production is often said to be determined by MC and MR. The firm is said to be making a profit if $MR > MC$, losing if $MR < MC$, and at the profit-maximizing stage if $MR = MC$. Figure 6.2 demonstrates the relationship between MR and MC in determining the profit/returns of a firm in production. The profit optimizing output is Q, the point where $MR = MC$ (MR intersects MC). At this point, the firm into corn production can increase the amount of corn production as long as the added revenue from producing one more bushel/acreage of corn outweighs the added cost of producing one more bushel/acreage.

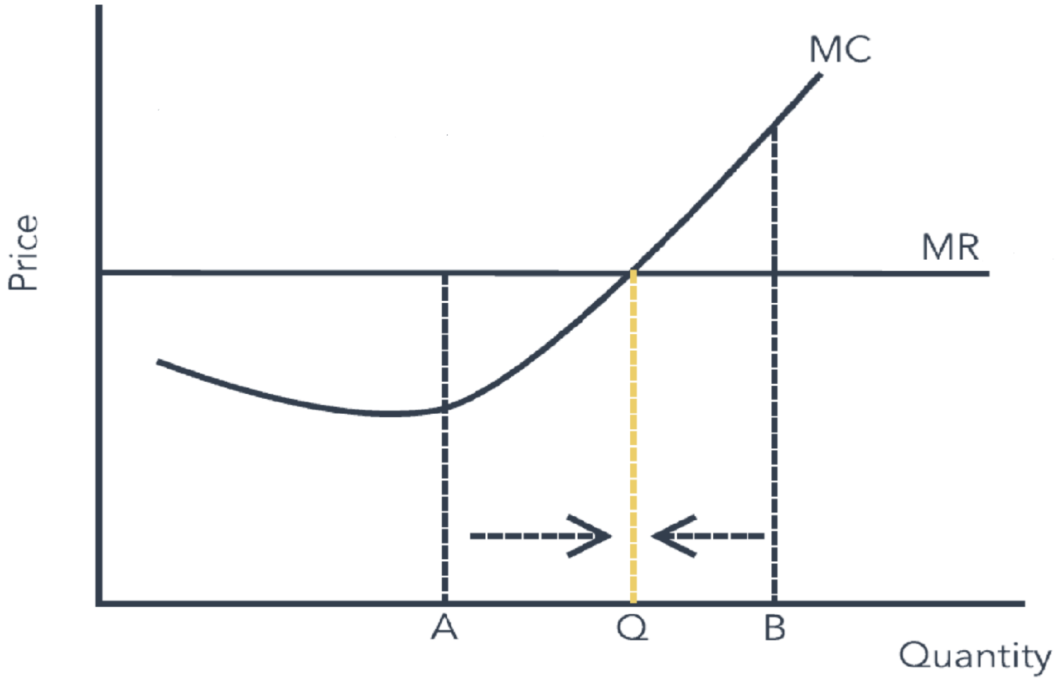

Figure 6.2: Relationship between Marginal Revenue (MR) and Marginal Cost (MC)

The above profit/return optimization principle suggested by most economists and used by most firms in production has some limitations. In the real world, it is tedious to know exactly your MR and MC of the last products sold [116]. For instance, it is difficult for firms in production to know the price elasticity of demand for their product which induces the MR. The above concept of profit optimization also depends on how other firms react to the price, especially in a perfectly competitive market, and if demand is inelastic. All this being equal ("ceteris paribus"), if you are the only firm to increase the price, demand will be elastic, and hence affecting the returns negatively. Therefore, the above profit-maximizing rule may not work in most cases, given that there are several other firms into corn production. The price, demand, and supply of corn can be affected by several other factors. These other factors drive the TR and TC of production of corn, thereby affecting the returns. For us to be able to determine these factors would provide a tremendous leap towards controlling or manipulating the TR and TC and hence optimizing the returns from corn production.

There are three variables of interest in production, i.e. price (P), quantity (Q), and cost (C). These 3 variables determine the total cost (TC) and total revenue (TR) of production, and hence the profitability (returns) of the firm into corn production. TC and RT, on the other hand, is determined by several factors. Therefore, the return of production (R_p) (Profit $) is a function of several attributable variables of TC and TR. Thus, one may ask what drives the returns of a firm into corn production? Our goal is to develop a data-driven nonlinear statistical model that predicts the returns from corn production in the U.S., given the set of values of the significant attributable variables.

In the present study, we developed a real data-driven statistical model of the significant attributable risk factors of corn production in the United States. The data consist of the returns from corn production from 1975-2018 in the U.S. There are 25 variables or attributable factors believed to be contributing to the returns from corn production by the United States Department of Agriculture (USDA). The data was filtered to fulfill all the analytical modeling assumptions. We identified the significant attributable variables or risk factors, and interactions contributing to the returns/profit from corn production. The significant attributable factors, including the interactions identified, were ranked based on the percentage of contribution to the returns from corn production, using the coefficient of determination (R^2) of the returns. The quality and accuracy of the proposed model was assessed based on the R^2 along with $R^2_{adjusted}$ statistic, the Akaike information criterion (AIC) of model selection, the prediction error sum of squares (PRESS), the root mean square error (RMSE), the variance inflation factor (VIF), the residual analysis, and comparison of the model with other models.

6.2 Data Description

The data used in this study was obtained from the United States Department of Agriculture (USDA) Economic Research Service. The data set consists of 25 attributable variables of the returns from corn production in the USA from 1975 to 2018. Figure 6.3 below displays the non-stationary time series of the returns on corn production in the US. We can see that on the average the U.S. experienced negative returns on corn production from 1975 to 2006, and from 2007 to 2013 there were positive returns on average. The U.S. experience the lowest returns on corn production in 1999 and the greatest returns in 2011. In 2012, the returns continuously decrease from positive returns to negative returns until 2014. Thereafter, the returns have remained negative even though it has been increasing up to the most recent returns in 2018. The decrease in the returns from 2012 was probably due to the effect of the acute drought on the yield and price of corn experienced in 2012. We further observed that the returns on corn production were rising after a fall in 2005 until 2009, where there was a decrease before a rise in 2010. The volatility of the returns during this period was probably largely due to the economic recession in 2007-2008. Table 6.1 below shows the detailed description of the 25 various variables presumed to be contributing to the returns of the U.S. corn production given by Figure 6.3.

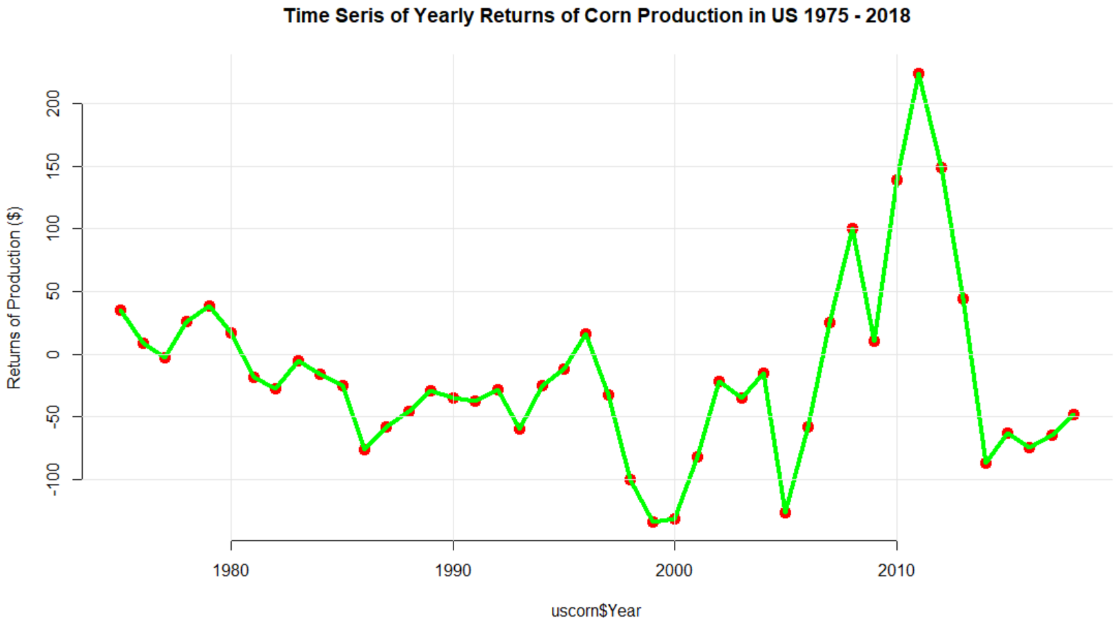

Figure 6.3: Time Series of the returns from corn production in U.S. 1975 - 2018

6.3 Parametric Analysis of the Production Returns

Before carrying out any statistical analysis and modeling, it is imperative to first perform parametric analysis. Parametric analysis guides us to find the right probability distribution of the response variable we are modeling or analyzing. It also guides us to make the correct decision on whether to transform the response variable, as well as for deciding on the correct choice of transformation. It is a statistical fallacy to employ non-parametric analysis or test if parametric distribution exists. It is important to recognize that parametric analysis is more robust and efficient than non-parametric analysis of any kind. However, a non-parametric test is desirable if the given data distribution has no parametric form. A good and detailed review and reference on parametric analysis can be found in our previous study on parametric and non-parametric analysis of the survival times of multiple myeloma patients [60].

The parametric analysis usually starts with a graphical representation of a histogram and display of descriptive statistics to investigate the probability distribution of the product returns. Figure 6.4 displays the histogram of the returns on corn production. In Table 6.2, we show the descriptive statistics of the product returns. The descriptive statistics show a mean

Table 6.1: VARIABLE RECORDED - Corn Production Cost and Returns Per Planted Acre, Excluding Government Payments from 1975 - 2018

Symbol	Variable Name: Contributable Variables/Risk Factors
R_p	Production returns
X_1	Value of primary product grain
X_2	Value of secondary products silage
	Operating Costs
X_3	Seed
X_4	Fertilizer
X_5	Chemicals
X_6	Custom services
X_7	Fuel, lube, and electricity
X_8	Repairs
X_9	Purchased irrigation water
X_{10}	Interest on operating capital
	Allocated Overhead
X_{11}	Hired Labor
X_{12}	Opportunity cost of unpaid labor
X_{13}	Capital recovery of machinery and equipment
X_{14}	Opportunity cost of land
X_{15}	Taxes and insurance
X_{16}	General farm overhead
	Supporting Information
X_{17}	Yield (bushels per planted acre)
X_{18}	Price (dollars per bushel at harvest)
X_{19}	Enterprise size (planted acres)
X_{20}	Dry-land (percent acres)
X_{21}	Irrigated (percent of acres)
	Economic Costs
X_{22}	Variable cash expenses
X_{23}	Capital replacement
X_{24}	Operating Capital
X_{25}	Other non-land capital

returns from production of -16.83, which is greater to the median returns from production of -26.28, indicating the production returns is right-skewed, as shown by the histogram, and given by the positive skewed and kurtosis values (i.e. skewed value is 1.17 and kurtosis value is 2.06). See [60] for an extensive explanation of skewness and kurtosis. The histogram further shows that most of the production of corn returns is between -120 and 40 dollars per planted acre.

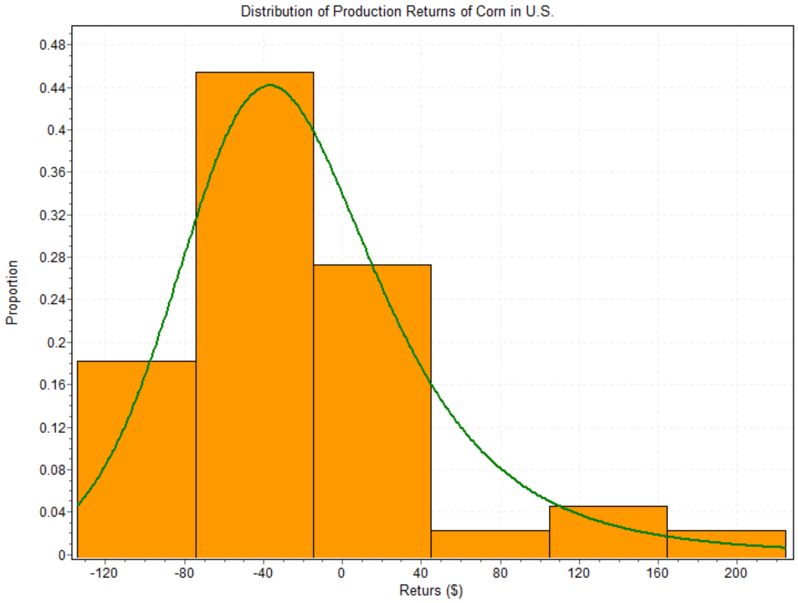

Figure 6.4: Histogram Showing the Returns on Corn Production in U.S. 1975-2018

Table 6.2: Descriptive Statistics of Corn Production Returns

Mean	Median	Std Err	Std Dev	Kurtosis	Skewness
-16.83	-26.28	10.68	70.82	2.06	1.17

Table 6.3: Goodness-of-fit Test of the 3P-Lognormal Distribution of the Survival Time.

Type of Test	$p - value$
Kolmogorov-Smirnov	0.86244
Anderson-Darling	0.54708
Chi-Squared	0.83494

After a close assessment of Figure 6.4 and Table 6.2, we concluded that the probability distribution that characterize the probability behavior of the returns from corn production in the United State from 1975 - 2018 follows the three-parameter log-logistic probability distribution. In Table 6.3, we perform three different goodness-of-fit tests to further assess the validity of the subject probability distribution. The test was based on the Kolmogorov-Smirnov, Anderson-Darling and Chi-Squared goodness-of-fit test. The three tests revealed a large $p - value$, meaning that we do not reject the null hypothesis that the distribution of the returns on corn production follows the 3p-log-logistic probability distribution. Thus, given random production returns, denoted by r, the pdf of the 3p-log-logistic probability distribution is given by

$$f(r;\alpha,\beta,\gamma) = \begin{cases} 0, & \text{if } r \leq 0 \\ \frac{\alpha}{\beta}\left(\frac{r-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{r-\gamma}{\beta}\right)^{\alpha}\right)^{-2}, & \text{if } r > 0 \end{cases} \tag{6.1}$$

where $\alpha > 0$ denotes continuous shape parameter, $\beta > 0$ is the continuous scale parameter, and γ is the continuous location parameter and $\gamma \leq r \leq +\infty$. Note that $\gamma \equiv 0$ gives the two-parameter log-logistic distribution. We employed the maximum likelihood estimation (MLE) method to estimate the parameters α, β and γ. The MLE method was used because it is more robust than other methods like the least-squares estimation and the method of moment [32]. See [60] for further information and references on the MLE method of parameter estimation. To compute the MLE of the parameters, we compute the derivative of the log-likelihood function and set to zero. For n observation from 3p-log-logistic probability distribution, denoted by $r_1, r_2, ..., r_n$, the likelihood function can be written as

$$\begin{aligned} L(\alpha,\beta,\gamma|r_i) &= \prod_{i=1}^{n} f(r_i|\alpha,\beta,\gamma) \\ &= \prod_{i=1}^{n}\left[\frac{\alpha}{\beta}\left(\frac{r-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right)^{-2}\right] \\ &= \left(\frac{\alpha}{\beta}\right)^{n}\prod_{i=1}^{n}\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right)^{-2}, \quad \forall r_i > \gamma. \end{aligned} \tag{6.2}$$

Now, we take the natural log of the likelihood function in Equation (6.2), given by

$$\begin{aligned}\ln Ł &= \ln L(\alpha, \beta, \gamma | r_i) \\ &= n \ln(\alpha) - n\ln(\beta) + (\alpha - 1)\sum_{r=1}^{n} \ln\left(\frac{r_i - \gamma}{\beta}\right) - 2\sum_{r=1}^{n}\ln\left(1 + \left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right).\end{aligned} \tag{6.3}$$

By differentiating Equation (6.3) with respect to α, β and γ, we have

$$\frac{\partial \ln Ł}{\partial \alpha} = \frac{n}{\alpha} + \sum_{r=1}^{n}\ln\left(\frac{r_i-\gamma}{\beta}\right) - 2\sum_{r=1}^{n}\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\ln\left(\frac{r_i-\gamma}{\beta}\right)\left(1+\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}, \tag{6.4}$$

$$\frac{\partial \ln Ł}{\partial \beta} = -\frac{n}{\beta} - (\alpha-1)\left(\frac{n}{\beta}\right) + 2\frac{\alpha}{\beta}\sum_{r=1}^{n}\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\left(1+\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}, \tag{6.5}$$

and

$$\frac{\partial \ln Ł}{\partial \gamma} = \left(-\frac{\alpha-1}{\beta}\right)\sum_{r=1}^{n}\left(\frac{r_i-\gamma}{\beta}\right)^{-1} + 2\alpha\sum_{i=1}^{n}\left(\frac{\left(\frac{r_i-\gamma}{\beta}\right)}{r_i-\gamma}\right)\left(1+\left(\frac{r_i-\gamma}{\beta}\right)^{\alpha}\right)^{-1}. \tag{6.6}$$

By setting Equation (6.4) - (6.5) to zero, the MLEs of the parameters of the 3p-log logistic distribution of the production returns given by Table 6.4 below.

Table 6.4: Parameter Estimates for the Three-Parameter Log Logistic Probability Distribution of the Returns of Production

Location ($\hat{\alpha}$)	Scale ($\hat{\beta}$)	Shape ($\hat{\gamma}$)
6.4564	223.59	-249.91

We substitute the parameter estimates in Table 6.4 into Equation (6.1) and obtained the pdf of the 3p-log-logistic probability distribution of the product returns, given by

$$f(r) = \begin{cases} 0, & \text{if } r \leq 0 \\ 0.0289\left(\frac{r+249.91}{223.59}\right)^{5.4564}\left(1+\left(\frac{r+249.91}{223.59}\right)^{6.4564}\right)^{-2}, & \text{if } r > 0. \end{cases} \tag{6.7}$$

After finding the pdf we can compute the cumulative frequency distribution, cdf, by taking the integral of the pdf in Equation (6.1). The cdf allows us to estimate the probability that the production firm obtained a certain amount of returns (i.e. $F_R(r) = P(r \leq R)$). So, the cdf of the 3p-log-logistic probability distribution of the product returns is given by

$$\begin{aligned} F_R(r;\alpha,\beta,\gamma) &= \int_0^r f(r;\alpha,\beta,\gamma)dr \\ &= \int_0^r \frac{\alpha}{\beta}\left(\frac{r-\gamma}{\beta}\right)^{\alpha-1}\left(1+\left(\frac{r-\gamma}{\beta}\right)^{\alpha}\right)^{-2} dr \\ &= \left(1+\left(\frac{r-\gamma}{\beta}\right)^{\alpha}\right)^{-1}. \end{aligned} \tag{6.8}$$

Substituting the parameter estimates given by Table 6.4, we have the cdf to be given by

$$F_R(r) = \left(1+\left(\frac{r+249.91}{223.59}\right)^{6.4564}\right)^{-1}. \tag{6.9}$$

We can also obtain the reliability of the production returns by deducting the cdf from one. Thus, the probability that the production firm yields beyond a certain amount of returns ($\hat{R}(r) = P(r > R) = 1 - P(r \leq R)$). Therefore, the reliability of production returns, (r) is given by

$$\hat{R}(r;\alpha,\beta,\gamma) = 1 - F_R(r) = 1 - \left(1+\left(\frac{r+249.91}{223.59}\right)^{6.4564}\right)^{-1}. \tag{6.10}$$

6.4 Developing the Statistical Model for the Returns of Corn Production

After the parametric analysis, we found the returns from corn production in the U.S. from 1975-2018 was right-skewed and followed the three-parameter log-logistic probability distribution. Now, we proceeded to develop a multivariate nonstationary statistical regression model for the product returns taking into consideration the 25 attributable factors presumed to be contributing to the returns from corn production in the U.S. given in Table 6.1. The statistical model was developed based on satisfying the major assumptions of the multivariate linear regression model. Firstly, there should be a linear relationship between the response, r (corn production returns), and the explanatory or attributable variables, given by

$$r_i = \tau + \sum_{i=1}^{k} \delta_i X_i + \sum_{i \neq j=1}^{k} \gamma_{ij} X_i X_j + \epsilon_i, \tag{6.11}$$

where the response variable $r_i = (r_1, ..., r_n)^T$, $\tau = (1, ..., 1)^T$ is the intercept or constant term, $\beta_i = (\delta_1, ..., \delta_k)^T$ is the coefficient parameter of the attributable factors X_i's, γ_{ij} is the coefficient parameter of interaction between i^{th} and j^th attributable risk factors, $\epsilon_i = (\epsilon_1, ..., \epsilon_n)^T$ denotes the model residual error term, $k = 25$ is the number of attributable factors given by Table 6.1, and $n = 43$ the sample size from 1975-2018. We assessed the Linearity by investigating the correlation matrix between the response and the continuous attributable factors given by Figure 6.5. The values of the correlation coefficient are bounded between -1 and 1, where -1 is a perfect negative correlation, and 1 is a perfect positive correlation. In the correlation diagram, the dark blue color means a strong positive (+ve) correlation (linear relationship/association) between the two variables, the light blue means moderate +ve correlation, and the white color means little or no correlation. The deep brown depicts strong negative (-ve) correlation and light brown color implies moderate -ve correlation. We can see that there is a presence of a moderate to a strong linear relationship between the response and most of the predictors, although some predictors X_i, $i = 2, 9, 10, 21$, showed little or no correlation with the response. We also see that there is a very strong correlation between some of the predictors, which can contribute to multicollinearity in a regression model. The problem of linearity would be addressed in the course of the model building process.

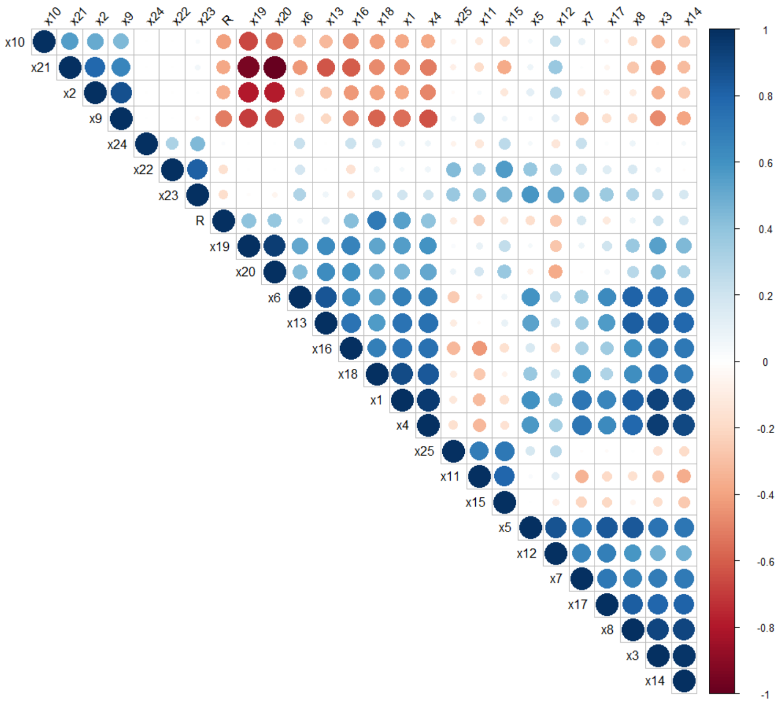

Figure 6.5: Correlation Matrix of the Response,R_p, and the Attributable Variables

Next, we investigated the assumption of multivariate normal distribution. The residuals of a linear regression model is expected to follow the Gaussian normal probability distribution, $\epsilon \sim N(0, 1)$ as $n \to \infty$. We noted discrepancies in the data given by skewed response r (see Figure 6.4 and Table 6.2), lack of linearity between the response and some predictors, and a near-perfect correlation between some predictors. However, we proceeded to fit a linear regression model to the original data to assess other assumptions and how well the 25 attributable variables fit the returns of production. A Durbin Watson test for autocorrelation shows that the errors are uncorrelated. However, there was a violation of the homoscedasticity of the residual. Another problem encountered from the initial fitted model was an insignificant intercept. Every linear regression model must have a significant intercept to adjust for the linearity between the response and the attributable variables. Not having a significant intercept implies that if all the predictors are zero, then the response variable is zero (i.e. no other variables contribute to the response apart from the given predictors); thus, a highly statistical fallacy.

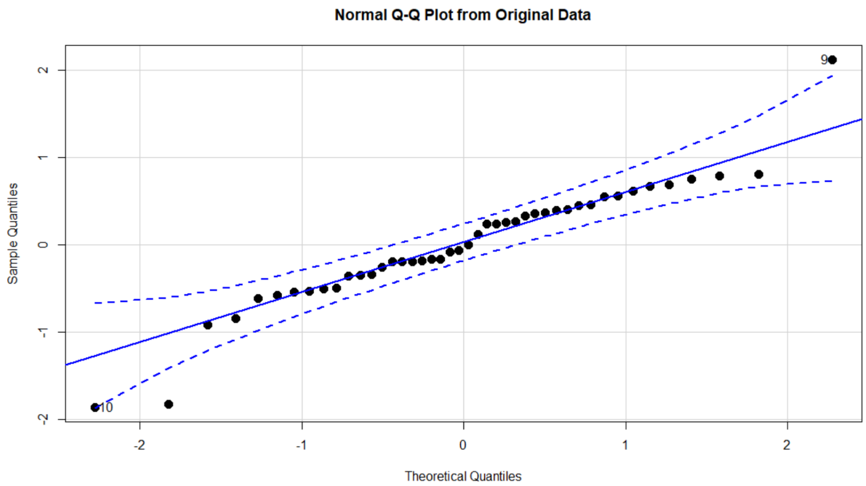

Figure 6.6: Testing the Normality of Residuals of the Original Data

In Figure 6.6, we tested for the assumption of the Gaussian normal probability distribution of residual errors in the model developed from the original data. We can see that most points fall within the 95% confidence bound, but few points are falling outside which could distort the normality of the errors. We performed a formal test from Shapiro Wilks resulting in a small $p - value = 0.02096$, which implies the violation of normal probability distribution.

Another problem of concern in statistical linear regression modeling is multicollinearity which should be giving close attention. There have been several arguments regarding the concern for multicollinearity if the main objective of the model is for prediction, such as our case for the prediction of the returns of production. We believe multicollinearity must be closely assessed, because in some cases it may affect the significance of the parameter coefficients of predictors, and distort the efficiency and accuracy of predictions by the statistical model leading to wrong or misleading decisions. Although, there are findings and suggestions by some researchers that multicollinearity does not affect the precise prediction and goodness-of-fit of the statistical regression model if it is mainly to make predictions [87, 88]. After fitting the initial model, we tested for the presence of multicollinearity among the predictors using the variance inflation factor (VIF). We found multicollinearity in some of the predictors. However, it is not surprising given the fact that in general, most economic variables tend to be highly correlated. Extremely high multicollinearity should not be overlooked even if the model is mainly for prediction.

Given the number of discrepancies encountered, we apply the Johnson transformation to transform the response variable, given by

$$R_T = \gamma + \eta \ln \left(\frac{r - \epsilon}{\lambda + \epsilon - r} \right), \tag{6.12}$$

where R_T denotes the transformed response, r is the non-transformed response, and γ, η, ϵ and λ are the transformation parameters. The Johnson transformation was chosen because it gives a better transformation of the response, R_p, than other forms of transformation like the log transformation and Box-Cox transformation. After transforming the response variable, we refit the model with all the 25 attributable variables, including their two-way interactions to the transformed response,R_T. We then employed the step-by-step backward elimination model selection method to select the significant contributing variables and the interactions. The backward elimination method is a more efficient model selection technique because the resulting mean square error (MSE) is less biased and prevents model overfitting thereby enhancing the model prediction performance. The method uses the Akaike information criterion, AIC to select the best model with the least AIC. The AIC estimates the relative amount of information loss in the model; hence the smaller the AIC the better the fit of the model. Therefore, given that we applied the best form of transformation to the response variable, and adopted the best model selection procedure in selecting the significant attributable variables and interactions, which resulted into the final model with the least AIC given by

$$
\begin{aligned}
R_T = {} & 9.424e^{-01} + 2.801e^{-02}1X_1 - 8.737e^{-02}X_4 \\
& - 6.225e^{-02}X_6 - 3.589e^{-02}X_7 - 1.447e^{-01}X_{11} \\
& - 5.173e^{-02}X_{14} + 2.082e^{-01}X_{24} - 4.223e^{-03}X_1 * X_{18} \\
& + 1.505e^{-02}X_4 * X_{18} + 9.248e^{-05}X_4 * X_{19} - 1.238e^{-02}X_4 * X_{22} \\
& + 6.140e^{-03}X_{14} * X_{18} - 9.953e^{-03}X_8 * X_{24},
\end{aligned}
\tag{6.13}
$$

along with the transformed response of the returns of production, R_T, given by

$$
R_T = -0.2680 + 0.9173 \ln\left(\frac{r + 38.0612}{-0.0708 - r}\right). \tag{6.14}
$$

The above final model in Equation (6.13) is our proposed model for the returns from corn production in the United State from 1975-2018 and includes seven individual attributable variables and six interaction terms. Where '$*$' denotes an interaction between two attributable variables. The model has an $R^2 = 0.9822$ along with $R^2_{adj} = 0.9745$, indicating a very good model. The coefficient of determination, R^2 along with the R^2_{adj} provides the proportion of variation in the response, R_T, explained by the seven identified significant attributable factors and the six interaction terms in Equation (6.13). Therefore, the higher R^2, the better the goodness-of-fit of the model. But the model must first fulfill all the required assumptions including having little or no multicollinearity. The analytical form of R^2 and R^2_{adj} is given by

$$
R^2 = 1 - \frac{SSE}{SST},
$$

and

$$
R^2_{adj} = 1 - \frac{SSE/(n-k)}{SST/(n-1)},
$$

where $SST = \sum_i (r_i - \bar{r})^2$, is called the total sum of squares is the proportional to the sample variance, and equals to the sum of SSR and SSE. $SSR = \sum_i (\hat{r}_i - \bar{r})^2$ is the regression sum of squares representing the variation explained by the proposed model and $SSE = \sum_i (r_i - \hat{r}_i)^2 = \sum_i e_i^2$; and r_i are the corn returns, $\bar{r} = \frac{1}{n} \sum_i^n r_i$, $\hat{t}_i$ is the estimated corn returns. Generally, the R^2 has the problem of increasing by increasing the number of parameters or predictors in the model. So, it is recommend to state R^2 along with R^2_{adj} to adjust for the degree of freedom of the model ($R^2_{adj} \leq R^2$). Note that $n - k$ denotes the degree of freedom of SSE and $n - 1$ is the degree of freedom of SST. The closer the R^2_{adj} to R^2, the better the good the goodness-of-fit of the model.

To use the proposed model Equation (6.13), we first put the values of the identified attributable variables and the interaction terms into the model, which results in the transformed response or corn returns, R_T. To obtain the real value of the corn returns, r, we find the anti-transformation or the r given by Equation (6.14). Thus, given the values of the identified factors contributing to corn production returns in the proposed model, we can precisely predict the returns to be earned with about 98% degree of accuracy. Table 6.4 below shows the ranking order of the statistical significance of each of the identified attributable variables and interaction terms according to the percentage of contribution to the returns from corn production in the U.S. from 1975-2018 based on the R^2 statistic. The opportunity cost of land is rank first, and the interaction of repairs and operation capital has been rank thirteen among the significantly identified attributable factors. We have a detailed and extensive analysis of the rankings in a later discussion.

Table 6.5: Rank of Contribution of Attributing Factors to returns from corn production in U.S. 1975-2018

Rank	**variable**	**Description**	$p-value$	R^2	**%Contribution**
1	X_{14}	Opportunity cost of land	$8.52e^{-09***}$	0.2116	21.54
2	X_7	Fuel, lube & electricity	$8.77e^{-05***}$	0.1846	18.79
3	X_6	Custom Services	$2.53e^{-04***}$	0.1796	18.29
4	X_1	Value of the primary product grain	$2.00e^{-16***}$	0.1538	15.66
5	X_4	Fertilizer	$5.76e^{-06***}$	0.141	14.36
6	$X_4 * X_{18}$	Fertilizer & Price	$8.72e^{-05***}$	0.0453	4.61
7	X_{24}	Operating capital	$8.41e^{-05***}$	0.0197	2.01
8	X_{11}	Hired Labor	$8.88e^{-10***}$	0.0166	1.69
9	$X_4 * X_{19}$	Fertilizer & Enterprise size	$8.34e^{-03**}$	0.0088	0.90
10	$X_1 * X_{18}$	Value of the primary product grain & Price	$1.17e^{-09***}$	0.0071	0.72
11	$X_{14} * X_{18}$	Opportunity cost of land & Price	$8.72e^{-05***}$	0.005	0.51
12	$X_4 * X_{22}$	Fertilizer & Variable cost expenses	$1.73e^{-02*}$	0.0048	0.49
13	$X_8 * X_{24}$	Repairs & Operating Capital	$1.19e^{-02*}$	0.0043	0.44
Total				0.9822	100

6.4.1 Validation of the Proposed Model

We validated the proposed model given by Equation (6.13) by first satisfying all the key assumptions of the model. First, we tested for linearity between the response variable and the continuous attributable variables using the partial residual plot given by Figure 6.7. From Figure 6.7, we can see that there is a well-defined linear relation between the response variable and the individual continuous attributable factors contributing to corn returns. Also, the problem of an insignificant intercept term of the initial model has been resolved. We now have a significant intercept term of the transformed model proposed given by $p - value = 3.432e^{-03}$ (i.es rejecting $H_0 : \tau = 0$), attesting to the linearity assumption of the model.

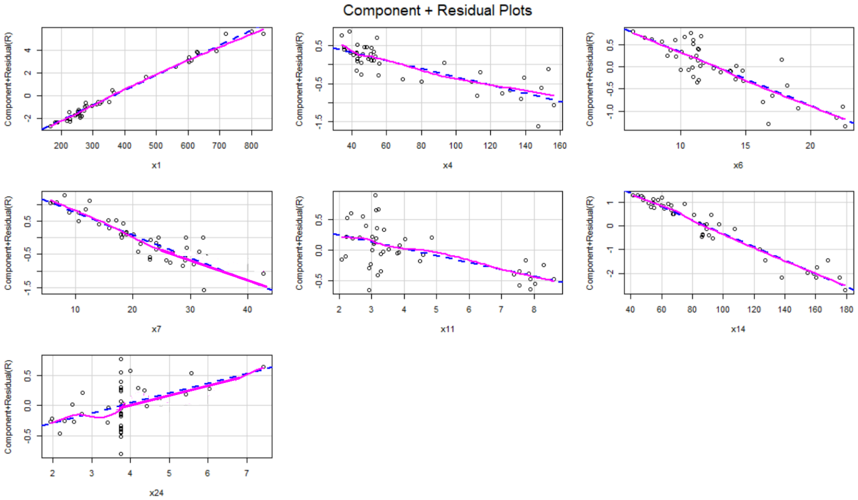

Figure 6.7: Assessing Linearity of the Proposed Model

Secondly, we investigated the presence of Gaussian normal probability distribution of the proposed model, given by normal plots in Figure 6.8. The first panel is the normal Q-Q plot of residuals with 95% confidence bounds and the second penal is the distribution of the studentized residuals. We can see from both panels that the assumption of the normal probability of the proposed model is well-preserved since all the residual point falls within the 95% bound of the Q-Q plot with no major outlier. We performed a formal test for normality using the Shapiro-Wilk's test, which resulted in a large $p - value = 0.9759$, indicating the proposed model residuals are Gaussian distributed. Thus, further affirming to the evidence of normality given by Figure 6.8.

Another key assumption that our proposed model satisfied is homoscedasticity (i.e. the residual errors should have constant variance). We plotted the residuals against fitted values and look for a pattern or trend given by Figure 6.9. If there is no pattern or trend, points in the plot are randomly scattered about the zero lines, and no major outlier is an indication of the presence of homoscedasticity. Figure 6.9 reveals that there is evidence of the homoscedasticity of residuals of the proposed model. In a further residual analysis, we found the residuals to have a mean of zero (i.e. $\bar{\epsilon} = \sum_{i=1}^{n} e_i \approx 0$) and a standard deviation ($s = 1/(n-1)\sum_{i=1}^{n}(\epsilon_i - \bar{\epsilon})^2$) of 0.7317. Also, the Durbin-Watson test was performed to

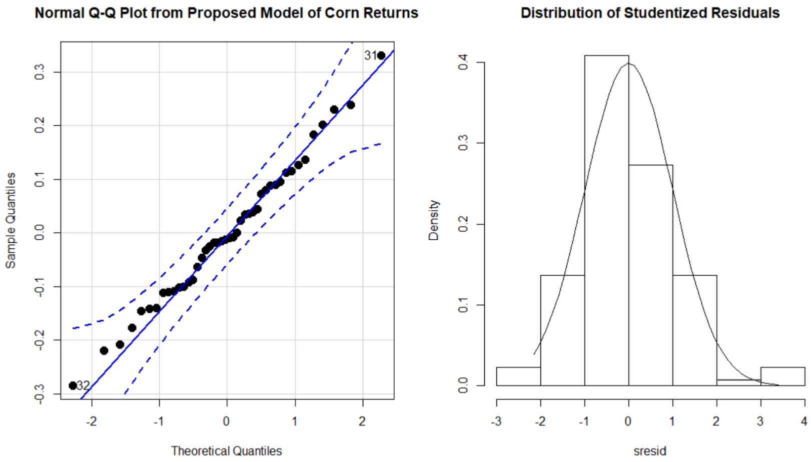

Figure 6.8: Assessing Normality of the Proposed Model

investigate the presence of autocorrelation among residuals. The test resulted in a large $p - value = 0.238$, indicating that the residuals are uncorrelated.

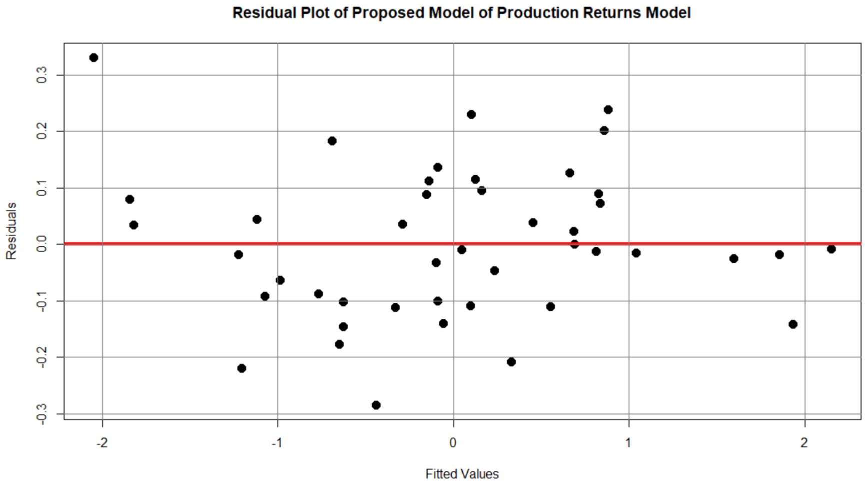

Figure 6.9: Assessing Homoscedasticity of the Proposed Model

As we stated earlier during the model building, multicollinearity was a problem we encountered. Although some have argued that multicollinearity does not affect the precise prediction of the model. However, we expect a statistical model with very small or no multicollinearity to perform better than models with high multicollinearity. Multicollinearity can cause the mean square error (MSE) to increase drastically and cause some predictors to be statistically insignificant, when in fact they are important in predicting the response. One commonly used technique for handling multicollinearity is by removing the redundant predictor(s) that are highly correlated with the other predictor(s). Very high multicollinearity among predictor variables can lead to overfitting, hence may result in a misleading decision. Our proposed model addressed the problem of multicollinearity after the transformation of the response variable, and careful selection of the attributable variables based on the stepwise backward elimination model selection procedure, thereby reducing its impact on the precision and accuracy of predictions.

Given that the proposed multivariate nonlinear regression model of corn returns is of high quality, given by $R^2 = 0.9822$, and validates all the key assumption, we further measure or validate the quality of the model base on the root mean square error, given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)}.$$

The RMSE measures the difference between the predicted value and the observed values. The smaller the RMSE the closer the predicted values are to the observed values, and the more accurate the model prediction. The proposed model has an RMSE of 0.1577, indicating a very good predictive accuracy.

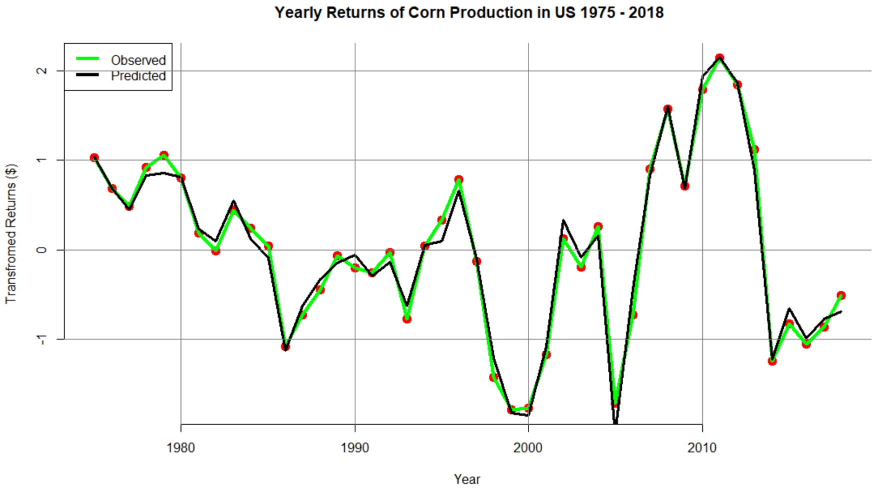

Figure 6.10: Assessing the Accuracy of Prediction by the Proposed Model

In Figure 6.10, we assessed how the accuracy of prediction by the proposed model. Our proposed model has great accuracy. We can see that the predicted values are very close to the observed value, which explains the high efficiency, precision, and robustness of the proposed model. We further performed a Kruskal-Wallis test to assess the difference between the observed returns and the predicted returns given by Table 6.6. The test shows that there is no difference between the two returns, given by the large p-value. Also, there is a very strong correlation between the observed values and the predicted values of $\mathsf{0.9911}$.

Table 6.6: Kruskal-Wallis rank sum test of the Difference in Observed Returns and Predicted Return of Corn Production.

Type of Test	**Corn Returns**	**Data: list(Observed, Predicted)**
Kruskal-Wallis	$chi - squared(\tilde{\chi}^2) = 6.9644e^{-05}$	$p - value = 0.9933$

Furthermore, the model was developed using 80% trained data and then assessed the prediction accuracy using the remaining 20% of the data. Table 6.7 shows the prediction of the returns from the 20% test data after we developed the model using the entire data and then 80% train data. A correlation coefficient between the test set and the predicted values revealed $\mathsf{0.958}$ and $\mathsf{0.976}$ using the trained model and the entire data model, respectively. Interestingly, the relation between the two model prediction revealed a correlation coefficient of $\mathsf{0.995} \approx \mathsf{1}$. This further attests to the high quality, efficiency, predictive accuracy, and precision of the proposed nonlinear analytical model given by Equation (6.13).

6.4.2 Evaluation of the Proposed Model

Another strategy we employed to assess the quality of the proposed model is by comparing it with other least square models. The coefficients of the proposed model for the returns from corn production are estimated using the ordinary least square (OLS) parameter estimation method. One major assumption of OLS is homoscedasticity and the absence of serial correlation. If the assumption is violated (similar to what we initially encountered before the transformation of the response), a transformed version of the OLS, namely, the gener-

Table 6.7: Comparison of Prediction of the Return from Corn Production Based on Train and Test Method

Returns	Predicted Values	
Observed Values	Entire data Model	Trained Model
1.0255	1.0408	1.0572
0.6871	0.6874	0.6764
-0.0141	0.0949	0.1099
0.3312	0.1004	-0.0080
-1.7665	-1.8461	-1.8539
0.1204	0.3286	0.4018
1.5682	1.5941	1.4535
1.1205	0.8815	0.8658
-0.8282	-0.6516	-0.5541

alized least square (GLS) is often recommended to give the best linear unbiased estimators provided the other assumptions are satisfied. The GLS regression model can be written as

$$\tilde{y} = \tilde{X}\beta + \tilde{\epsilon}, \tag{6.15}$$

where $\tilde{y} = \Sigma^{-1}y$ denotes the response, $\tilde{X} = \Sigma^{-1}X$ is the model matrix of predictors, $\tilde{\epsilon} = \Sigma^{-1}\epsilon$ is the model residuals, and β is the GLS estimator given by $\hat{\beta}_{GLS} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\tilde{y} = (X^TV^{-1}X)^{-1}X^TV^{-1}y$. V is symmetric and positive definite defined as $V = \Sigma\Sigma^T$, where Σ is an invertible variance covariance matrix. In a situation where the errors are uncorrelated, but not necessarily homoscedastic, the weighted least square (WLS) is often used to obtain the best unbiased estimators. When V is diagonal, the errors are uncorrelated but may not have equal variance. We can express $V = dia(1/w_1, ..., 1/w_n)$, where w_i are weights and $\Sigma = \sqrt{1/w_1}, ..., \sqrt{1/w_n}$, and we can regress $\sqrt{w_i}x_i$ on $\sqrt{w_i}y_i$.

One of the initial problems in the course of the model building process was an unequal variance of the errors. We applied the Johson transformation on the response leading to the selection of the final model that satisfies all the assumptions. We compared the quality of our proposed model with the GLS and WLS based on the root mean square error (RMSE) and the Akaike information criteria (AIC) given by Table 6.7. The proposed model performed much better than the other two methods used.

Table 6.8: Proposed Model Comparison with Other Least Square Models.

Rank	Method	RMSE	AIC
1	Proposed model(OLS)	0.1577	-24.520
2	WLS Model	0.171	-18.072
3	GLS Model	0.1759	-10.823

6.5 Discussion

The production of corn plays a key role in the economy of the United States of America. The U.S. is the world's leading producer of corn, with about 80 million acres (32 million ha) of land exclusively dedicated to corn production. The U.S. agricultural sector is predominantly corn production, playing an essential role in the ethanol production industry, distillery industry, livestock industry, beverage alcohol industry, among others. Approximately 13% of the U.S. annual corn yield is exported to more than 73 different countries across the globe, a report by U.S. grains council. It is therefore imperative to investigate the returns from corn production in the United States. The amount the corn production industry earns after all revenues and costs is an essential motivation for how they plan their production each year. Therefore, the industry needs to know the key elements or factors contributing to their returns at the end of each production circle. Knowing these key contributors to the returns of the corn production will aid the industry to plan rationally and judiciously to their favor, thereby increasing the product returns. It would further serve as a boost in the U.S. economy of corn production and stiffen its competitiveness in the world's economy of corn production.

In the present study, we developed a data-driven nonlinear statistical regression model to predict the returns of the production of corn in the U.S. The initial model building process takes into account 25 elements or factors presumed by the U.S. department of agriculture

(USDA) to be contributing to the returns from corn production in the US. The following questions were asked during the model building process. Are all the 25 factors significant? Are there any significant interacting factors? How much percent is the significant factors contributing? What is the percentage of contribution by each significant factor? How much percent of contribution to the returns is by unknown or confounding factors? These are highly intriguing and essential questions our developed model addressed.

The model building process started by considering all the 25 factors published by the USDA as contributing to the returns from corn production in the US. However, after rigorous and careful investigating analysis we found 7 out of the 25 factors to be statistically significantly contributing individually to the returns from corn production, as well as 6 interaction terms. We utilized the best form of transformation on the skewed returns (i.e. the Johnson transformation) and the best model selection technique (stepwise backward elimination) to identify the significant contributable factors to the corn returns. The final proposed model that precisely and accurately predict the returns from corn production in the U.S. is given by Equation (6.13) in a transformed form. To use the model, we replace the identified predictor variables (i.e. the seven individual attributable factors and the six interaction terms) with real values to predict the transformed returns. We then utilize Equation (6.14) to transform back to the original values of the corn returns.

Now, how do we justify the goodness-of-fit of the model? Firstly, the proposed model satisfies all the key assumptions of a linear statistical regression model. It addresses the problem of heteroscedasticity and serial correlation initially encountered. The model has a coefficient of determination, R^2 of 98.22%. Thus, 98.22% of the variation in the returns from corn production is explained by the identified seven individual attributable factors and the six interaction terms. In other words, 98.22% of corn production returns are contributed by the identified attributable variables of the proposed model, and the remaining 1.78% is contributed by other unknown or confounding factors. Although, multicollinearity may not be considered problematic in a predictive model like ours because it does not affect the precision of prediction of the model [87, 88]. However, multicollinearity can cause some predictors to be insignificant when in fact they are important, causing model overfitting or underfitting. The proposed model lessens the impact of multicollinearity.

We ranked the identified attributable factor and the interactions according to the individual percentage of contribution to the returns from corn production in the US, given by Table 6.5. In other words, we ranked from the most important contributor to the least contributor of the identified factors to the returns from corn production. It is extremely important to consider the rankings to enable farmers or industries into corn production to allocate resources effectively and efficiently towards maximizing the returns. The opportunity cost of land was ranked first (21.54%), followed by fuel, lube and electricity (18.78%), custom services (18.29%), the market value of the grain (15.66%), and fertilizer (14.36%) was ranked fifth. The opportunity cost of land is the benefit forgone for trading off the land for cultivation of corn in the U.S. over other economic purposes or use. Given that all other factors remain constant, the lesser the opportunity cost of land for corn production, the higher the likelihood of the returns. Thus, the expected returns from the production of corn can increase as long as it cost less to cultivate more acres of land for corn production. Also, the investment in corn production would increase if it promises more profit or economic benefits than other alternative economic investments. This is extremely important and useful information we can extract from our model, attesting to the quality of our model. On the contrary, all things being equal, if all the other factors like operating capital, labor, fertilizer, etc, are readily adequately available except the availability of adequate land for the cultivation and expansion of corn production, the amount of returns earned is likely to fall. Given that the U.S. is the world's leading producer and exporter of corn, not trading off the land cultivated for corn production for other economic purposes is a major strategy for increasing the returns. Rather expanding the acreage of cultivated land is essential to keep the U.S. as a continuous world's leader of corn production, thereby increasing the market size of corn, and hence the returns.

Interestingly, the top 5 ranking factors contribute to 88.64% of the returns from corn production in the US. We expected more contribution from operating capital and labor hired ranked seventh and eighth, and contributing 2.01% and 1.69%, respectively, to the returns from corn production. Though operation capital was identified to be highly statistically significantly contributing to the returns from corn production, we expected more in terms of its percentage of contribution given the fact that the growth and expansion of the profitability level of most economic industries or businesses depend largely on the operating capital. The rank of fuel, lube & electricity as second contributing to 18.79% of the return is not surprising due to the technological advancement in the production of corn. Cardwell (1982) [89] pointed out the strong evidence of technological progress effect on the economics of corn production. Another intriguing finding by our model is, fertilizer interacts with three different other contributing factors (price, enterprise size, and variable cost expense) not contributing as an individual factor to the returns. Also, price (dollar per bushel harvest) which was not found as a significant contributable individual term interacted significantly with three factors (fertilizer, the value of the grain and opportunity cost of land) that were found as individually contributing to the returns. Similarly, repairs were not individually contributing to the returns, but significantly contribute as it interacted with the operating capital.

Most research in statistical modeling turns to ignore the inclusion of interactions between attributable variables because they are either difficult to find or interpret. However, not including interaction in a model when they significantly contribute to the response variable can distort the robustness and efficiency of the model, thereby weakening the effectiveness, predictive accuracy, and useful information that can be extracted from the model. To have a significant interaction term implies that both attributable variables together have a significant influence on the response variable (the returns), though one or both may or may not be individually significant.

The value of the coefficient of the attributable variables can be interpreted as the change in the response variable (i.e. returns from corn production) brought about by a unit change in the value of the attributable variable. All this being equal, for a positive coefficient, we can maximize the returns/profit by increasing the value of the attributable variable. Whereas, for a negative coefficient, we can maximize the returns by decreasing the value or impact of that attributable variable. For instance, the value of the coefficient for the opportunity cost of land is -0.05173, given by Equation (6.13). This means by holding all the other factors constant, a unit decrease in the opportunity cost of land would increase the returns from corn production by 0.05173, and vise versa. The coefficient of the interaction between fertilizer & price (dollar per bushel harvest) is 0.01505, meaning a unit change in either the fertilizer or price would result in 0.05105 change in the returns.

To further evaluate the quality of the proposed model, we used the model to predict the returns from corn production in the U.S. from 1975-2018 and compared with the observed or original values of the returns, given by Figure 6.10. The observed returns are given in green color and the predicted returns are given in black. We can see that the proposed model did great to closely predicting the exact values of the observed returns given by the data. We further computed the correlation coefficient between the two returns to assess the strength of the relationship, resulting in a very strong correlation. We also tested whether there was a difference between the two returns based on the Kruskal-Wallis test, resulting in a very large p-value (no difference), which goes to affirm the result given by Figure 6.10 and the correlation coefficient. The proposed model was compared with other least squares models (i.e. the generalized least squares and the weighted least squares), given by Table 6.8. The criteria of comparison of the three models were based on the root mean square error (RMSE) and the Akaike information criterion (AIC). A model with least RMSE (captures the remaining amount of unexplained variation in the returns) and least AIC (measures the amount of information not captured by the model) is considered the better. Our proposed model has the least RMSE and AIC, hence the best choice of implementing our proposed model. The finding of the proposed model would serve as a strategy or guide for increasing the returns earned by industries or farmers into corn production in the U.S. and the world at large.

6.6 Contribution

Corn production plays a major role in the agricultural economics of the united states. Finding strategies to maximizing the returns is a key way to influence more investors into the production of corn, and sustaining the U.S. position as the world's leading producer of corn. The best way for corn farmers or industries to increase their returns is knowing the major factors that contribute to the returns to enable them to proportionate resources more effectively and efficiently. In the present study, we developed a data-driven multivariate nonlinear statistical models that identified seven significant individual contributable (risk) factors and six significant contributable interaction terms that accurately predict the returns from corn production in the U.S. from 1975-2018. The identified factors includes **the opportunity cost of land**, **fuel, lube and electricity**, **custom services**, **the market value of the grain**, **fertilizer**, **operating capital**, **hired labor**, and the interaction factors including **fertilizer & price**, **fertilizer & enterprise size**, **market value of grain & price**, **opportunity cost of land & price**, **fertilizer & variable cost expense**, and **repairs & operating capital**. The quality of the proposed model was evaluated by satisfying the model assumptions, and base on very high coefficient of determination (R^2 along with R^2_{adj}) statistic, the least root mean square error (RMSE) statistic, the least Akaike information criterion (AIC) of model selection, and the minimum variance inflation factor (VIF).

Our study offers five major usefulness to the economics of corn production. Firstly, given the set of real values of the significant identified contributable factors, we can precisely estimate/predict the returns from corn production with a 98.22% degree of accuracy. Secondly, we identify individual and interaction factors significantly contributing to the returns from corn production. Thirdly, we obtained the ranks of the identified contributable factors to the corn returns from the highest to the least percentage contributor, with the opportunity cost of land appearing as the top contributor to the returns from corn production. Fourthly, we can perform surface response analysis or optimization analysis to identify the value of the attributable factors that are necessary to maximize the returns from corn production. Thus, for a given contributable factor we can analyze ways to maximize the returns either by increasing or decreasing the impact of the contributable factor, holding the other factors constants. Fifthly, we can create confidence bound with a desirable level or degree of confidence to monitor the returns from corn production. For example, for a 95% confidence interval, if the returns fall below the confidence bound could create investor panic, hence there would be the need for some instantaneous and critical adjustment in the production process through rigorous and careful analysis of the identified contributable factors in the model. On the contrary, if the returns fall within or above the confidence bound could further boost investors' motivation and trust in the economics of corn production.

Finally, the proposed model is cost-effective for the subject area. For corn production firms to maximize their returns/profit, they do not have to spend a huge amount of resources on variables or factors that do not contribute to the returns. Hence, there is no doubt about the tremendous importance the current study brings to improving the economics of corn production in the United States. It would help corn farmers or industries to plan rationally and judiciously towards allocating resources effectively and efficiently to maximize their returns. The proposed statistical model can be applied to monitor and evaluate the returns/profit in other fields of production. Our Statistical model building process for the returns from corn production can be followed to develop a similar model for other production sectors or economies.

In appendix A, we discussed how the production firm or industry can utilize our proposed analytical model to maximize the returns of production of corn.

Chapter 7: Desirability Function Approach for Optimizing the Returns of Corn Production in the U.S.

Corn production plays a vital role in the agricultural economic growth and development of the U.S and the world at large. To ensure that corn farmers, firms, industries, etc, attain maximum returns/profit from production, this chapter presents a surface response optimization analysis for maximizing the returns/profit from corn production in the U.S utilizing the desirability function approach. The study uses 44 years of data from 1975-2018, obtained from the U.S Department of Agriculture (USDA) Economic Research Service. We first identified a statistical model with ten different significant risk factors out of 25 published by the USDA (see Chapter 6), resulting in R^2 of 98% and predict the returns with high accuracy. We obtained the optimal/maximum value of the returns with a desirability function of $0.99 \approx 1$, implying that the optimum values of the risk factors are robust and effective in maximizing the returns. A 95% confidence region we obtained further suffices that the optimal point of the response is significant. The optimization process employed in maximizing the returns from corn production has been well-validated and provides more flexibility and efficiency for production profit maximization.

We organized this chapter as follows: Section 7.1 introduces and review some literature of studies on subject area; Section 7.2 presents the optimization method of the returns from corn production; Section 7.3 presents discusses the findings in this study; and finally, the research contributions of this chapter is captured in Section 7.4.

7.1 Introduction

The United States continues to play a key leading role in the world's production and exportation of corn. Corn, also known as "maize" serves several purposes to humans and animals including food and industrial products like cereal, alcohol, sweeteners, byproduct feeds, livestock feed, manufacture of ethanol, and its co-product, biomass fuel, among others [68]. For an extensive review of corn production in the U.S see [68, 105, 69, 70]. Corn is one of the major crops that dominate the agricultural production economies not only because of the widespread of subsistence and commercial usefulness, but also its profitability or returns. Corn turns to enjoy a considerable market demand throughout the year, hence considerable revenue from production. The returns from production is often said to be influenced by the forces of demand and supply, which are often uncontrollable factors, in particular, in a perfectly competitive market. The controllable factors that impact the returns from production are factors that contribute to the cost and revenue of production. In our previous study on corn production [105], we investigated the factors that influence the returns or profitability of corn production utilizing a data of 25 risk factors of cost and revenue obtained from the U.S Department of Agriculture (USDA). We found thirteen risk factors, which include seven individual and six interaction terms that predict the returns with 98% accuracy. The identified factors were ranked according to the contributing effect on the returns. After we have found the factors that contribute to the returns, one intriguing question will be on how to optimize or maximize the returns. Thus, how do we obtain the optimal or maximum returns based on the thirteen identified risk factors? What are the target or optimal values of each identified risk factor to achieve the maximum returns/profit from the production of corn?

Greg Ibendahl discussed the profit-maximizing of the economics of corn production using quantities and prices of all inputs and outputs [91]. Jeff Coulter [92] reviewed research by the University of Wisconsin on corn silage planting date trials. They found that planting between April 21 and May 6 produced maximum grain yields, hence maximum revenue. Hubner Seed

[93] conducted a study to evaluate the economic impact of different management inputs and practices on corn yield and profitability. They found early relative maturity products planted at the standard rate with QuickRoots Dry Planter Box Corn and additional nitrogen (33KQN) to be the most profitable treatment for corn production. Another concept often used by economists to determine the profitability of a firm in production is the concept of cost and revenue. This concept outlines that the profit-maximizing stage of a production firm is where the marginal cost of production equals the marginal revenue. This method has several limitations, some of which we pointed out in Chapter 6. These findings or concepts of obtaining the optimal returns from production are less efficient and can result in making the wrong decision about the firm profitability.

In the present chapter, we proposed a more robust and efficient means to obtain the optimal value of the returns of US corn production utilizing the model developed in Chapter 6. We performed an optimization analysis to maximize the returns from corn production utilizing the desirability function method of response surface methodology (RSM). The RSM is typically a combination of design of experiment, regression models, and optimization. It accesses the combination and relationship between risk factors and the response variable(s) for optimization purposes. The method was proposed by George E. P. Box and K. B. Wilson in 1951. The objective of RSM is to optimized the response variable(s) by maximizing, minimizing, or obtain a target value. It finds the solutions/values of the controllable risk factors which result in the best value of the response or responses simultaneously. The optimization in RSM can be categorized into three types, namely, single response surface optimization (SRSO), dual response surface optimization (DRSO), and multiple response surface optimization (MRSO). The SRSO optimize the mean response based on the combination of a set of values of the risk factors or independent variables subject to constraints. The DRSO optimizes one response subject to constraints of the other response. Thus, the optimum response is based on the constraints and trade-offs between the two competing responses. The MRSO optimizes one mean response subject to the constraints or trade-offs among one or more responses. In RSM, there are two major approaches/techniques for response optimization; namely, *constraint optimization problem approach (COPA)* and *desirability function approach (DFA)*. Li J et al [98] used the desirability function approach as an optimization strategy for microbial glutamine production. Ryad Amdoun et al [99] also used the desirability optimization method to predict two antagonist responses in biotechnological systems. This study involves the SRSO type and uses DFA to maximize the returns of corn production in the United States. We validated the efficiency of the optimization approach based on the value of the desirability function, R^2, R^2_{adj}, and R^2_{pred} statistic, and the 95% confidence interval (CI) and prediction interval (PI) of the optimal point of the returns. Please, see Section 6.2 of Chapter 6 for detail description of the data used in this chapter.

7.2 Proposed Method for Optimization of Returns from Corn Production

7.2.1 Overview of Desirability Function Approach

A response surface optimization problem with single or multiple responses $Y = \hat{y}_j$ and input variables or risk factors $X = x$ can be expressed as follows:

$$\text{Optimize } [\hat{y}_1(x), \hat{y}_2(x), ..., \hat{y}_r(x)],$$
$$\text{subject to } x \in \Omega$$

where $\hat{y}_j$ for $j = 1, 2, ..., r$ is the estimated j^{th} response variable and r is the number of responses. For $j = 1$ means a single response problem, $j = 2$ implies dual response problem, and $j = 2$ or more implies a multiple response problem. Ω is the experimental space of x.

The desirability function approach (DFA) is a response optimization strategy for optimizing the response as a function of controllable input factors. It transforms the fitted response into free-scale value, called desirability. More extensive literature and review can be obtained from [94, 95, 96]. DFA accesses the combination of the controllable input variables or risk factors that optimize the defined goal of the response variable(s). Different objective of the response requires different desirability function be employ. The optimization objectives can be to maximize, minimize, or obtain a target value of the response variable [97]. If the objective is to maximize the response (i.e. the larger the better), the desirability function is defined as

$$d_j(\hat{y}_j) = \begin{cases} 0, & \hat{y}_j < L_j \\ \left(\frac{\hat{y}_j - L_j}{T_j - L_j}\right)^{\eta}, & L_j \leq \hat{y}_j \leq T_j \\ 1, & \hat{y}_j > T_j. \end{cases} \tag{7.1}$$

If the objective is to minimize the response (i.e. the smaller the better), the desirability function is defined as

$$d_j(\hat{y}_j) = \begin{cases} 1, & \hat{y}_j < T_j \\ \left(\frac{U_j - \hat{y}_j}{U_j - T_j}\right)^{\eta}, & T_j \leq \hat{y}_j \leq U_j \\ 0, & \hat{y}_j > U_j. \end{cases} \tag{7.2}$$

And, if the objective is to obtain a target value of the response, the two-sided desirability function is defined as

$$d_j(\hat{y}_j) = \begin{cases} 0, & \hat{y}_j < L_j \\ \left(\frac{\hat{y}_j - L_j}{T_j - L_j}\right)^{\eta_1}, & L_j \leq \hat{y}_j \leq T_j \\ \left(\frac{U_j - \hat{y}_j}{U_j - T_j}\right)^{\eta_2}, & T_j \leq \hat{y}_j \leq U_j \\ 0, & \hat{y}_j > U_j. \end{cases} \tag{7.3}$$

Where η_1, η_2, and η represent the weighted parameters that define the shape of the individual desirability function $d_j(\hat{y}_j)$. For $\eta_1 = \eta_2 = 1$ implies the shape of $d_j(\hat{y}_j)$ is linearly increasing, $\eta_1 < 1$ and $\eta_2 < 1$ implies the shape of $d_j(\hat{y}_j)$ is concave, and $\eta_1 > 1$ and $\eta_2 > 1$ implies the shape of $d_j(\hat{y}_j)$ is convex. T_j is the target value of $\hat{y}_j$. L_j and U_j denotes the lower bound and the upper bound of the response $\hat{y}_j$, respectively. To obtain the overall desirability function, we use the formula

$$D = \left[\prod_{j=1}^{r} d_j(\hat{y}_j)\right]^{1/r} = [d_1(\hat{y}_1)d_2(\hat{y}_2)...d_r(\hat{y}_r)]^{1/r}. \tag{7.4}$$

D ranges from zero to one (i.e. $0 \leq D \leq 1$). $D = 1$ is the ideal case of optimality and indicates the optimum point of the response. Thus, the input variables or the risk factors are very effective in optimizing the response. $D = 0$ implies that the response variable(s) is/are outside the acceptable region, which means the controllable input factors are doing poorly in optimizing the response (the case of undesirable response).

7.2.2 Statistical Analysis for Optimization of the Production Returns

The optimization problem for this study involves a single response, and the objective is to maximize the returns from corn production. Therefore, we utilized the desirability function defined by Equation (7.1) to optimize the response variable. The optimization process we adopted involves the following steps:

1. Building the regression model that significantly predicts the continuous response/target variable(s) with high accuracy.

2. Defining the constraints or limits of the response(s) and that of the input/risk factors.

3. Identifying the desirability function appropriate to optimize the response base on the response optimization objective.

4. Executing the function and obtain the optimum value of the response(s), values of the input variables, and the value of the desirability function.

5. Validating the Optimization process.

The results of the outlined optimization process for the returns from corn production is as follows:

Table 7.1: The Constraints of the Response and Risk Factors

Returns	Input/Risk Factors	
$-134.03 \leq y \leq 224.31$	$167.03 \leq X_1 \leq 836.58$	$41.19 \leq X_{14} \leq 179.15$
	$34.09 \leq X_4 \leq 156.51$	$1.00 \leq X_{18} \leq 6.79$
	$6.28 \leq X_6 \leq 22.69$	$189 \leq X_{19} \leq 280$
	$5.76 \leq X_7 \leq 42.64$	$0.05 \leq X_{22} \leq 0.59$
	$2.08 \leq X_{11} \leq 8.61$	$1.94 \leq X_{24} \leq 7.42$

Step 1

The statistical model that accurately predicts the returns from corn production in the U.S. with 98% accuracy is given by

$$\begin{aligned} \hat{y} &= 121.6 + 1.263X_1 - 5.74X_4 - 4.541X_6 - 2.408X_7 - 7.46X_{11} \\ &\quad - 2.304X_{14} - 2.13X_{18} - 0.193X_{19} - 19.00X_{22} + 6.24X_{24} - 0.096X_1 * X_{18} \\ &\quad + 0.613X_4 * X_{18} + 0.010X_4 * X_{19} + 0.91X_4 * X_{22} - 0.086X_{14} * X_{18} \end{aligned} \tag{7.5}$$

Note that the risk factors that were not significant individually, but significant as it interacts with others were included in the model for the response optimization because they cannot be separated. Hence, we have ten individual risk factors, including the factors that made-up interaction terms.

Step 2

The constraints or limits of the returns from corn production and the ten input variables are given in Table 7.1 below.

Step 3

The optimization objective is to maximize the returns from corn production. Thus, the larger the returns the better. Therefore, we utilized the desirable function defined in Equation (7.1).

Step 4

After the optimization process has been executed, the optimum estimated response value and the corresponding values of the input variables, and the value of the desirability function is given in Table 7.2.

Table 7.2: Optimal Values and the Desirability Function

$\hat{y}$	X_1	X_4	X_6	X_7	X_{11}	X_{14}	X_{18}	X_{19}	X_{22}	X_{24}	$d(\hat{y})$
850.773	836.58	34.09	6.28	41.19	2.08	5.76	1.00	280	0.05	7.42	0.99

Step 5

We validated and assessed the optimization process-ability to achieve the maximum returns from corn production based on the value of the desirability function $d(\hat{y})$, the coefficient of determination of the optimum model R^2 (the variation in the response explained by the input factors), and the model prediction accuracy R^2_{pred} given by Table 7.3. The 95% CI and PI was used to structure the following hypothesis test about the significance of the optimal value of the response.

H_0 : the optimal point is significant.

H_a : the optimal point is not significant.

Also, display in Figures A.1 and A.2 in the Appendix are contour plots and their corresponding 3D plots of the combination of two risk factors in maximizing the response holding other factors constant.

Table 7.3: Optimization Process Validation

R-sq	R-sq(adj)	R-sq(pred)	$d(\hat{y})$	95% CI	95% PI
98.72%	98.04%	94.80%	0.99	(722.5, 979.1)	(720.9, 980.7)

7.3 Discussion

The economic viability of corn in the U.S and the world at large cannot be underestimated, as it plays a pivotal role in agricultural economics and the economic development of most countries. Corn production in the U.S contributes to 95 percent of total feed grain production and usage [100]. There has been an increased revenue for corn farming in the

U.S from 2009-2014, with the U.S recording a revenue of 63 billion dollars in 2014 [101]. However, the profit or returns from corn production in the U.S have seen a downslide, as the returns per planted acre continue to decrease since 2011, shown in Figure 6.2 of Chapter 6. The returns represent the amount of dollars per acre that the individual farmer or firm saves after all revenues and costs. The higher the returns earned, the greater the incentive to continue in the economics of production of corn. It is important to note that investors plan their investment based on the returns or profits they can earn from a particular portfolio. The trade-offs are that the investor will invest more in areas that will earn him/her more profit, and vice versa. The main objective of every individual, firm, industry, organization, or cooperation into production is to maximize profit or returns, which also means that they must strive to minimize costs and maximize revenues. The intriguing question here is, how can the players or investors of corn production maximize returns based on the combination of the various risk factors of cost and revenue?

In the present study, we employed a response surface optimization analysis to maximize the returns from corn production utilizing the desirability function approach or method. A data constituting 25 risk factors of the returns from corn production from 1975-2018 was obtained from the USDA Economic Research Service and used to develop a statistical model that gave an R^2 of 98% [105]. The model consisted of seven individual risk factors and six interaction terms. To optimize the response (the returns) of the model, we refitted the model including all the individual risk factors and that of the interaction terms. Hence, we used ten individual risk factors and five interaction terms to initiate the optimization process. We then obtained the values of the constraints for the returns and each of the individual input variables, given in Table 7.2. After the execution of the optimization process, we obtain the optimal or maximum value of the returns from corn production to be 850.77 dollars per acre cultivated, along with a desirability function value of 0.99. Notice that the desirability function measures the effectiveness of the controllable input variables or risk factors in maximizing the returns earned. Note also that the desirability function is between

0 and 1. Also, the values of the input factors that achieve the optimal point is given in Table 7.2. We further performed a validation process to assess the quality and the robustness of the optimization process in attaining the optimal (maximum) returns. In Table 7.3, we obtained the validation estimates, which include the R^2 along with its adjusted of 98.72% and 98.04%, respectively. We obtained a prediction accuracy of 94.80% and 95% confidence interval (CI) and prediction interval (PI) of (722.5, 979.1) and (720.9, 980.7), respectively. From the CI and PI, we performed a hypothesis test to investigate whether the optimal value of the returns was significant at a 5% level of significance. Both intervals included the optimal value of 850.77, sufficing that the optimal point is significant and meaningful for decision making. The CI is essential for the production firms to assess whether they are within the optimal region of their returns.

We also displayed contour and 3D plots given by Figures A.1 and A.2 in the Appendix to investigate the combination of two input/risk factors in maximizing the returns, holding other factors constant. The deep green color region of the contour plot denotes the region of the optimality of the returns. Thus, the deeper the green color, the closer the returns approaches the optimal point. Not all combinations of the risk factors are shown. We only focus on the major risk factors that highly contribute to the returns. For instance, the contour plot of the enterprise size (X_{19}) and the application of fertilizer (X_4) in Figure A.1 shows that we can maximize the returns from corn production by increasing the enterprise size more than the application of fertilizer. Thus, firms must continue to invest more to expand the acreage of land for the production of corn to continue maximizing the returns they earn than the investment in fertilizer. Also, the contour plot of the value of the primary product grain (X_1) and the opportunity cost of land (X_{14}) in Figure A.2 suggest that the production firm is deemed to maximize its returns if the value of the primary product grain increase and the opportunity cost of land for corn production decreases. This is true because most investors will invest more in corn production as long as the market value for it continues to increase and the opportunity cost of investment lower than other areas of investment. Furthermore, the coefficients or parameters of the risk factors provide additional information about the impact each risk factor or interaction has on the production returns. The factors X_1, X_{24}, $X_4 * X_{18}$, $X_4 * X_{19}$, and $X_4 * X_4$ are positively associated with the returns, and the factors X_4, X_6, X_7, X_{11}, X_{14}, X_{18}, X_{19}, X_{22}, $X_1 * X_{19}$, and $X_{14} * X_{18}$ are negatively associated with the returns from the production of corn in the U.S. The present study provides information necessary for the profit-maximizing decision of a production firm.

7.4 Contribution

We proposed an optimization strategy to maximize the returns/profit from the production of corn in the U.S based on the desirability function approach of response surface optimization methodology. We obtained the optimal/maximum returns to be 850.77 dollars per acre of land cultivated with a 95% confidence region of (722.5, 979.1) using a statistical model with R^2 of 98% and high prediction accuracy of 94%. We obtained a desirability function value of $0.99 \approx 1$, which indicates that the optimal value is robust and efficient. This also suffices that the identified optimum values of the input variables or risk factors are effective in obtaining the maximum returns. The optimization process we performed searched and obtained an arbitrary optimum value of the response. However, the production firm can set its optimal value or target they wish to attain, and the optimization process will search and produce the values of the risk factors needed to obtain the target value of the response at a given desirable function and 95% region of confidence.

It is important to note that the maximum value of the returns we obtained uses the identified risk factors of the statistical model developed from 1975-2018 U.S Corn data. Also, other confounding factors could affect the attainment of the optimal value like weather, economic policies, natural disasters, competitive products, etc, hence the optimal value may not be exact when the approach is applied in a different setting or area. Furthermore, a high-quality statistical model is needed to obtain a high desirability function, hence the optimal point of the response.

Finally, this is the first time the desirability function method is used in the optimization of the returns from corn production. This method is much efficient and flexible to apply than other production profit maximization/optimization techniques or concepts. Our study, therefore, provides a robust and more advanced method for obtaining the maximum returns from corn production.

Chapter 8: A Stochastic Model that Monitors the Returns of Crop Production in the United States.

Corn is one of the most viably productive and economically versatile crop in the agricultural economies of the United State (US). How a corn production firm monitors, assesses and evaluates its returns play a major significant role in regulating the production process to ensure a successful operationality, functionality, longevity, and continuity of it. Most research studies on returns of corn production have focused on comparative analysis of factors contributing to the returns, which is insufficient to present a holistic assessment of dynamism in the returns. To be able to effectively evaluate the changes in the returns of corn production based on it's increasing, decreasing, or remaining unchanged is necessary for ensuring the correct and reliable decisions are made about the firm's production process. In the present Chapter 8, we developed a data-driven time-dependent analytical model using a nonhomogeneous Poisson process (NHPP) or the Power Law Process (PLP) to monitor and evaluates the returns of corn production in the US from 1975 to 2018. The proposed approach uses a $\beta - \mathit{factor}$ obtained from the failure intensity function of the NHPP to monitor, assess, and evaluate the returns of the corn production by assessing changes in $\hat{\beta} >, <, = 1$. Generally, the returns of corn production in the US was found to be depreciating, prompting the need for some adjustment to the ongoing production process. We compared our method to the classical approach of comparing marginal revenue (MR) to marginal cost (MC) and market price to average cost (AC) of production. The proposed production returns monitoring approach proved to be more practical, effective, and efficient, providing a much improved approach to evaluating the returns and making a reliable strategic decision about the production process of US corn production.

The organization of this chapter is as follows: Section 8.1 introduces and review some literature of studies on the subject area; Section 8.2 presents the materials and methods analysis in this study; Section 8.3 presents the analytical results; Section 8.4 discusses the findings in this study; and Section 8.5 represents the research contribution of this study.

8.1 Introduction

To meet the growing demand and consumption of corn of the increasing population in the United States (US) and the world at large requires hefty policy changes in how corn production returns is monitored and evaluated [102]. Corn is the most viably productive and economically versatile crop in the US. The usefulness and the importance of corn production in the US and the world as a whole cannot be underestimated. Corn, often referred to as "maize" serves several usefulness to humans and animals including food and industrial products like high fructose corn syrup, sweeteners, cereal, beverage, alcohol, manufacture of ethanol, byproduct feeds, livestock feed, and its co-product, biomass fuel, among others [68]. The agriculture sector and the landscape of America is dominated by corn. It has been a pillar of American agriculture for decades [125]. Corn production is a major economic product in the agricultural economic growth and development of the US. The US remains the global leading producer of corn producing over 333 million tonnes in 2009 with 96 million acres (39 million ha) of land reserved for corn production. There are 80 million acres (32 million ha) of land allocated exclusively to corn cultivation in the United States [68]. As high as 95% of corn farms in the US are owned by families, with more than 30% of them operated by women. Iowa State is the largest producer of corn in the US, producing 2.36 billion bushels (172.0 bu/acre), with corn value of production of US$14.5 billion in 2011 [106]. A more extensive review and history of corn production in the U.S can be found in [68, 69, 70, 72, 105]

The main goal of every firm in production is to maximize returns. The returns, often defined as total revenue (TR) minus total cost (TC) is the end product of production, consisting of the combination of several attributable factors of TR and TC, which determines and influences the decision making process of a production firm at the end of any given production year. The US Department of Agriculture (USDA) has several of the factors influencing corn returns published on their website at https://www.ers.usda.gov/. Other factors such as economic crisis and natural factors like weather (climate change) can result in fluctuation or unstable corn prices, hence an impact on corn returns [77, 78]. The value of a type of corn and its returns may also depend on the location, amount of bushels cultivated and the quality of the corn product [76]. In 2015, the Agricultural Marketing Service of USDA reported the cost of a bushel of corn as $3.50 [110]. A cautious and conscientious analysis of the corn returns is necessary to ensure the smooth running of industries into corn production. Bruce A. Babcock and Gregory R. Pautsch (1998) evaluated the effect on rates and returns of moving from uniform to variable fertilizer rates on Iowa Corn. Their results showed modest increases in the gross returns over fertilizer costs, ranging from $7.43 to $1.52 per acre [111]. In 2017, the Prude University Center for Commercial Agriculture examined the gross revenue for corn in a case farm developed for West Central Indiana tropical corn rotation using the trend-adjusted historical yields; 2017 projected price used for crop insurance adjusted for basis ($3.70 per bushel); and the ratio of historical harvest to projected prices [112]. They further evaluated corn cost and returns for a low, average, and high productivity soil; in which high productivity soil gave higher corn earning or returns than average and low productivity soil [112]. David W Archer et al (1987) investigated the economic risk, returns, and input use under Ridge (RT) and Conventional Tillage (CT) in the northern corn belt of the US. They found economic returns to be significantly higher at the highest nitrogen treatment levels with the highest average net returns to land and management were $78 per hectare for RT at the high nitrogen treatment level (RT-H) followed by $59 per hectare for CT at the high nitrogen treatment level (CT-H) [113]. As a means to improve the returns of corn, Elizabeth Nolan and Paulo Santos (2019) presented evidence that most genetically modified hybrid technologies can improve the yield distribution of corn based on stochastic dominance analysis of corn in the US [114].

Most literature and research studies on returns of corn production have focused on comparative analysis of contributing factors. No research work has been conducted on monitoring and evaluating the returns based on a combined significant factors impact on the return. Given that the return of corn production is influenced by TR and TC of production, the TR and TC incurred by a production firm are in turn influenced by factors such as technology; internal and external factors such as government and international policies; as well as factors of market demand and supply. The returns of production can influence a firm's decision to change the production process by making adjustments or changes to the factors of production. The decision a firm makes based on its returns at a given period can have both positive and negative economic impacts. A positive economic impact could be the employment of additional factors of production and a negative economic impact could mean laying off existing factors of production or even shutting down. Therefore, the production return is a measure of the operational existence of production firms and must be reliably dependable to enable firms to make the correct decisions that are not misleading. Generalizing decisions based on factor-comparative analysis and evaluation is not sufficient to present a holistic assessment and monitoring in the dynamism of corn returns, because it does not take into consideration all the significant production factors influencing the returns. Also, it would be a wrong strategy or approach if a firm based its decision solely on a particular/single production period returns. Most often the return in the present is influenced by past returns, and that of the future is influenced by the present. Carl Zulauf (2015) [115] found the net cash return per acre of corn to have grown on average during each of three different periods: 1975-1995, 1996-2006, and 2007-2014. However, the reliability of such growth was not justified. It is important to assess the rate of change in the reliability growth of the production returns to enable firms to decide on whether the return of production is increasing, decreasing, or remaining unchanged over a period of time. The closest approach to this effect that has been employed by most firms or economists over the years is the profit-maximizing concept of comparing the average cost (AC) of production with the price or marginal cost of production (MC) with the marginal revenue (MR). The application of this concept has worked in some instances, especially for monopolistic firms and in some cases firms in perfectly competitive markets, but in general, has proven to be ineffective and obsolete [116]. The inevitable question is, how can firms make robust decisions about the production returns? In the this chapter, we proposed a time-dependent analytical model or stochastic model that assesses and monitors the end returns of corn production as a function of all the significant factors affecting the returns. The innovative method or approach used in the current study is more effective and provides an enhanced assessment of the returns, resulting in a reliable decision.

8.2 Materials and Methods

8.2.1 Data Preview

The data used in our analysis is in reference to the US Department of Agriculture (USDA) from 1975 to 2018 and available upon request. The data contains twenty-five factors reported to be contributing to the returns of corn production in the US. The Returns (R_p) were calculated by deducting Cash Expenses (TC) from the Gross Value of Production (TR). In Table 8.1, we display the descriptive analysis of the forty-four years returns of US corn production. It showed a negative average return of corn production of 16.83 dollars per planted acre, a median negative return of 26.28 dollars per planted acre, and a standard deviation of 70.8 dollars per planted acre. Comparison of the mean and median returns shows a right-skewed probability distribution, as indicated by the value of kurtosis and skewness. In Figure 8.1, we show the trend analysis of the production returns of corn in the US from 1975 to 2018. There was almost a steady fluctuations in the returns from 1975 to 1996. The return was at its lowest point in 1999 (-134.03 dollars per planted acre).

Thereafter, the returns generally increased and hit the highest peak in 2011 (224.31 dollars per planted acre). It then falls from 2011 to 2014 and thereafter started to rise.

Table 8.1: Descriptive Statistics of Corn Production Returns

Mean	Median	Std Err	Std Dev	Kurtosis	Skewness
-16.83	-26.28	10.68	70.82	2.06	1.17

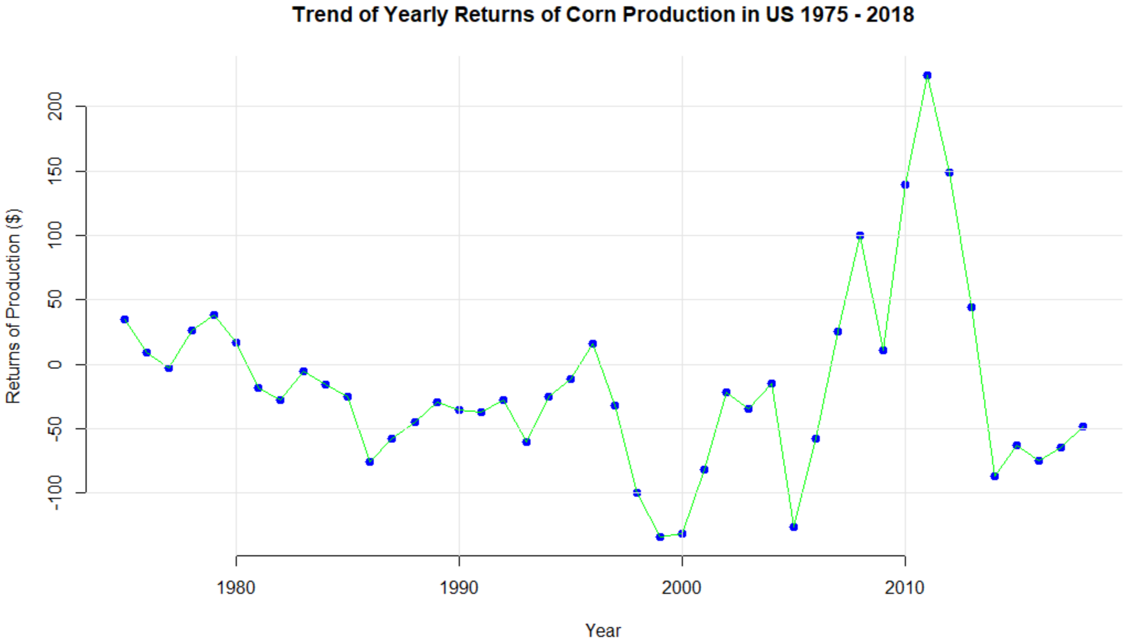

Figure 8.1: Assessing the Accuracy of Prediction of the Corn Production Model

The evaluation of the returns proceeded after a well-developed high quality model that predicts the returns of corn production in the US with 99% accuracy, identifying thirteen statistically significant contributable factors, including seven main or individual contributable factors and six two-way interaction contributing factors to the returns. The model is given by

$$
\begin{aligned}
\hat{R}_p &= 9.424e^{-01} + 2.801e^{-02}1X_1 - 8.737e^{-02}X_4 \\
&- 6.225e^{-02}X_6 - 3.589e^{-02}X_7 - 1.447e^{-01}X_{11} \\
&- 5.173e^{-02}X_{14} + 2.082e^{-01}X_{24} - 4.223e^{-03}X_1 * X_{18} \\
&+ 1.505e^{-02}X_4 * X_{18} + 9.248e^{-05}X_4 * X_{19} - 1.238e^{-02}X_4 * X_{22} \\
&+ 6.140e^{-03}X_{14} * X_{18} - 9.953e^{-03}X_8 * X_{24}.
\end{aligned}
$$

Table 8.2 identifies the thirteen contributable factors with their rank of contribution to the returns in parentheses.

Table 8.2: Description and Rank of the Identified Contributing Factors to the Returns

Main/Single Factors	Interacting Factors
(X_{14}) Opportunity cost of land (1)	(X_4X_{18}) Fertilizer & Price (6)
(X_7) Fuel, lube and electricity (2)	(X_4X_{19}) Fertilizer & Enterprise size (9)
(X_6) Custom Services (3)	(X_1X_{18}) Value of the primary product grain & Price (10)
(X_1) Value of the primary product grain (4)	($X_{14}X_{18}$) Opportunity cost of land & Price (11)
(X_4) Fertilizer (5)	(X_4X_{22}) Fertilizer & Variable cost expenses (12)
(X_{24}) Operating capital (7)	(X_8X_{24}) Repairs & Operating Capital (13)
(X_{11}) Hired Labor (8)	

Also, given in Figure 8.2 is the prediction of the original and predicted returns of corn production in the US from 1975 to 2018. We can see that the proposed model almost predicted the returns perfectly. Therefore, we proceeded with our analysis of evaluation and monitoring of the predicted returns using a nonlinear time-dependent stochastic model known as Nonhomogeneous Poisson Process, NHPP. The goal is to determine when a given return shows increased, decreased, or unchanged over a given period of production and the impact on the production process of corn.

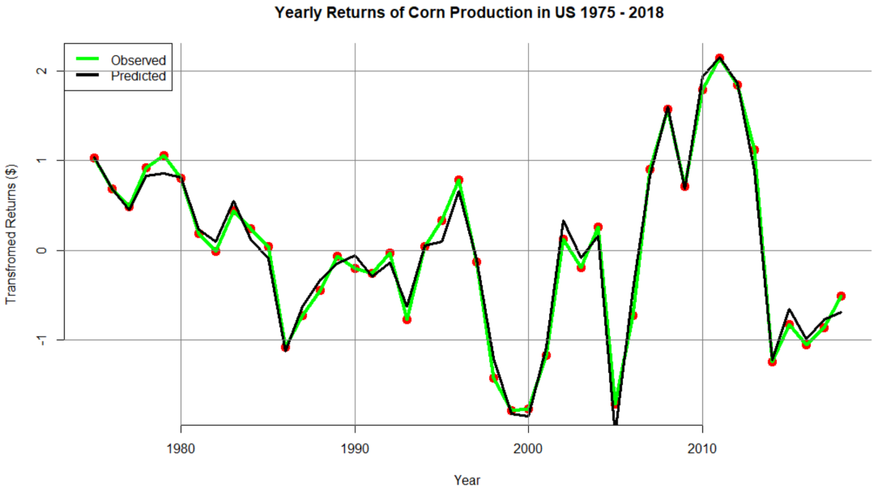

Figure 8.2: Assessing the Accuracy of Prediction by the Proposed Model

8.2.2 A Brief Overview of Nonhomogeneous Poisson Process (NHPP)

The NHPP is also known as the Power Law Process (PLP) or the Weibull process, has been utilized in reliability growth modeling of time-dependent phenomena. [117] [118]; [119]; [120]; among others, have addressed some of the fundamental aspects of reliability growth of repairable systems. However, in corn production, firms would want to assess and evaluate the reliability of the returns of corn production to aid their decision-making process. It is the reliability of contributable factors (tangible and intangible) that drive the production returns. If a factor is less effective or ineffective, the production returns can be greatly negatively affected. For instance, a production firm would like to know the reliability of the production process of the returns based on whether they are making profit or loss, whether to increase, decrease or produce the same quantity of the product, the quality, and the useful life of the production factors such as equipment/machinery, etc. The NHPP contains an intensity function that plays a tremendous role in regulating and measuring the rate of change of the reliability growth process of a phenomenon as a function of time. The intensity function is given by

$$V(t;\beta,\theta) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1}, \tag{8.1}$$

for $\beta > 0$, $\theta > 0$ and $t > 0$. Where β and θ denotes the shape and scale parameter, respectively, and t is the time dependent phenomenon under consideration, proposed by [120][121]. In a NHPP, if there are n failures of a given phenomenon at a time interval $(0, 1]$, then the probability is given by

$$P(x = n; t) = \frac{\exp\{-\int_0^t V(x)dx\}\{\int_0^t V(x)dx\}^n}{n!}, t > 0, \tag{8.2}$$

where $V(t)$ is the intensity function of the process given by (1), which is expressed in a reduced form as

$$P(x = n; t) = \frac{1}{n!}\exp\left\{-\frac{t^{\beta}}{\theta}\right\}\left(\frac{t^{nt}}{\theta}\right). \tag{8.3}$$

equation (8.3) is called the nonhomogeneous Poisson process (NHPP)/Power Law Process (PLP)/Weibull process. Given n failure times $T_1, T_2, ..., T_n$ of a NHPP, where $T_1 < T_2 < ... < T_n$, then the truncated conditional probability distribution function, $f_i(t|t_1, ..., t_{i-1})$, in the Weibull process and is given by

$$f_i(t|t_1, ..., t_{i-1}) = \frac{\beta}{\theta}\left(\frac{t}{\theta}\right)^{\beta-1}\exp\left\{-\left(\frac{t}{\theta}\right)^{\beta} + \left(\frac{t_{i-1}}{\theta}\right)^{\beta}\right\}, t_{i-1} < t. \tag{8.4}$$

The maximum likelihood method of parameter estimation was utilized to estimate the parameters β and θ in equation (8.4). The likelihood function for equation (8.4) when $T_1 = t_1$, $T_2 = t_2, ..., T_n = t_n$ can be expressed as

$$\begin{aligned} \text{Ł} = L(t; \beta, \theta) &= \prod_{i=1}^{n} f_i(t_i|t_1, ..., t_{i-1}) \\ &= \left(\frac{\beta}{\theta}\right)^n \exp\left\{-\left(\frac{t_n}{\theta}\right)^{\beta}\right\}\prod_{i=1}^{n}\left(\frac{t_i}{\theta}\right)^{\beta-1}. \end{aligned} \tag{8.5}$$

The largest failure time or the largest value of the phenomenon often influences the shape parameter, β. To estimate β we equate the partial derivative of Ł with respect to β and letting it equal to zero. By solving for β, we obtained the expression as

$$\begin{aligned} &\frac{\partial Ł}{\partial \beta} = 0; \\ &and \\ &\hat{\beta}_n = \frac{n}{\sum_{i=1}^{n} log\left(\frac{t_n}{t_i}\right)}. \end{aligned} \tag{8.6}$$

The scale parameter, θ is a function of β. Thus, θ is influenced by β. Similarly, we estimate θ by equating the partial derivative of Ł with respect to θ and replacing the estimate of β, the expression of θ is written as

$$\begin{aligned} &\frac{\partial Ł}{\partial \theta} = 0; \\ &and \\ &\hat{\theta}_n = \frac{t_n}{n^{1/\hat{\beta}}}. \end{aligned} \tag{8.7}$$

For an extensive review of the theory of NHPP and its application to reliability, see, Tsokos, C.P, (1995). Now, given the estimates of $\hat{\beta}$ and $\hat{\theta}$, we can estimate the value of the intensity function, $V(t)$ or failure intensity used in modeling the reliability growth of a phenomenon at any given time t. $V(t)$ measures the rate of change in reliability growth as a function of time as a phenomenon depletes or improves. A decrease in $V(t)$ implies that the rate of failure of the phenomenon is reducing. In other words, there is an improvement in the reliability of the phenomenon. This also implies that $\hat{\beta} < 1$. A rise in $V(t)$ implies that there is an appreciation in the failure rate of the phenomenon, therefore a reduction in the reliability of the phenomenon, implying that $\hat{\beta} > 1$. This means that there is rapid depletion in the system or phenomenon, hence, the need for high-level adjustments. When $V(t)$ remains unchanged, it means that $\hat{\beta} = 1$ or approximately 1; thus, the system reliability is the same. Therefore, the behavior of the change in the reliability and dependability model dynamics of a phenomenon is influenced by the shape parameter $\hat{\beta}$ of the intensity function. That is, we can monitor and assess reliability changes behavior of a given time-dependent phenomenon by monitoring the changes in $\hat{\beta}$ of the intensity function. We now showed how the $\hat{\beta}$ of the intensity function has been utilized to monitor and evaluate the returns of corn production process in the US.

8.2.3 Using the $\beta - factor$ of the Intensity Function to Monitor and Evaluate Production Process of Corn based on the Returns

Now, we utilize the NHPP to monitor, assess and evaluate the corn production process of a firm based on the returns of production. The return is the end outcome of production, hence, the most important component that strongly influences a firm's decision-making process and regulates its operationality. The returns could mean a profit or loss to the production firm. The returns of production, R is a time-dependent phenomenon (i.e. changes with time) derived from the total revenue (TR) and the total cost TC, which are both time-dependent variables. Given that TR and TC are function of time t, we can express R as

$$\hat{R}_t(X) = T\hat{R}_t(X) - T\hat{C}_t(X), t > 0. \tag{8.8}$$

Where the predicted or estimated production returns $\hat{R}_t$ with 99% accuracy is a function of time, given by R_p in Section (8.2.1). For instance, $\hat{R}_1$ for year 1 of a production firm is given by $\hat{R}_1(X) = T\hat{R}_1(X) - T\hat{C}_1(X)$, year 2 is $\hat{R}_2(X) = T\hat{R}_2(X) - T\hat{C}_2(X)$, and year n is $\hat{R}_n(X) = T\hat{R}_n(X) - T\hat{C}_n(X)$. $\hat{R}_t$ is negative if $T\hat{C}_t > T\hat{R}_t$, positive if $T\hat{C}_t < T\hat{R}_t$, and zero if $T\hat{C}_t = T\hat{R}_t$. X denotes the contributable factors. Note that both $T\hat{R}_t$ and $T\hat{C}_t$ are given by the identified contributable factors to $\hat{R}_t$, given by Table 8.2. If $T\hat{C}_t > T\hat{R}_t$ implies the production firm is incurring loss. If $T\hat{C}_t < T\hat{R}_t$ implies the production firm is making some profit. The production firm is said to be at break-even stage if $T\hat{C}_t = T\hat{R}_t$.

Given that we rearrange $\hat{R}_t$ such that $\hat{R}_1 < \hat{R}_2 < ... < \hat{R}_n$ in ascending order of magnitude from the least value of $\hat{R}$ to the largest value as a function of time, then the probability distribution behavior of $\hat{R}_t$ can be said to follow the nonhomogeneous Poisson process (NHPP) or the power law process (PLP), given by equation (8.4). Given that the $\hat{R}_t$ after is rearranged from smallest to highest follow the NHPP, we can compute the intensity function using equation (8.1) to evaluate the failure intensity or changes in the production process as a function of time, and based on R, given by

$$\hat{V}(\hat{R}_t; \beta, \theta) = \frac{\hat{\beta}}{\hat{\theta}} \left(\frac{\hat{R}_t}{\hat{\theta}} \right)^{\hat{\beta}-1}. \tag{8.9}$$

Consequently, we can find an estimate of the shape parameter, β, and thus, evaluate the failure intensity in the production process based on the returns (i.e. the intensity function, $V(\hat{R}_t)$). The estimate of the shape parameter $\hat{\beta}$ and the scale parameter $\hat{\theta}$ of the production process based on the predicted returns of corn production $\hat{R}_t$, which is a function of time can respectively be expressed as

$$\begin{gathered} \hat{\beta}_n = \frac{n}{\sum_{i=1}^{n} log\left(\frac{\hat{R}_n}{\hat{R}_i}\right)}, \\ and \\ \hat{\theta}_n = \frac{\hat{R}_n}{n^{1/\hat{\beta}}}. \end{gathered} \tag{8.10}$$

For $\hat{R}_1 < \hat{R}_2 < ... < \hat{R}_n$ and n is the duration of the production. Equation (8.10) mimics equations (8.6) and (8.7). Now, given that we have computed the $\hat{\beta}$ based on $\hat{R}_t$, we interpret the changes in $\hat{\beta}$ as follows: If $\hat{\beta} < 1$ means that R of production is rising. In other words, the rate of failure of production or $\hat{V}(\hat{R}_t)$ is depreciating, hence the production process or factors may require no changes. If $\hat{\beta} > 1$, then the R of production is decreasing. This implies that $\hat{V}(\hat{R}_t)$ is increasing, hence there is the need for some adjustment in the factors of production. Finally, if $\hat{\beta} = 1$ (which is rare) delineate constant R. That is, $\hat{V}(\hat{R}_t)$ remains unchanged, and the production firm may decide whether or not to make changes to improve the production process by making changes in the values identified contributable factors in Table 8.2. It is important to note that the nonhomogeneous Poisson process or the power law process is often used to model the reliability growth of repairable and nonrepairable systems. An example is modeling the reliability of a software system proposed by Freeh

N. Alenezi and Christ P. Tsokos (2019). Thus, the NHPP allows for the assessment of the useful life of machines, software, equipment, etc. used in a system. Therefore, NHPP will be more efficient in assessing the production process, which makes use of several technologies.

Interestingly, the cost of production is a function of several factors including the cost of buildings, machines, repairs, equipment, among others. Whereas the returns of production R is a function of the cost of production TC. Thus, there is a very strong correlation between R and both TC and TR. So, using NHPP to monitor the production process of corn based on returns provides a much robust result to aid and guide the decision-making process of the US Corn production firms. Both the original returns R and the estimated or predicted returns $\hat{R}_t$ are in transform form using Johnson transformation. The Johnson transformation suppresses or stabilizes the impact of large values on the returns during the model building stage. To estimate the parameters of the intensity function of the production returns, we further transformed the predicted returns $\hat{R}_t$ to obtained only positive values that take the logarithm function of $\hat{\beta}$ expression in equation (8.10), given by

$$\hat{R}_t = \hat{R}_p + max(\hat{R}_p), \tag{8.11}$$

where $max(\hat{R}_p)$ is the maximum value of the transformed returns.

8.3 Results

In the present section, we show the results of applying the NHPP to monitor and evaluate corn production in the US. Figure 8.3 shows the rank of the predicted returns of corn production $\hat{R}_t$ in ascending order of magnitude with the year of production as a time index on the horizontal axis. The probability distribution behavior of the corn production returns in Figure 8.3 follows the nonhomogeneous Poisson process (NHPP), thus, the power-law process (PLP), given by $\hat{R}_1 < \hat{R}_2 < ... < \hat{R}_{44}$.

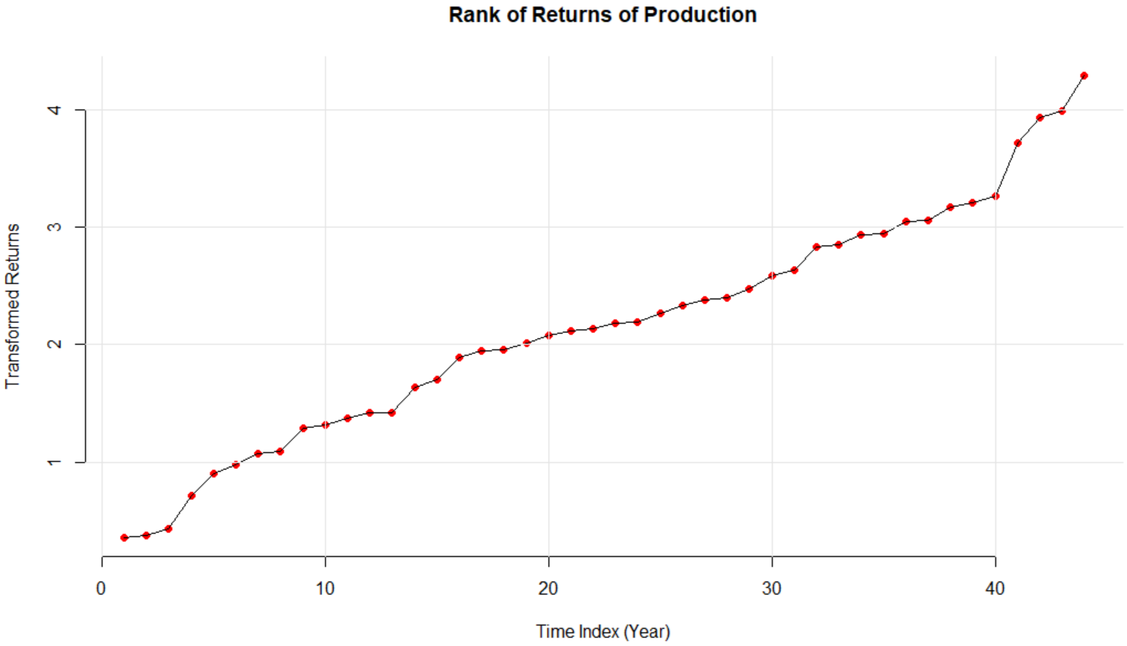

Figure 8.3: Plot of Ranking of Returns of Corn Production

We computed the failure rate or intensity $\hat{V}(\hat{R}_t; \hat{\beta}, \hat{\theta})$ as a function of production returns with parameters β and θ. We estimated β and θ as given by equation (8.10). Consequently, we can evaluate the production returns by assessing the behavior or changes in $\hat{\beta}$. Now, we estimate the $\hat{\beta}$ of the entire forty-four years of corn production in the US, given by Table 8.3. We see that $\hat{\beta} = 3.047 > 1$, meaning the failure rate of production is increasing, hence the production returns R is falling. Thus, the forty-four years of corn production show that the rate of failure intensity of the production of corn is rising. We can see from Figures 8.1 and 8.2 that R generally is falling, which exactly corresponds to the value we had for $\hat{\beta} > 1$ after the forty-four years of corn production. This is further supported by the negative mean (-16.83 dollars per planted acre) and median (-26.28 dollars per planted acre) returns of US corn production over the forty-four year period of production, given in Table 8.1. This finding suggests the need to implement strategic changes to reverse the corn production process of the US by assessing the impact of the identified contributable factors in Table 8.2. This justifies the high quality and efficiency of our analysis of monitoring the production returns using the NHPP.

Table 8.3: $\hat{\beta} - \mathit{factor}$ Evaluation of Forty-Four Years of Corn Production of the US

Estimates	
$\hat{\beta}$	$\hat{\theta}$
3.047	0.473

In Table 8.4 and Figure 8.4, we monitor and evaluate the corn production returns from 2008 to 2018 using the $\beta - \mathit{factor}$ estimates. Notice that to compute the estimates, we can take into account the previous years' returns of production of the firm, given that the firm has been in operation for more than a year. For instance, the estimates for the fourth year take into account the returns of production of the previous three years. This is a very intriguing feature of using the NHPP because we can evaluate both the current and previous years of the firm's production process, and how it could be affecting the firm's total production returns entirely. From Table 8.4, $\hat{\beta} < 1$ means that the failure intensity of the production process is decreasing, implying the return of corn production is increasing. This further means the ongoing production process is reliable or dependable. Also, $\hat{\beta} > 1$ implies the failure intensity or rate of the production process is deteriorating. Thus, the return of production is decreasing and the ongoing production process is failing or depleting. This is exactly the case by assessing Figure 8.4. We see from Figure 8.4 that the production returns were in the ascendancy from 2005 to 2008 where $\hat{\beta} = -2.645 < 1$. In 2009, shows the returns fall with $\hat{\beta} = 1.839 > 1$. That is, the returns fall from \$99.98 per planted acre to \$10.52 per planted acre. From 2009, the returns started to rise, given by the negative $\beta - \mathit{factor}$ in 2010 and 2011, where the highest returns of production for the 44 years of corn production in the US were attained. After 2011, the returns of production started to fall again. At this stage, $\beta - \mathit{factor}$ remains less than one until 2014 where there was a swift fall of the returns from \$44.06 in 2013 to -\$86.62 and the $\beta - \mathit{factor}$ estimate rose very high above 1, given by $\hat{\beta} = 13.340 > 1$. This means no major adjustments are required in the production process until 2014, which shows the production process is in critical condition and needs immediate adjustments. The corn production returns of the US have generally been rising after 2014, given by $\beta - \mathit{factor}$ less than 1 and as shown in Figure 8.4. We can also see that as the returns of corn production increase, the $\beta - \mathit{factor}$ estimates attain higher negative values (i.e. falling). Whereas the $\beta - \mathit{factor}$ estimates attain smaller negative values (i.e. rises) as the production returns are falling.

Table 8.4: $\hat{\beta} - \mathit{factor}$ Yearly Evaluation of Corn Production Returns of US 2008-2018

		Estimates		
Year	Returns ($)	$\hat{\beta}$	$\hat{\theta}$	$\hat{V}(\hat{R}_t)$
2008	99.98	-2.645	5.361	-88.118
2009	10.52	1.839	0.456	19.886
2010	139.19	-1.819	8.543	-69.886
2011	224.31	-7.427	2.726	-199.574
2012	148.98	-2.635	5.359	-88.118
2013	44.06	-1.302	15.302	-57.858
2014	-86.62	13.340	2.840	20.819
2015	-62.73	-7.883	3.190	-33.092
2016	-74.56	-3.852	3.656	-14.462
2017	-64.28	-17.141	2.941	-59.627
2018	-48.10	-3.307	3.780	-27.204

8.4 Discussion

The amount of dollars left after all revenue and cost is vital in influencing the decision-making process of a firm in corn production. This is referred to as the returns, representing the difference between the total revenue and total cost, which can be a gain or loss. Every firm in production aims to maximize profit and minimize costs or losses. The continuous operational existence of a production firm heavily depends on how best it regulates or manages the returns. Many firms are struggling and going through difficulties about the best strategy to monitor and evaluate their production process. The usual strategy has been assessing the total revenue (TR) and total cost (TC) or the marginal cost (MC) and marginal revenue (MR) of a given production period and making decision based on when $MR < MC$, $MR > MC$, or $MR = MC$. This strategy may work for just a particular production year but may flop as production continues. It only considers the production process of the current

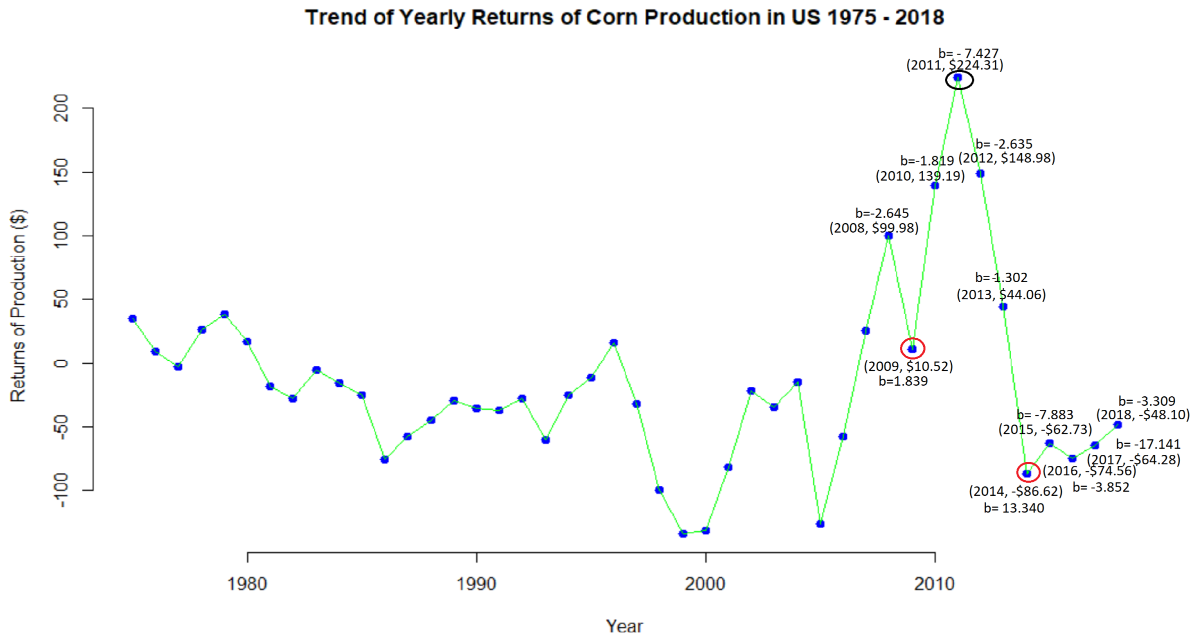

Figure 8.4: Plot of $\hat{\beta} - \mathsf{factor}$ Yearly Evaluation of Corn Production Returns, 2008-2018

period and not the prior years of production. Just considering the TC and TR to make decision about the production process can be misleading. For instance, if a firm makes negative returns (loss) in year 1 and positive returns (profit) in year 2, does it means it is profitable, and hence the production process is in good shape? Evaluating only year 2 production returns may not aid the firm in making the correct decision. The firm may still be in danger although it made a profit in year 2. Therefore, making decision by just evaluating MC and MR over a period of production may not be reliable or dependable. In the real world, it is tedious for firms to know exactly where their MR equals MC of the last product sold to adjust the production process [116]. MR and MC concept of monitoring the profitability of a production firm has the limitation of largely relying on trial and error strategy [122]. Another known strategy of monitoring and evaluating the production process has been comparing the market value or price of the product to the average cost (AC) of production. This approach is also problematic because in most cases the market price of a product is out of the control of the production firm [123]. It may work well for monopolistic production firms, but generally has a lot of flaws. So, what strategy can production firms adopt to enhance the decision-making of the production process and ensure that correct and reliable decisions are made about the returns of production? The current study introduces a time-dependent analytical model called the nonhomogeneous Poisson process (NHPP) or power-law process (PLP) to monitor and evaluate the production process of US corn production based on the returns.

The proposed time-dependent analytical model for monitoring and evaluating the production process using the returns of production can be said to be an advanced version of the strategy comparing or evaluating TC and TR or MR and MC. Firms do not have to worry about the price of the product, which in most cases is out of their control. The return, which is the end outcome of production plays a tremendous role in the functionality and the existence of every production firm. Before we applied the proposed model to monitor the returns of corn production in the US, an analytical model was first developed to identify key contributable factors that predict the returns of corn production in the US with 99% accuracy[105]. It is important to base our assessment on well-defined and well-validated returns of production. The present study focused on assessing the predicted returns and using it as a catalyst to monitor and evaluate the production process of corn in the US. The nonhomogeneous Poisson process (NHPP) is often used in the assessment of the reliability growth of a system or phenomenon which may be repairable, nonrepairable, tangible, or intangible. It allows for the detection of component defects, flaws, repairs, removal, or replacements in the production process. The NHPP is used in modeling the reliability process of a phenomenon that is time-dependent, in our case the returns of corn production R_t. The forty-four years returns of US corn production were rearranged in ascending order of magnitude to follow the probability distribution behavior of NHPP. We then estimated the intensity function $V(R_t; \beta, \theta)$, which measures the failure intensity of the production process as a function of R_t with the parameter estimates of $\hat{\beta}$ and $\hat{\theta}$. To monitor and evaluate the production returns R_t, we utilized the $\beta - \mathit{factor}$ which measures the shape parameter of the intensity function. We applied the Johnson transformation on R_t to minimize the impact of large outlying values; then further transformed it to eliminate negative values to accommodate the logarithmic function of the β formula in equation (8.10). The $\hat{\beta}$ was then calculated utilizing the MLE method of parameter estimation, and then interpreted based on when $\hat{\beta} > 1, < 1$ or $= 1$ as given by Section 8.2.3 of this study.

We applied the method utilizing the $\beta - \mathit{factor}$ to evaluate the forty-four years returns of corn production in the US from 1975 to 2018, given by Table 8.3; and further evaluate the yearly returns from 2008 to 2018, given by Table 8.4. We can see in Figure 8.4 that the returns of corn production in the US generally falls, which is consistent with the identified $\hat{\beta} = 3.047 > 1$, indicating that the failure intensity of the production process of corn in the US is increasing, resulting in decreasing returns. This triggers the need for immediate turnaround and serious adjustments in the corn production system in the US. The finding is also consistent with the corn returns projection by Central Illinois University of Illinois Ag Economists; Carl Zulauf; and Ohio State University Ag Economists, who reported decreasing returns of corn production after 2018, prompting the need for more Federal Aid [124]. The result was expected due to the economic uncertainties during the 44 years period of production. In a report published by Scientific American in 2013, Jonathan Foley outlined the need for a rethink of America's system of corn production in the fact that the current corn system is highly vulnerable to shocks [125]. We further monitored and assessed the yearly returns of US corn production from 2008 to 2018. After 2008 shows a swift fall in 2009, given by $\hat{\beta} = 1.839 > 1$, implying that the production process is failing with returns falling. This is no surprise and was expected because of the economic downturn caused by the "Great Recession" between 2007 and 2009 which resulted in a decline of global GDP by 5.1% and 10% peak of global unemployment in October 2009 [126]. The USDA also reported in 2009 that the world economic crisis in 2008 had major consequences on US agriculture as a result of the downslide of export demand and reduction in commodity prices [127]. After 2009, it is shown that the failure intensity of the production process started falling, given by $\beta - \mathit{factor}$ less than one until 2014 where the $\beta - \mathit{factor}$ became greater than one (i.e.

an increased failure intensity of production process). The increasing returns after 2009 were expected due to economic recovery after the recession period. From 2011 to 2014, it is shown that the return was falling, which is indicated by the rising value of the $\beta-\mathsf{\mathit{factor}}$ from -7.427 to 13.340 in 2014. An intriguing observation to note is the fact that the $\beta-\mathsf{\mathit{factor}}$ does not suddenly become greater than one with a decreased returns. Instead, an increasing $\beta-\mathsf{\mathit{factor}}$ hints and prompts the production firm the need to take a closer look at the ongoing process for some possible defects or failure before it gets worse where $\beta-\mathsf{\mathit{factor}} > \mathsf{1}$. However, a fall in the returns of production does not necessarily mean a failure or decline in the efficiency of the ongoing production process but could be a result of impact from other confounding contributing factors beyond the control of the corn production firm or out of the realm of the current process such as weather. The switch turnaround after the dip in the returns from 2008 to 2009 was most likely because the US agriculture was not severely affected by the economic crisis due to record-high agricultural exports, prices, and farm income in 2007 and 2008, which put U.S. farmers on solid financial ground., echoed by the USDA in March 2009 [127]. In 2014 shows the corn production process requires serious adjustments, given by the high $\beta-\mathsf{\mathit{factor}}$ of 13.340. After 2014, the returns have remained negative. However, the general trend after 2014 shows the returns are increasing, as also indicated by the values of the $\beta-\mathsf{\mathit{factor}}$ been less than one. Notice that a negative return does not always mean a failing production process. This is another significant feature the profit maximization concept of using the MR and MC lacks. In an article by Smriti, Chang stated empirical evidence by Hall and Hitch shows that businessmen have not heard of marginal cost and marginal revenue[123]. Also, Tejvan Pettinger (2019) outlined in an article on profit maximization that firms may have other social objectives leading to increasing their profitability and do not care about the behavior of MR and MC [122].

Regarding the monitoring strategy of comparing AC and price of the product, it explains that a firm may be in a critical condition if the price falls below AC, and allows the firm to consider shutting down. This raises a strong argument or critique of what if the rise in AC over price has nothing to do with the current production process, rather as a result of unusual/seasonal events during that period such as the economic recession or weather condition. Also, the suggestion of shutdown as mitigation to a firm is not a good strategy to resolve most production predicaments, especially with firms who are still recovering their fixed cost of production. Therefore, the $\beta - \mathit{factor}$ of the intensity function of NHPP applied to corn production in this study provides a better and more advanced production monitoring and evaluation strategy. It uses the returns predicted from a well-structured and well-validated model with 99% accuracy taking into account other confounding factors affecting the returns. The monitoring process is also time-dependent and takes into consideration the prior returns of production of corn in the US. The application of the NHPP in production is very essential and interesting because not only are we concerned with monitoring the changes in the returns of corn production, we may also be concerned about the assessment of the technologies been used in the production of the corn. Thus, we can as well assess the reliability of the technologies used in the production process. This makes applying the NHPP in production a powerful technique for monitoring and evaluation of the production process. Our proposed production process monitoring and evaluation method or strategy can applied to any type of production or business operation.

8.5 Contribution

We have proposed an analytical method for monitoring, assessing, and evaluating the production returns of corn production in the US from 1975 to 2018 using the nonhomogeneous Poisson process (NHPP). We have shown that the current method is more useful, effective, and efficient to apply compared to the existing concept of comparing MR and MC or market price and the average cost of production, which is mostly dependent on trial and error technique. The NHPP can model the reliability growth of production taking into consideration the failure intensity of the production process. The failure intensity, on the other hand, is influenced by the shape parameter ($\beta - \mathit{factor}$). By assessing the changes in the $\beta - \mathit{factor}$, we monitored and evaluated the returns of US corn production and its effect on the production process. That is if $\hat{\beta} > 1$, means the return is failing. Thus, the failure intensity of the production process is said to be accelerating, hence immediate attention and adjustment would be required. If $\hat{\beta} < 1$, means failure intensity of the production process is decreasing, an indication of increasing returns. Therefore, no major improvement or changes to the current process may be required. This also means the firm's production is in good shape.

Finally, if $\hat{\beta} = 1$, or approximately equal one (which rarely happens) indicates a constant return. Thus, the failure intensity of production remains the same. At this point, the current process may or may not require any alteration or improvement. The application of this method does not require us to know the market price of the product (marginal revenue), which is mostly out of the control of the firm. All we need to know is the returns, predicted from significantly identified contributable factors of the corn production model. These factors are usually available to all corn production firms. The US corn production returns were found to be decreasing. Therefore, we recommend the need for some adjustment in the production process to enable the US to stay at the top as the global leading producer of corn. The current method can also be used to monitor the factors of production, including the technologies used in production from which the total cost is incurred. It assesses the defects, breakages, worn-outs, repairs, or removal in the production system. Hence, we have no doubt the effectiveness, quality, and usefulness of our proposed analytical method in monitoring and evaluating the production returns, has proven to be an improved strategy for production management in any production firm/industry/factory. The future research study will focus on state-wise monitoring and evaluation of the returns of corn production.

Chapter 9: Future Research Work

9.1 Big Data Analysis of Human DNA Impact on Cancer Survivorship

The human cells are faced with several thousands of DNA. These DNA repairs preserve the functionality and survival of the cells. The repair of the DNA can result in an unstable genome and cancer. Our future study on cancer survivorship will focus Big data analysis taking into consideration the various types of human DNA impact on the length of cancer survival. Given the characterized DNA of an individual diagnosed with cancer may influence its development or growth rate, consequently the length of survival of cancer. Our interest will be to investigate the association of type of DNA on the length of patients' survival of particular cancer.

9.2 Assessment of the Stage Cancer Survivorship

Given that the length of survival of a type of cancer is greatly influenced by the stage of cancer, whether insitu, regional, or distance/malignant/invasive. Our interest will be to assess the contributing risk factors at the various stages of cancer and how they impact the length of survival time.

9.3 Application of Nonhomogeneous Poisson Process (NHPP) / Power Law Process (PLP) to the Survival Time

How do we tell whether the survival time of an individual diagnosed with a kind of cancer increases, decreases, or remains unchanged? In a further study, we want to investigate how we can utilize the intensity function of the NHPP to evaluate the changes in the survival times of patients diagnosed with cancer. The objective is to provide in-depth information about the rate of survival of a given cancer.

9.4 Application of Simulations of the NHPP in Monitoring the Production Process

Future studies in monitoring the returns of production will focus on adopting some simulation techniques (for example, the Markov Chain Monte Carlo Simulation, MCMC) to assess the behavior or changes in the returns. This will be a further approach to justify the robustness of using the NHPP in assessing the production process.

9.5 The Impact of External Factors or Indicators on Corn Production in the U.S.

Besides the internal controllable factors or indicators of crop production of a firm that influences the returns, there are several exogenous factors (known to be confounding factors) that may greatly impact the production returns. Our goal will be to investigate how exogenous factors such as GDP, inflationary rate, employment rate, government policy, among others, influence the production returns compared to the internal factors. Thus, to assess which factors have a greater impact on the returns of crop production.

References

[1] Max Roser and Hannah Ritchie (2019). “Cancer”, Our World in Data. https://ourworldindata.org/cancer

[2] Raab MS, Podar K, Breitkreutz I, Richardson PG, Anderson KC (July 2009). ”Multiple myeloma”. Lancet. 374 (9686): 324–39. doi:10.1016/S0140-6736(09)60221-X. PMID 19541364

[3] Ferri, Fred F. (2013). Ferri’s Clinical Advisor 2014 E-Book: 5 Books in 1. Elsevier Health Sciences. p. 726. ISBN 978-0323084314.

[4] SEER Cancer Facts: Myeloma. National Institute. Bethesda, MD, https://seer.cancer.gov/statfacts/html/mulmy.html

[5] About Multiple Myeloma. American Cancer Society. https://www.cancer.org/cancer/multiple-myeloma/about/what-is-multiple-myeloma.html

[6] “Plasma Cell Neoplasms (Including Multiple Myeloma) Treatment”. National Cancer Institute. 1980-01-01. Retrieved 28 November 2017. https://www.cancer.gov/types/myeloma/patient/myeloma-treatment-pdq#section/all

[7] Van de Donk NW, Mutis T, Poddighe PJ, Lokhorst HM, Zweegman S (2016). ”Diagnosis, risk stratification and management of monoclonal gammopathy of undetermined significance and smoldering multiple myeloma”. International Journal of Laboratory Hematology. 38 Suppl 1: 110–22. doi:10.1111/ijlh.12504. PMID 27161311

[8] World Cancer Report 2014. World Health Organization. 2014. pp. Chapter 5.13. ISBN 978-9283204299.

[9] Roberts, DL; Dive, C; Renehan, AG (2010). ”Biological mechanisms linking obesity and cancer risk: new perspectives”. Annual Review of Medicine. 61: 301–16. doi:10.1146/annurev.med.080708.082713. PMID 19824817

[10] Dutta AK, Hewett DR, Fink JL, Grady JP, Zannettino AC (2017). ”Cutting edge genomics reveal new insights into tumour development, disease progression and therapeutic impacts in multiple myeloma”. British Journal of Haematology. 178 (2): 196–208. doi:10.1111/bjh.14649. PMID 28466550

[11] Landgren O, Kyle RA, Pfeiffer RM, Katzmann JA, Caporaso NE, Hayes RB, Dispenzieri A, Kumar S, Clark RJ, Baris D, Hoover R, Rajkumar SV (28 May 2009). ”Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma: a prospective study”. Blood. 113 (22): 5412–7. doi:10.1182/blood-2008-12-194241. PMC 2689042. PMID 19179464

[12] Korde N, Kristinsson SY, Landgren O (2011). ”Monoclonal gammopathy of undetermined significance (MGUS) and smoldering multiple myeloma (SMM): novel biological insights and development of early treatment strategies”. Blood. 117 (21): 5573–5581. doi:10.1182/blood-2011-01-270140. PMC 3316455. PMID 21441462

[13] Kyle RA, Rajkumar SV (2008). ”Multiple myeloma”. Blood. 111 (6): 2962–72. doi:10.1182/blood-2007-10-078022. PMC 2265446. PMID 18332230.

[14] McCarthy, P. L; Holstein, S. A; et al (July 27, 2017). “Lenalidomide Maintenance After Autologous Stem Cell Transplant in Newly Diagnosed Multiple Myeloma: a Meta-Analysis”. J Clin Oncol. 35 (29): 3279–3289. doi:10.1200/JCO.2017.72.6679. PMC 5652871. PMID 28742454.

[15] Sonneveld, P (July 16, 2012). "Bortezomib induction and maintenance treatment in patients with newly diagnosed multiple myeloma". J Clin Oncol. 30 (24): 2946–55. doi:10.1200/JCO.2011.39.6820. PMID 22802322

[16] Siegel, R. L., Miller, K. D. Jemal, A. Cancer Statistics, 2017. CA Cancer J. Clin. 67, 7–30 (2017)

[17] Cancer Facts Figures (2014) American Cancer Society, http://www.cancer.org/acs/groups/content/@research/documents/webcontent/ acspc-04215.pdf.Accessed6March2014

[18] Becker N (2011) Epidemiology of multiple myeloma. Recent Results Cancer Res 183:25-35

[19] Meyer, Bruce D. (1990). "Unemployment Insurance and Unemployment Spells" (PDF). Econometrica. 58 (4): 757–782.

[20] Brian G.M.Durie et al. Pretreatment Tumor Mass, Cell Kinetics, and Prognosis in Multiple Myeloma Department of internal Medicine. College of Medicine, University of Arizona, Tucson, Ariz. 85724.© I 980 by Grune Stratton. Inc.

[21] Shaji K. Kumar et al. Improved survival in multiple myeloma and the impact of novel therapies.

[22] Feigl, P. and Zelen, M. [1965]. Estimation of exponential survival possibilities with concomitant information. Biometrics 21, 826-38.

[23] John M. Krall, Vinceent A. Uthoff, John B Harley: A set-up procedure for selecting variables associated with survival. Biometrics 31, 49-57, March 1975.

[24] Harley, J. B. [1971]. Ten years of experience in multiple myeloma at the West Virginia University Hospital. In preparation. Morgantown, West Virginia.

[25] Peto, Richard; Peto, Julian (1972). "Asymptotically Efficient Rank Invariant Test Procedures". Journal of the Royal Statistical Society, Series A. Blackwell Publishing. 135 (2): 185–207. doi:10.2307/2344317. JSTOR 2344317.

[26] Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics, 2002. CA Cancer J Clin. 2005;55:74–108.

[27] Multiple Myeloma Research Foundation (MMRF)
https://themmrf.org/multiple-myeloma/prognosis/

[28] Rogers, C. S., Gilnack, M., and Fitz III, H. C. (1983), "Monitoring of Coral Reefs with Linear Transects: A Study of Storm Damage," Journal of Experimental Marine Biology and Ecology, 66, 285-300.

[29] Giampaolo Merlini, Jan G. Waldenstrom, and Suresh D. Jayakar. A New Improved Clinical Staging System for Multiple Myeloma Based on Analysis of 123 Treated Patients. University of Lund, Malmo General Hospital, 5-214 01 Malmo, Sweden. (c)/ 980 by Grune Stratton, Inc.0006-4971/80/5506–0022$0I.00/0

[30] Doane, David P., and Lori E. Seward. "Measuring Skewness: A Forgotten Statistic?" Journal of Statistics Education 19.2 (2011): 1-18.

[31] Sharma, R.; Bhandari, R. (2015). "Skewness, kurtosis and Newton's inequality". Rocky Mountain Journal of Mathematics. 45 (5): 1639–1643.

[32] Chambers, Raymond L.; Steel, David G.; Wang, Suojin; Welsh, Alan (2012). Maximum Likelihood Estimation for Sample Surveys. Boca Raton: CRC Press. ISBN 978-1-58488-632-7.

[33] Calitz, F. (1973), "Maximum Likelihood Estimation of the Parameters of the Three Parameter Lognormal Distribution - a Reconsideration," Australian Journal of Statistics, 9, 221-226.

[34] Cohen, A. C. J. (1951), "Estimating Parameters of Logarithmic-Normal Distributions by Maximum Likelihood," Journal of the American Statistical Association, 46, 206-212.

[35] Cohen, A. C. J., and Whitten, B. J. (1980), "Estimation in the Three-Parameter Lognormal Distribution," Journal of the American Statistical Association, 75, 399-404.

[36] Chen, C. (2006), "Tests of Fit for the Three-Parameter Lognormal Distribution,"Computational Statistics Data Analysis, 50, 1418-1440.

[37] Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958;53:457–481.

[38] Daniel S. et al, Medication use and multiple myeloma risk Los Angelos County.

[39] Alexanian R., Balcerzak S., Bonnet J.D., Gehan EA, Haut A,Hewlett J.S., Monto R.W.: Prognostic factors in multiple myeloma. Cancer36:1192-l2Ol, 1975

[40] Durie B.G.M., Salmon S.E.: A clinical staging system for multiple myeloma. Cancer 36:842-854, 1975

[41] Bergsagel D.E.: Plasma cell myeloma: Prognostic factors and criteria of response to therapy, in Staquet Mi (ed): Cancer Therapy. New York, Raven, 1975, pp 73-87

[42] H. Akaike, "A new look at the statistical model identification," IEEE Trans. Autom. Control, vol. 19, no. 6, pp. 716–723, 1974.

[43] Cox, David R (1972). "Regression Models and Life-Tables". Journal of the Royal Statistical Society, Series B. 34 (2): 187-220.

[44] L. Douglas Case; et al (June 2002). "Interpreting Measures of Treatment Effect in Cancer Clinical Trials". The Oncologist. 7 (3): 181–187.

[45] Brody, Tom (2011). Clinical Trials: Study Design, Endpoints and Biomarkers, Drug Safety, and FDA and ICH Guidelines. Academic Press. pp. 165–168.

[46] Kleinbaum DG. Survival Analysis: A Self-Learning Text. New York: Springer, 1997;1-324.

[47] Kalbfleisch JD, Prentice RL. The Statistical Analysis of Failure Time Data. New York: John Wiley Sons, 1980;1-321.

[48] L. Douglas Case, Gretchen Kimmick, Electra D. Paskett, Kurt Lohman and Robert Tucker. Interpreting Measures of Treatment Effect in Cancer Clinical Trials. The Oncologist 2002, 7:181-187. doi: 10.1634/theoncologist.7-3-181

[49] Sikander Ailawadhi et al. Disease and outcome disparities in multiple myeloma: exploring the role of race/ethnicity in the Cooperative Group clinical trials.

[50] Peto, Richard; Peto, Julian (1972). "Asymptotically Efficient Rank Invariant Test Procedures". Journal of the Royal Statistical Society, Series A. Blackwell Publishing. 135 (2): 185–207. doi:10.2307/2344317. JSTOR 2344317.

[51] Lütkepohl, H. Xu, F. Empir Econ (2012) 42: 619. https://doi.org/10.1007/s00181-010-0440-1

[52] Glantz, Stanton A.; Slinker, B. K. (1990). Primer of Applied Regression and Analysis of Variance. McGraw-Hill. ISBN 978-0-07-023407-9

[53] Draper, N. R.; Smith, H. (1998). Applied Regression Analysis. Wiley-Interscience. ISBN 978-0-471-17082-2.

[54] Doane, David P., and Lori E. Seward. "Measuring Skewness: A Forgotten Statistic?" Journal of Statistics Education 19.2 (2011): 1-18.

[55] Efron, B.; Tibshirani, R. (1993). An Introduction to the Bootstrap. Boca Raton, FL: Chapman Hall/CRC. ISBN 0-412-04231-2.

[56] Efron, B. (1979). "Bootstrap methods: Another look at the jackknife". The Annals of Statistics. 7 (1): 1–26.

[57] Costa G, Engle R, Taliente F: Criteria defining risk and response in multiple myeloma. Proc Am Assoc Cancer Res 10:15, I 969 (abstr)

[58] Kiang DT, Goldman A, Fortuny I, Theologides A, Kennedy BJ: Prognostic factors in multiple myeloma. Proc Am Assoc Cancer Res 14:107, 1973 (abstr)

[59] Kyle RA, Bayrd ED: The Monoclonal Gammopathies-Multiple Myeloma and Related Plasma-Cell Disorders. Springfield, Ill, Charles C Thomas, I 976, p 159

[60] Lohuwa Mamudu, Chris P. Tsokos (2020). Parametric and Non- Parametric Analysis of the Survival Times of Patients with Multiple Myeloma Cancer. Open Journal of Applied Sciences, 10, 118-134. DOI: 10.4236/ojapps.2020.104010

[61] Lohuwa Mamudu, Chris P Tsokos, Otunuga Oluwaseun E (2020). Survival Analysis of Multiple Myeloma Cancer Using the Cox-PH Model. Medical Clinical Research Journal ISSN 2577 – 8005. doi.org/10.33140/MCR.05.07.05.

[62] M. Lohuwa and C. Tsokos (2020). Data-Driven Statistical Modeling and Analysis of the Survival Times of Multiple Myeloma. Health Science Journal 14:1.. DOI: 10.36648/1791-809X.14.1.693

[63] Mamudu, L. and Tsokos, C. (2021) A New Statistical Modeling Approach for Survival Analysis of Cancer Patients—Multiple Myeloma Cancer. Open Journal of Applied Sciences, 10, 365-378. doi: 10.4236/ojapps.2021.104027.

[64] Kruskal; Wallis (1952). ”Use of ranks in one-criterion variance analysis”. Journal of the American Statistical Association. 47 (260): 583–621.

[65] Corder, Gregory W.; Foreman, Dale I. (2009). Nonparametric Statistics for Non-Statisticians. Hoboken: John Wiley Sons. pp. 99–105.

[66] Douglas G. Altman and Bianca L. De Stavola (1994). Practical problems in fitting a proportional hazards model to data with udated measurements of the covariates. Statistics in Machine, Vol 13, Issue 4, Pages 301-341. doi.org/10.1002/sim.4780130402

[67] Ian Ford, John Norrie, and Susan Ahmadi (1995). Model inconsistency, illustrated by the cox proportional hazards model. Statistics in Machine, Vol 14, Issue 8, Pages 735-746. doi.org/10.1002/sim.4780140804

[68] "Corn Production by Country in 1000 MT". Index Mundi. Retrieved June 3, 2013. https://www.indexmundi.com/agriculture/?commodity=corn

[69] United States Department of Agriculture Economic Research Service. https://www.ers.usda.gov/topics/crops/corn-and-other-feedgrains/

[70] "Production and Use". Iowa Corn organization. Retrieved 6 March 2014. https://www.iowacorn.org/corn-production/

[71] Elliott, Foster Floyd (1933). Fifteenth census of the United States. Census of agriculture. Types of farming in the United States. United States Government Printing Office. pp. 47–.

[72] Smith, C. Wayne (2004). Corn: Origin, History, Technology, and Production. John Wiley Sons. pp. 134–. ISBN 9780471411840.

[73] "A story of technology and innovation". Corn Farmers Coalition Organization. Archived from the original on 19 March 2013. Retrieved 3 June 2013.

[74] Amelia Urry, April 20, 2015. http://grist.org/food/our-crazy-farm-subsidies-explained/

[75] Wile, Rob (July 18, 2012). "11 Wild Facts About Corn In America". Business Insider.

[76] Foreman, Linda F. "Characteristics and Production Costs of U.S. Corn Farms". United States Department of Agriculture. Retrieved June 4, 2013.

[77] Giola, Vincent (May 5, 2008). “The Importance of the Corn Economy”. The National Ledger.

[78] Meyer, Gregory (May 21, 2013). ”Corn prices tumble amid intense planting”. The Financial Times.

[79] “GRAINS: Soggy U.S. weather propels corn and soybean prices”. Reuters. May 28, 2013.

[80] “Corn for Home Heat: A Green Idea That Never Quite Popped”. March 2, 2015. Retrieved July 7, 2017.

[81] Mark Clayton (January 28, 2008). “Christian Science Monitor”. Retrieved October 6, 2014.

[82] Anna-Lisa Laca Nov 28, 2017. Factors Influencing Global Grain Production. https://www.agweb.com/article/factors-influencing-global-grain-production-\ \NAA-anna-lisa-laca

[83] Wescott, P. C., & Jewison, M. (n.d.). Economic Research Services. Weather Effects on Expected Corn and Soybean Yields . Retrieved March 10, 2014, from http://www.usda.gov/oce/forum/past_speeches/2013_Speeches/Westcott_Jewison.pdf

[84] Gardebroek, C., Hernandez, M. (n.d.). 123 EAAE Seminar. Price Volatility and Farm Income Stabilisation. Retrieved March 10, 2014, from http://ageconsearch.umn.edu/bitstream/122551/2/maitredhotel.pdf

[85] Wallander, S., Claassen, R., Nickerson, C. (n.d.). USDA ERS - The Ethanol Decade: An Expansion of U.S. Corn Production, 2000-09. USDA ERS - The Ethanol Decade: An Expansion of U.S. Corn Production, 2000-09. Retrieved April 23, 2014, from http://www.ers.usda.gov/publications

[86] Mamudu, L. and Tsokos, C.P. (2020) Parametric and NonParametric Analysis of the Survival Times of Patients with Multiple Myeloma Cancer. Open Journal of Applied Sciences, 10, 118-134.
https://doi.org/10.4236/ojapps.2020.104010

[87] M. H. Kutner, Applied linear statistical models., The McGraw Hill/Irwin series operations and decision sciences, McGraw-Hill Irwin, 2005.

[88] Abu Sheha, M. and Tsokos, C. (2019) Statistical Modeling of Emission Factors of Fossil Fuels Contributing to Atmospheric Carbon Dioxide in Africa. Atmospheric and Climate Sciences, 9, 438-455. doi: 10.4236/acs.2019.93030.

[89] Cardwell, V. B. "Fifty Years of Minnesota Corn Production: Sources of Yield Increase." Agronomy J. 74(November/December 1982):984

[90] Box, G.E.P.; Wilson, K.B. (1951). "On the Experimental Attainment of Optimum Conditions". Journal of the Royal Statistical Society: Series B. 13 (1): 1–45.
doi:10.1111/j.2517-6161.1951.tb00067.x

[91] Greg Ibendahl. "Economics of Corn Production".
http://www2.ca.uky.edu/agcomm/pubs/id/id139/economics.pdf

[92] Jeff Coulter, Extension agronomist, Reviewed in 2018."Strategies to optimize corn silage production".
https://extension.umn.edu/corn-planting/strategies-optimize-corn-silage-production

[93] Hubner Seed, January 2019. "Management Practices for Optimizing Yield and Productivity in Corn".
https://www.hubnerseed.com/en-us/research-library/management-practice-\
\for-optimizing-yield-and-productivity-in-corn.html

[94] Derringer G, Suich R. Simultaneous optimization of several response variables. J Quality Technol. 1980;12:214-219.

[95] Jeong IJ, Kim KJ. An interactive desirability function method to multiresponse optimization. Eur J Oper Res.2009;195(2):412-426. doi: 10.1016/j.ejor.2008.02.018.

[96] Harrington ECJr. The desirability function. Ind Quality Control.1965;21:494-498.

[97] Pasandideh SHR, Niaki STA. Multi-response simulation optimization using genetic algorithm within desirability function framework. Appl Math Comput. 2006;175(1):366-382. doi: 10.1016/j.amc.2005.07.023.

[98] Li J, Ma C, Ma Y, Li Y, Zhou W, Xu P. Medium optimization by combination of response surface methodology and desirability function: an application in glutamine production. Appl Microbiol Biotechnol. 2007;74(3):563-571. doi:10.1007/ s00253-006-0699-5.

[99] Ryad Amdoun, Lakhdar Khelifi, Majda Khelifi-Slaoui, Samia Amroune, Mark Asch, Corinne Assaf-Ducrocq, Eric Gontier. The Desirability Optimization Methodology; a Tool to Predict Two Antagonist Responses in Biotechnological Systems: Case of Biomass Growth and Hyoscyamine Content in Elicited Datura starmonium Hairy Roots. Iranian J Biotech. 2018 January;16(1):e1339. DOI:10.21859/ijb.1339

[100] USDA Economic Research Service, February 2020. “Feedgrains Sector at a Glance”. https://www.ers.usda.gov/topics/crops/corn-and-other-feedgrains/feedgrains-sector-at-a-glance/

[101] Published by Statista Research Department, Mar 31, 2014. “Revenue of corn farming (NAICS 11115) in the United States from 2009 to 2014(in billion U.S. dollars)”. https://www.statista.com/statistics/296374/revenue-corn-farming-in-the-us/

[102] Sandhu, H. et al. The future of agriculture and food: evaluating the holistic costs and benefits. Anthropocene Rev. 6, 270–278, https://doi.org/10.1177/2053019619872808 (2019).

[103] Jonathan Foley (2013). “It’s Time to Rethink America’s Corn System”. Scientific American. Retrieved March 5, 2013. https://www.scientificamerican.com/article/time-to-rethink-corn/

[104] Smith, C. Wayne (2004). Corn: Origin, History, Technology, and Production. John Wiley Sons. pp. 134. ISBN 9780471411840.

[105] Mamudu Lohuwa, Tsokos Chris P.2 (2020) Data-Driven Statistical Modeling and Analysis of the Returns Production in the United State. Provisional Patent No. 20A079PR (292107 – 8610)

[106] “Iowa Agriculture Quick Facts 2011”. Iowa Department of Agriculture and Land Stewardship. Archived from the original on 18 June 2015. Retrieved 5 June 2013. http://www.iowaagriculture.gov/quickfacts.asp

[107] Meyer, Gregory (May 21, 2013). “Corn prices tumble amid intense planting; Rapid pace makes record US harvest more likely”. Financial Times. Archived from the original on 2020-08-17. Retrieved August 17, 2020.

[108] Giola, Vincent (May 5, 2008). ”The Importance of the Corn Economy”. The National Ledger.

[109] Foreman, Linda F. “Characteristics and Production Costs of U.S. Corn Farms”. United States Department of Agriculture. Retrieved June 4, 2013.

[110] Agricultural Marketing Service, USDA (2015). “Commercial grain prices”. https://www.ams.usda.gov/mnreports/gx_gr113.txt

[111] Bruce A. Babcock and Gregory R. Pautsch (1998). Moving from Uniform to Variable Fertilizer Rates on Iowa Corn: Effects on Rates and Returns. Journal of Agricultural and Resource Economics , December 1998, Vol. 23, No. 2 (December 1998), pp. 385-400. Published by: Western Agricultural Economics Association.
https://www.jstor.org/stable/40986990

[112] Michael Langemeier, Center for Commercial Agriculture, Prude University (2017). 2017 Gross Revenue Scenarios For Corn. Retrieved May 8, 2017.
https://ag.purdue.edu/commercialag/home/resource/2017/05/2017-gross-revenue-scenarios-for-corn/

[113] David W Archer, Joseph L Pikul, Walter E Riedell. Economic risk, returns and input use under ridge and conventional tillage in the northern Corn Belt, USA. Soil and Tillage Research, Volume 67, Issue 1, 2002, Pages 1-8, ISSN 0167-1987
https://doi.org/10.1016/S0167-1987(02)00016-8

[114] Nolan E, Santos P (2019) Genetic modification and yield risk: A stochastic dominance analysis of corn in the USA. PLoS ONE 14(10): e0222156.
https://doi.org/10.1371/journal.pone.0222156

[115] Zulauf, C (2015). "Per Acre Net Cash Return to U.S. Corn and Soybeans since 1975: Part I." farmdoc daily (5):181, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, October 1, 2015.
https://farmdocdaily.illinois.edu/2015/10/per-acre-net-cash-return-us-corn-soybeans-1.html

[116] Prakeet Agarwal. Profit Maximization Rule. Intelligent Economist. Retrieved from June 30, 2019.
https://www.intelligenteconomist.com/profit-maximization-rule/

[117] Tsokos, C.P. (1995). Reliability Growth; Nonhomogeneous Poisson Process, Recent Advances in Life-Testing and Reliability, TA 169.3.R43, 1995, CRC Press.

[118] Ascher, H.E., Lin, T.T.Y., and Siewiorek, D.P. (1992), Modification of: Error log analysis: Statistical modeling and heuristic trend analysis, IEEE Transaction on Reliability, 41, 599-601.

[119] Ascher, H.E. and Feingold, H. (1984), Repairable Systems Reliability: Modeling, Inference, Misconceptions and Their Causes, Marcel-Dekker, New York.

[120] Bain, L.J. and Engelhardt, M (1991), Statistical Analysis of Reliability and Life Testing Models, Marcel-Dekker, New York.

[121] Basin, W.M. (1969), Increasing hazard functions and overhaul policy,Proceedings of the 1969 Annual Symposium on Reliability, Chicago, IEEE, 8, 173-178.

[122] Tejvan Pettinger (2019). “Profit Maximization”. Economics-Helping to Simplify Economics. Retrieved 16 July 2019.
https://www.economicshelp.org/blog/3201/economics/profit-maximisation/

[123] Smriti Chad. “Profit Maximisation Theory: Assumptions and Criticisms— Economics”.
https://www.yourarticlelibrary.com/economics/profit-maximisation-theory-\
\assumptions-and-criticisms-economics/28998

[124] Gary Schnitkey, Krista Swanson, Jonathan Coppess, and Nick Paulson, University of Illinois Ag Economists; and Carl Zulauf, Ohio State University Ag Economist (2020). Corn, Soybeans: Farm Profitability – More Federal Aid Needed. Published by AgFax. Retrieved June 11, 2020.
https://agfax.com/2020/06/11/corn-soybeans-farm-profitability-more-federal-aid-needed/.

[125] Jonathan Foley (2013). “It’s Time to Rethink America’s Corn System”. Published by Scientific American. Retrieve March 5, 2013. https://www.scientificamerican.com/article/time-to-rethink-corn/

[126] Wikipedia. “Great Recession”.
https://en.wikipedia.org/wiki/Great_Recession

[127] Mathew Shane, William M. Liefert, Mitch Morehart, May Peters, John Dillard, David Torgerson, and William Edmondson (2009). The 2008/2009 World Economic Crisis: What It Means for U.S. Agriculture. USDA Economic Research Service. Retrieved March 2009.

[128] Alenezi, F.N. and Tsokos, C.P. (2019). The effectiveness of the Square Error and Higgins-Tsokos Loss Functions on the Bayesian Reliability Analysis of Software Failure Times under the Power Law Process. *Engineering, 11, 272-299. https://doi.org/10.4236/eng.2019.115020*

Appendix A: Contour and 3D-Plots of Risk Factors Combination Effect

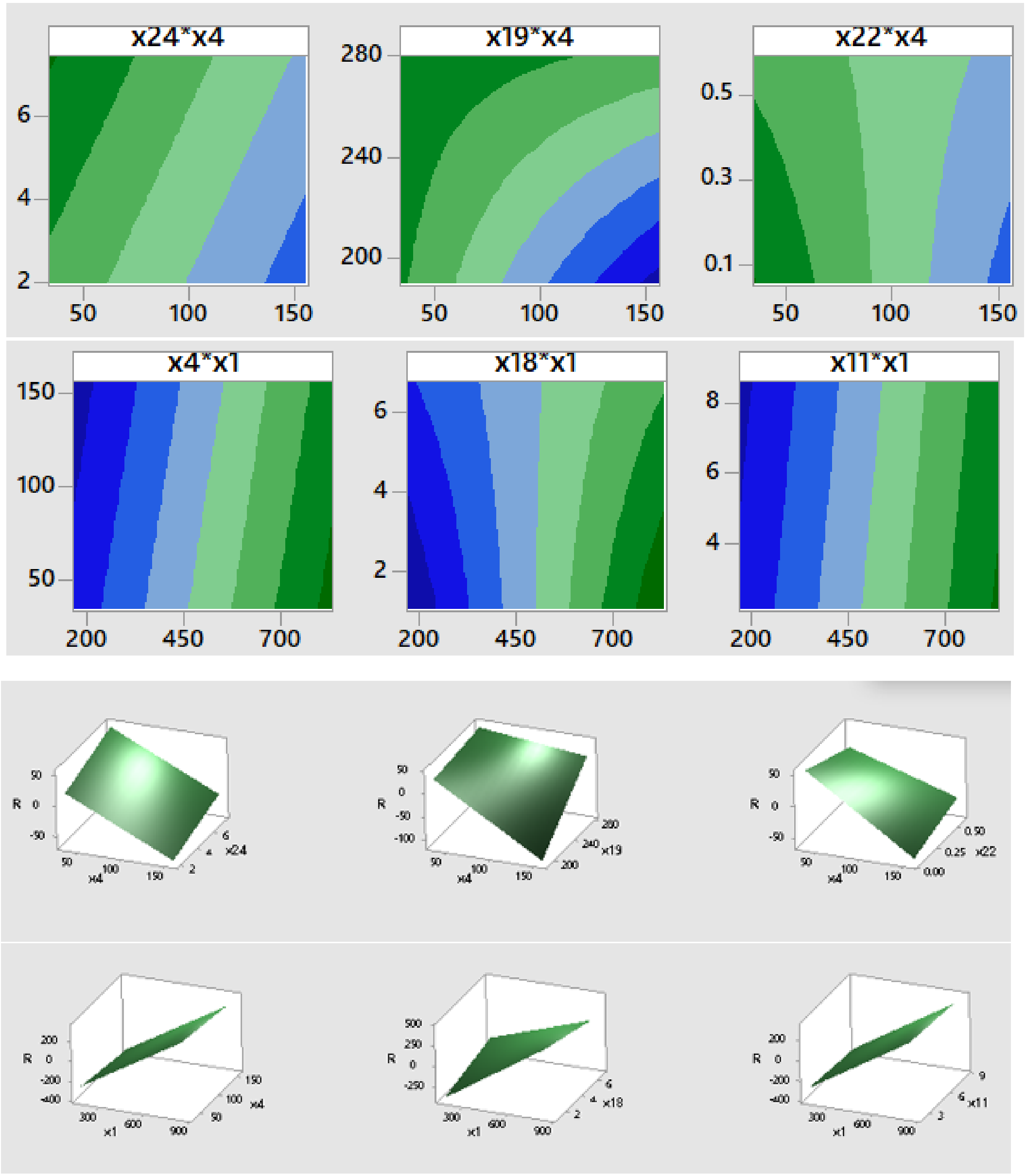

Figure A.1: Combination of Risk Factors Effect on the Returns

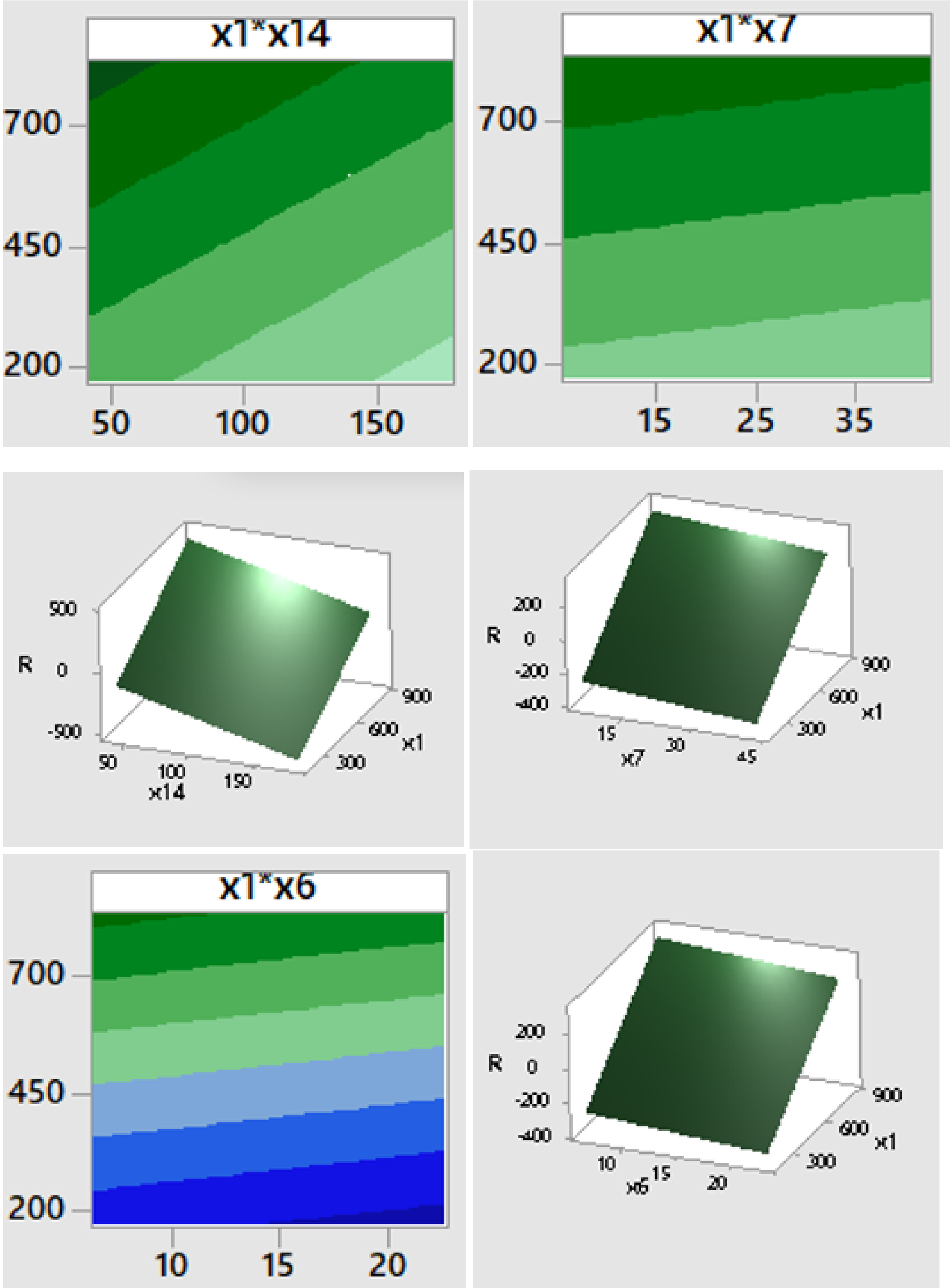

Figure A.2: Combination of Risk Factors Effect on the Returns