ASSESSING ESTIMATION OF REGRESSION TO MEAN FOR NON-NORMAL POPULATIONS THROUGH TRANSFORMATIONS

By

Hina Tariq

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

بسم الله الرحمن الرحيم

In the Name of Allah The Most Merciful and The Most Beneficent

ASSESSING ESTIMATION OF REGRESSION TO MEAN FOR NON-NORMAL POPULATIONS THROUGH TRANSFORMATION

By

Hina Tariq

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF PHILOSOPHY IN STATISTICS

Supervised By

Dr. Manzoor Khan

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2023

Declaration

I “Hina Tariq” hereby solemnly declare that this thesis titled, “ Assessing estimation of regression to mean for the non-normal populations through transformation”.

- This work was done wholly in candidature for a degree of M.Phil Statistics at this University.
- Where I got help from the published work of others, this is always clearly stated.
- Where I have quoted from the work of others, the source is always mentioned. Except for such quotations, this thesis is entirely my research work.
- Where the thesis is based on work done by myself jointly with my supervisor, I have made clear exactly what was done by others and what I have suggested

Dated:__________________________ Signature:__________________________

Acknowledgement

Boundless praise and gratitude to Allah Almighty, who is the most gracious, most Benevolent, and kind, peace and blessings of Allah be upon the Holy Prophet Hazrat Mohammad (S.A.W) and his pure and pious progeny, who are the source of knowledge and guidance for the entire world forever.

After this, I would like to express our heartfelt gratitude to all those people, especially my supervisor, **Dr. Manzoor Khan**, not only for his marvelous guidance and teaching but also for his scientific discussion that gave me the possibility to complete my thesis. It has been a great honor for me to work with him. I am deeply indebted to my supervisor whose help, stimulating suggestions, knowledge, patience, experience, and encouragement helped me throughout the study and analysis of the thesis in the pre and post-research period. The practical guidance in the field of regression that the name gave is very valuable to me. I feel highly privileged to take this opportunity to express my heartiest gratitude and a deep sense of indebtedness to the Honorable Supervisor **Dr. Manzoor Khan**. Also, thanks to all the department teachers **Dr. Ijaz Hussain**, **Dr. Abdul Haq**, **Dr. Sajid**, **Dr. Ismail**, and **Mam Maryam** for guiding me in all my research work.

I extend my heartfelt gratitude to my beloved parents, **Mr. and Mrs. Tariq Firdous**, whose hands have remained lifted in prayer for me always. May ALLAH bestow His blessings upon them abundantly. I also wish to express warm appreciation to my sisters – **Nida, Kiran, Komal, and Mehak** – for their unwavering support. A special note of thanks goes to my best friend and brother, **Muhammad Usman**, for standing by my side.

Additionally, I am deeply grateful to my closest companions, **Aaina and Rabyia**. Last but not least, my sincere thanks go out to all my classmates, juniors, seniors, siblings, and friends, whose unwavering support has been a constant encouragement throughout my journey to complete my thesis. **(Ameen)**

Dedication

*I dedicated this effort of mine to my Parents for raising me to believe that anything is possible and to myself for making everything possible along with all the respected teachers and my all sisters, especially **Mehak Tariq**, and brothers for their love, encouragement, support, and affection that make me able to get this success.*

Contents

List of Tables

List of Figures

Abstract

Regression to the mean is the phenomenon in which extreme values of a variable in one sample are discovered to be less extreme and more similar to the population mean following re-measurement. The first measurement is exceptional solely by accident. When individuals are chosen based on certain criteria or a cutoff point to determine the significance of an intervention effect, the observed change, known as the total effect, is made up of treatment and RTM effects. RTM must be taken into consideration in order to appropriately quantify the intervention effect. In this research, we aim to develop novel methods for estimating RTM effects in non-normal populations, transforming non-normal data into normal using Box-Cox transformations and bivariate normal data methods. The inverse transformation of RTM is then applied to bring it to the original units of the data. The second approach is motivated by the percentile change caused by the RTM, which is defined as the difference between the first quantile point and the mean of the data on the second occasion. The same amount of percentile change is used to identify the two quantile points whose difference is the RTM effect in that original population. We used several distributions to meet our research goals and evaluate the effectiveness of our proposed methods.

Chapter 1

Introduction

Regression toward the mean. That is, in any series of random events an extraordinary event is most likely to be followed, purely due to chance, by a more ordinary one.

– Leonard Mlodinow

1.1 Regression to Mean

Regression to the mean (RTM) is a statistical phenomenon that can make natural variation in repeated data appear to be due to treatment or intervention effects. It unusually happens when large or small measurements tend to be followed by measurements that are closer to the true population mean. This is also called “reversion to mediocrity” or “reversion to the mean”. It is a well-known phenomenon in several domains, including medicine, psychology, sports, economics, and education. Galton (1886) was the first to establish the idea of RTM at the end of the nineteenth century. Galton noticed that the heights of the offspring of tall parents were closer to the population average than those of their parents. The heights of children of short parents tended to be closer to the average population than those of their parents. To delve deeper into the evolution of the RTM concept over time, the readers are referred to Stigler (1997).

Galton coined the phrase ”RTM” to understand this phenomenon better. Some examples of RTM are below.

Assume a researcher wants to see how a new trading method affects stock returns. The researcher picks a set of stocks with a history of high volatility and trading volume. The stock returns are higher after using the new trading method. The efficacy of the technique is predicted to rise, but other factors may also be at play. Stock return variations can be influenced by various sources of variation, including market circumstances, investor state of mind, and random fluctuations. Probably, the rise in stock returns observed after applying the new trading technique could be influenced by variables other than the strategy's effect.

For example, the initial baseline measures might have been influenced by ideal market conditions or a strong investor mindset, resulting in better stock returns even before the new strategy was implemented. In this instance, the reported boost in stock returns may not be attributable to the treatment effect of the new trading method but may instead be partially explained by random market movements or outside causes. In this case, the idea of RTM also applies. If the selected stocks have historically poor returns, it is feasible that their returns would organically grow over time due to RTM even without any intervention. As a result, any reported rise in stock returns following the adoption of the new trading method should be carefully examined to establish the true impact of the strategy, taking into account the role of RTM and other complicating variables.

Investors frequently examine stock performance to make investing decisions. Assume ABC is a private trading company. ABC's stock price has increased by 40 percent in the last month as a result of the favorable market state of mind and high demand. Investors have taken notice of the unbelievable increase in value over such a short period.

However, based on the concept of RTM, investors should consider that ABC's future returns are likely to be more moderate compared to the recent exceptional growth. The stock's price may not continue to rise at the same rapid pace in the coming weeks or months. Regression to the mean suggests that the stock's performance will regress or move closer to its average or typical rate of growth.

Similarly, WXY's stock price has lately dropped significantly. Despite the stock's significant decrease, due to RTM, the WXY's future returns may improve. Over time, the stock's price may recover or stabilize, bringing it closer to its average or projected performance. By taking RTM into account, investors can reduce their expectations and make more rational decisions, knowing that significant price fluctuations in the stock market are likely to return to average or mean performance over time.

James (1973) investigated the effect of RTM in uncontrolled clinical trials, and concluded that regression effects were present due to biological variation over time and variation in the measurement. The author derived a formula to quantify and estimate the RTM effect. He argued to separate RTM and treatment effects in uncontrolled clinical trials. Chinn and Heller (1981) derived algebraic formulae to calculate the RTM effect for a measurement whose population mean and variance change over time and for sub-populations selected according to the initial value. The formulae were applied to the plasma cholesterol levels of participants in a dietary intervention study, enabling the effects of an intervention to be separated from the secular change and the RTM effect.

Healy and Goldstein (1978) reviewed the concept of RTM. The author described the concept of RTM in basic terms and illustrated how it appeared in studies of mental and physical development. Browne et al. (1999) investigated the effect of RTM on neurological symptoms following a coronary artery bypass graft procedure. At the discharge evaluation, the group's mean performance on the Rey Auditory Verbal Learning Test had decreased and remained below baseline after three months. The average performance of the Trail Making Test (Part A) showed a practically significant decrease upon discharge before increasing at three months. RTM had a considerable influence on single-case criteria of cognitive impairment with a disproportionately large percentage of high-baseline performers classified as handicapped.

Barnett et al. (2005) explained the RTM concept and demonstrated methods to deal with it. The authors thought that the RTM effect could be reduced by enhancing the design of research and employing suitable statistical methods. Bush et al. (2006) explored the prevalence of RTM in contrast to the efficiency of capital markets. By utilizing Fortune Magazine's ranking of highly regarded American companies to differentiate between positive and negative firms, the authors showcased the findings that: (1) the least admired companies outperformed the most admired firms, (2) a portfolio comprising the most admired firms demonstrated better performance than the market, and (3) a portfolio of the least admired companies also exhibited superior performance compared to the market.

1.2 Consequences of Overlooking RTM

The results of data analysis and decision-making could be misleading if RTM is ignored. To avoid determining wrong conclusions and creating inaccurate projections, it is essential to comprehend and resolve these statistical phenomena. Here are a few important consequences:

- **False Attributions**: Ignoring RTM might lead to incorrectly associating an effect with specific causes or treatments, thereby obscuring the real nature of the data.
- **Misleading projections**: Ignoring RTM might lead to overly optimistic or pessimistic projections, as extreme values tend to regress toward the mean later upon remeasurement.
- **Incorrect intervention evaluation**: Ignoring RTM in intervention evaluation might bias results since extreme pre-intervention ratings may regress toward the mean.
- **Unrealistic Expectations**: Failure to take RTM into account might result in inflated expectations which could result in bad decisions or strategies since dramatic successes or failures cannot be long-lasting.

Recently, researchers have reported the RTM effect in different research areas such as health (Moore et al., 2019; Wang et al., 2020; Cochrane et al., 2020; Kypri, 2020), measurements of geographic atrophy growth rate (Biarnés and Monés, 2020), and economic forecasting (Pritchett and Summers, 2014).

1.2.1 Methods for RTM Quantification

Many researchers have developed methods for quantifying RTM. Most of the method are based on the assumptions of bivariate normality of the pre-post variables. Formulae were derived for quantifying and estimating RTM by James (1973), Gardner and Heady (1973) , Davis (1976), and Johnson and George (1991). Recently, researchers have worked out estimating the RTM effect for discrete distributions including the Poisson distribution (Khan and Olivier, 2018) and the bivariate binomial distribution (Khan and Olivier, 2019).

Beath and Dobson (1991) used Saddle point and Edgeworth approximations to estimate RTM for non-normal populations. The methods had certain limitations like producing negative results for probability mass function and becoming complicated. Similarly Müller et al. (2003) developed non-parametric technique for quantifying RTM. However, their formulae does not help in decomposing the total effect into RTM and treatment effect. John and Jawad (2010) introduced a kernel density based methods for estimating the RTM effect.

1.2.2 Identification of RTM through graphs

The RTM impact may be displayed using a simple scatter plot that compares follow-up and baseline sales measurements. In Figure 1.1, the x-axis represents baseline sale measurement, while the y-axis represents the difference between follow-up and baseline sales. The solid line shows that the follow-up and baseline values are completely in line (i.e., there has been no change). The higher line represents the treatment group, while the gap between the regression lines denotes a potential treatment effect. The higher line represents the treatment group, and the space between the regression lines denotes a potential treatment effect. The dotted lines were created by performing a linear regression of the change values on baseline values with a group covariate. In the figure, some RTM is visible because subjects with unusually low baseline results have been more likely to improve (so that change values are probably above the solid line), and subjects with unusually high baseline results have tended to decline (so that change values are probably below the solid line). Since there was less variation in the group mean between the measurement intervals in the placebo group, this pattern is more visible in that group.

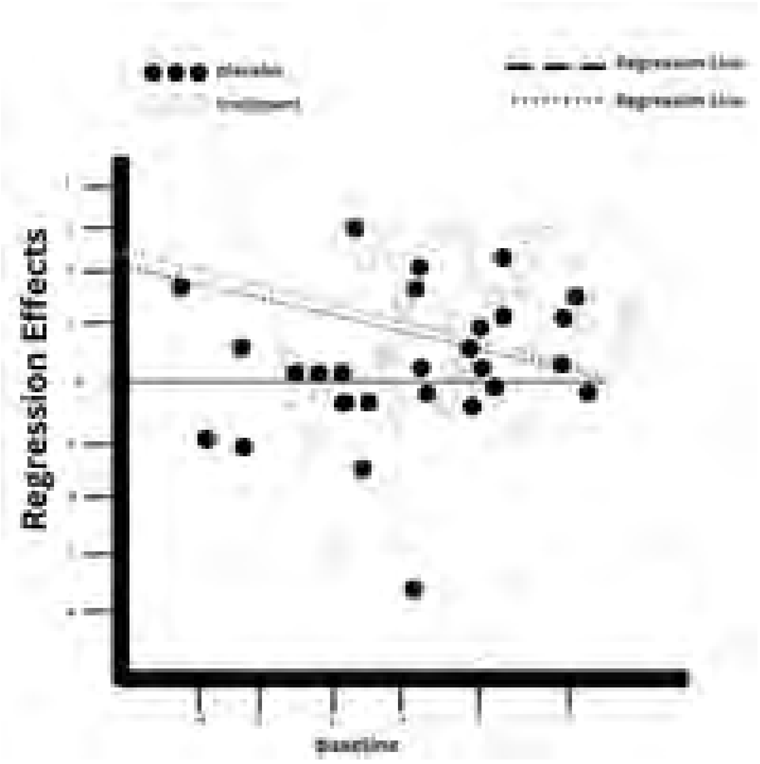

Figure 1.1: RTM effects in figure sales in baseline and follow-up measurement with true mean and variation

1.2.3 Reducing the effect of RTM

Yudkin and Stratton (1996) discussed several approaches that could be used in experiment design to mitigate the RTM effect. The RTM can be reduced by using a research design that randomly allocates participants to the control and active groups. This will have an equal impact on both groups' responses. The estimated treatment effect is the difference between the treatment and placebo groups after RTM correction. The RTM plus placebo effect is responsible for the change in the placebo group. James (1973) highlighted the importance of the control group in decreasing the RTM effect.

Gardner and Heady (1973) studied a second technique for reducing the RTM effect and recommended that people be selected using two or more baseline assessments and that the number of follow-up measurements be increased. The RTM effect is caused by the random component, and it is highly rare to detect extraordinary events (very good or very bad) on the second measurement.

Additionally, the RTM might be decreased during the data analysis stage when the cut-off point has been chosen; hence, it is suggested to use ANCOVA with a group co-variate.

1.3 Box Cox Transformation

Box and Cox (1964) introduced Box-Cox transformation. It is a statistical method called the Box-Cox transformation that can be used to reduce the variance of a dataset or bring the data closer to the assumption of normality. When data violate the assumption of normality (i.e., are not normally distributed) or show heteroscedasticity (unequal variances across several groups or levels of a variable), the Box-Cox transformation might be used. The transformation can be applied to continuous positive-valued data.

To transform datasets with non-normal data into a normal, Sakia (1992a) suggested a statistical technique known as Box-Cox transformation. This change may significantly increase the accuracy and dependability of linear regression modeling. It required you to take the natural logarithm of a variable and improve it to power (lambda) measured by MLE. The lambda value will be determined by how skewed the data is, which means that a new lambda will be used for each data set. This transformation can be used in regression, ANOVA, and a variety of other applications where non-normal data must be transformed into normal form.

There are many reasons to use the Box-Cox transformation, but here are the three most important reasons to use it: To stabilize the variance, improve normality, and make patterns in the data more easily recognizable. The most fundamental goal of the Box-Cox transformation is to choose an appropriate power transformation that maximizes data normality or equalizes variances between groups. A parameter called lambda λ defines the transformation by specifying the kind and degree of the transformation that will be applied to the data. The Box-Cox transformation is defined as follows:

$$w(\lambda) = \begin{cases} \frac{w^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \ln(w), & \lambda = 0 \end{cases} \tag{1.1}$$

where w represent the original data, and w(λ) represent the transformed data.

The Box-Cox transformation often involves assessing the probability function for several lambda values and selecting the lambda value that optimizes the likelihood. You may do this by applying several optimization techniques. The Box-Cox transformation's purpose is to generate a dataset that is improved in terms of symmetry and variance homogeneity, making the assumption of statistical analysis valid. However, the interpretation of the revised information may differ from the original data.

Although additional transformations, such as the inverse, square root, and logarithmic ones, are available, the Box-Cox transformation provides a versatile approach that combines all of them and may determine which one is appropriate for the data under consideration.

1.4 Motivation of the study

Much of the past research on RTM has focused on normal populations. This research, however, focuses on assessing RTM for non-normal populations. The objective of this study is to develop novel methods for estimating RTM effects in data that don't follow a normal distribution. The aim is to transform the non-normal data using the Box-Cox transformations and using the methods developed for bivariate normal data to quantify RTM. After quantification, take the inverse transformation of RTM to bring it to the original units of the data.

The second approach is motivated by the percentile change caused by the RTM. RTM can be defined in terms of percentile change as the difference between the first quantile point (usually the mean of the truncated data) and the mean of the data on the second occasion. After identifying the percentile change, the same amount of percentile change is used for identifying the two quantile points whose difference is the RTM effect in that original population.

1.5 Research Objectives

- In the case of RTM, Convert non-normal data to normal data using the Box-Cox transformation, then transform the RTM using the Inverse Box-Cox transformation.
- Study the effect of the percentage change.

1.6 Outlines of the thesis

In this thesis, we worked to assess the estimation of RTM for a non-normal population. The structure of the thesis consists of four chapters. In Chapter 2 the existing methods and relevant literature on RTM for normal distribution and Box-Cox are discussed. Chapter 3 contains an assessment of RTM through Box-Cox Transformation, a simulation study, and a discussion. Chapter 4 contains an assessment of RTM for discrete distribution, a simulation study and a discussion. Chapter 5 contain the conclusion of the work and recommendation for future work.

Chapter 2

Literature

Several techniques have been put forth in past research to quantify the impact of RTM in various situations. Researchers have offered their suggestions for how to decrease the impact of RTM and assess treatment effects. We provide a quick summary of all of these strategies and techniques in this chapter.

2.1 RTM Effect Under Normal Distribution

James (1973) discovered the RTM effect for bivariate normal distribution. He assumed that the observed variable is made up of both true and random error components. Let X_i is the effect size on i^{th} measurement of the same subject and X_0 is the true measurement then:

$$X_i = X_0 + e_i$$

where e_i is random error and $i = 1, 2, ...$

James (1973) discovered that the observed value was composed of two components namely the biological effect and measurement error and the true value. James (1973) suggested that RTM be separated from the observed change to accurately estimate the true treatment effect to avoid incorrect conclusions.

James derived the formula of RTM by assuming that the pre and post-variables $X_i \sim N(\mu, \sigma^2)$ with $cov(X_1, X_2) = \sigma_0^2$ for $i = 1, 2$. Considering the null treatment effect, the RTM effect equals the condition difference between the pre and post-treatment means. The resulting RTM effect formula is

$$R\left(x_0\right) = \frac{\sigma(1-\rho)\phi\left(z_0\right)}{1-\Phi\left(z_0\right)} = (\frac{\sigma_e^2}{\sqrt{\sigma_0^2+\sigma_e^2}}) * (\frac{\phi\left(z_0\right)}{1-\Phi\left(z_0\right)}). \tag{2.1}$$

James determined that if the treatment was effective, the model connecting the pre-post variables could be expressed as

$$X_2 - \mu = \gamma\rho\left(X_1-\mu\right) + e_1. \tag{2.2}$$

For $\gamma < 1$, the treatment effect is non-zero. The observed conditional difference in a bivariate normal distribution was shown to be

$$E(Z_1 - Z_2 \mid Z_1 > z_0) = \frac{(1-\gamma\rho)\phi(z_0)}{1-\Phi(z_o)}, \tag{2.3}$$

where Z_i is z-score for $i = 1, 2$.

James established the above formula in equation 2.3 for the overall proportionate decrease attributable to RTM and treatment as well. This will partition the total effect into true and RTM effects however it fails when the pre-post measurements variables are independent.

Similar to James, Gardner and Heady (1973) worked on the derivation of the RTM effect, but along with bivariate measures, the authors also investigated the effect of multiple measurements on RTM. The authors assumed the normal distribution of pre-post variables with $rho = \sigma_{/}\sqrt{\sigma_0^2+\sigma_1^2}$. The subjects selected based on the right cut-off point, i.e. $X_i > x_o$ follow the univariate truncated normal distribution with mean;

$$E\left(X_i \mid X_i > x_0\right) = \mu + \sigma\frac{\phi\left(z_0\right)}{1-\Phi\left(z_0\right)}. \tag{2.4}$$

Similarly, the mean of X_0 given that the observation is greater than the cutoff point is

$$E\left(X_0 \mid X_i > x_0\right) = \mu + \frac{\sigma_0^2}{\sigma} \cdot \frac{\phi\left(z_0\right)}{1-\Phi\left(z_0\right)}. \tag{2.5}$$

Since $\sigma > \sigma_0^2/\sigma$ unless $\sigma_e^2 = 0$, therefore it is clear from the above equations 2.4 and 2.5 that the observed mean of observation is always greater than their true mean due to RTM.

Gardner and Heady (1973) derived the RTM formula for multiple measurement n on the same subject before application of the treatment and is given by

$$R\left(x_0\right) = E\left(\bar{X} - X_0 \mid \bar{X} > x_0\right) = \frac{\sigma_e^2/n}{\sqrt{\sigma_0^2 + \sigma_e^2/n}} \cdot \frac{\phi\left(z_{0n}\right)}{1 - \Phi\left(z_{0n}\right)}, \tag{2.6}$$

where $\bar{X} = \sum x_i/n$ is the sample mean of n multiple measurements. James' derivation of RTM is a special case of equation 2.6 for $n = 1$. However, the RTM effect approaches zero when n becomes sufficiently large.

Davis (1976) worked on the study design to reduce the RTM effect. Let the mean of all multiple measurements and a follow-up observation be $\bar{X}$ and X^*, respectively such that $\bar{X} \sim N(\mu, \sigma_o^2 + \sigma_e^2/n)$ and $X^* \sim N(\mu, \sigma_o^2 + \sigma_e^2)$. The RTM effect is derived as shown below:

$$R\left(x_0, n\right) = E\left(\bar{X} - X^* \mid \bar{X} > x_0\right) = \frac{\sigma_e^2/n}{\sqrt{\sigma_0^2 + \sigma_e^2/n}} \cdot \frac{\phi\left(z_{0n}\right)}{1 - \Phi\left(z_{0n}\right)} \tag{2.7}$$

which is the same as derived by Gardner and Heady (1973). Using the first observation X_1 as a classification baseline measurement, i.e., choosing a subject based on the event $X_1 > x_0$, and the second observation X_2 on the same subject as the baseline from which the treatment effect may be assessed could be useful to mitigate the RTM effect (Davis, 1976). Let X_3 be the post-treatment measurement, then the author derived the RTM formula by taking the conditional expectation of the truncated bivariate distribution and derived the formula

$$R\left(x_0, \rho_{12}, \rho_{13}\right) = E\left(X_2 - X_3 \mid X_1 > x_0\right) = \left(\rho_{12} - \rho_{13}\right) \cdot \sigma \frac{\phi\left(z_0\right)}{1 - \Phi\left(z_0\right)}, \tag{2.8}$$

where the correlation coefficient between (X_1, X_i) are represented by ρ_{1j} for $j = 2, 3$. The RTM effect becomes zero when the two correlation coefficients are equal with baseline measurement, thereby does not require multiple measurements for reducing the RTM effect.

So far, the observed values were assumed to have consisted of two components, measurement error and the second is biological variables such as emotional and other influences during the recording of observation. Johnson and George (1991) extend the previous model and included the subject effect, $S_i \sim N(0, \sigma_s^2)$. Hence the model becomes;

$$Y_{ij} = X_0 + S_i + E_{ij} \tag{2.9}$$

where $i = 1, 2, ...m$, $j = 1, 2, ...n$ and Y_ij represents the j^{th} replicate measurement at the i^{th} study time. The correlation between S_i and S_k is positive and independent of random error E_i and baseline measurement X_0. Under Equation.2.9 the RTM formula derived by Johnson and George (1991) is

$$R_T\left(y_0\right) = \frac{\left(1-\rho_s\right)\sigma_s^2 + \sigma_e^2/n}{m\sigma_{\bar{y}}} \cdot \frac{\phi\left(z_1\right)}{1-\Phi\left(z_1\right)} \tag{2.10}$$

The above equation 2.10 represents the total RTM effect due to measurement error and subject effect. The measurement error of RTM can be reduced by either number of replication n or by increasing the number of repeated measurements. While a larger number of repeated measurements m taken at various times reduces the regression impact caused by subject variability.

The detailed work on RTM under normal distribution was recently done by Khan and Olivier (2022). The authors partitioned the total effect into true treatment and RTM effects found the MLE and checked their properties such as unbiasedness, consistency, and normality. The RTM effect was depicted for both positive and negative correlations. They derived the RTM effect in a pre-post measurement case in which the pre-variable is composed of true and random parts i.e. $X_1 = X_0 + \epsilon_1$ and Post-variable $X_2 = a + bX_0 + \epsilon_2$, where $a + bX_0$ is the true part and ϵ_2 is the random component. The total effect is quantified as follows;

$$T(x_0, \theta) = (\mu_1 - \mu_2) + (\sigma_1 - \rho\sigma_2)\frac{\phi\left(z\right)}{1-\Phi\left(z\right)}. \tag{2.11}$$

The first part on the right-hand side, $(\mu_1 - \mu_2)$, of the above equation is the average treatment effect, and the second term is the RTM effect. The authors also derived variance of RTM var$(X_1 - X_2 \mid X_1 > X_0)$ as

$$\text{var}\left(X_1 - X_2 \mid X_1 > x_0\right) = \sum_{i=1}^{2} \text{var}\left(X_i \mid X_1 > x_0\right) - 2 \times \text{cov}\left(X_1, X_2 \mid X_1 > x_0\right). \tag{2.12}$$

The maximum likelihood estimators of the total, RTM, and true effect were derived as

$$\hat{T}_r\left(x_0, \boldsymbol{x}\right) = \hat{\mu}_1 - \hat{\mu}_2 + \frac{\phi\left(\hat{z}_0\right)}{1 - \Phi\left(\hat{z}_0\right)} \cdot \left(\hat{\sigma}_1 - \hat{\rho}\hat{\sigma}_2\right), \tag{2.13}$$

$$\hat{R}_r\left(x_0; \boldsymbol{x}\right) = \left(\hat{\sigma}_1 - \hat{\rho}\hat{\sigma}_2\right) \cdot \frac{\phi\left(\hat{z}_0\right)}{1 - \Phi\left(\hat{z}_0\right)}, \text{ and} \tag{2.14}$$

$$\hat{\delta}(\boldsymbol{x}) = \hat{\mu}_1 - \hat{\mu}_2. \tag{2.15}$$

The distribution of RTM and true treatment $\hat{\delta}(x)$ were shown to be asymptotically normal and unbiasedness and consistency of the estimators were established. The simulation study shows that the RTM and intervention estimates are close to the true value in all cases while James (1973) method gave poor estimates.

2.2 Box-Cox Transformation

Box and Cox (1964) introduced Box-Cox transformation. It is a statistical method called the Box-Cox transformation that can be used to stabilize the variance of a data set or make the data satisfy the assumptions of normality. When data violate the assumption of normality (i.e., are not normally distributed) or show heteroscedasticity (unequal variances across several groups or levels of a variable), the Box-Cox transformation could be utilized. The transformation can be applied to continuous and positive-valued data. Sakia (1992b) proposed a parametric power transformation technique to reduce errors such as non-additive, non-normality, and heteroscedasticity.

In Daimon (2011) studied the Box-Cox power transformation. It is used to change the distributional shape of a piece of data to make it more normally distributed so that tests and confidence limits that need normality can be applied properly. Yang (1996) conducted a study on the Box-Cox's conditional method to inference after transformation selection demonstrates that the T-statistic obtained when the transformation is estimated using Box-Cox is asymptotically similar to that obtained when the transformation is assumed to be known. Rahman (1999) proposed a new method for estimating the Box-Cox transformation using the maximization of the Shapiro-Wilk W statistic, which forces the data to get closer to normal as much as possible. A comparative study is also presented with the normal-based likelihood and artificial regression model procedures.

Tommaso and Helmut (2011) investigated whether transforming a time series leads to an improvement in forecasting accuracy. The authors proposed a nonparametric approach for estimating the optimal transformation parameter and conducted an extensive recursive forecast experiment on a large set of seasonal macroeconomic time series. As the forecast horizon increased, the evidence in favor of a transformation became less strong.

Osborne (2010) studied Box-Cox as an automated procedure that may be used in SPSS and SAS, and it may be a best practice for normalizing data or stabilizing the variance.

Sarkar (1985) proposed the maximum likelihood method of estimation for the parameters of the Box-Cox model showing the seriousness of the problem of heteroscedasticity in the context of this transformation. The authors also suggested how to separate the problem of non-linearity from the influence of stabilization of error variance in an estimate of the transformation parameter.

Hossain (2011) reviewed the role of the Box-Cox transformation technique in estimation, testing, inference, and model selection. An attempt was made to bridge the gap by providing an analytical bibliography for model selection. Gaudry and Laferriere (1988) studied Box-Cox transformations on linear regression models, which could be interpreted as simple power transformations, leading to non-degenerate solutions when estimated without a regression constant.

Proietti and Lütkepohl (2013) examined if a time series transformation might increase forecasting precision. They conducted a comprehensive recursive forecast experiment and suggested a nonparametric method for estimating the ideal transformation value. The results demonstrated that at the one-step-ahead horizon, the Box-Cox transformation generated superior forecasts than the untransformed data. Another study conducted byOzgur Asar and Dag (2014) used seven goodness-of-fit tests and a search algorithm to estimate this Box-Cox parameter. Simulation studies showed that Shapiro–Wilk and the artificial covariate method were more effective than Pearson Chi-square. An R package called AID was proposed for implementation.

Chapter 3

Assessing RTM Through Transformation

3.1 Estimation of RTM for Non-Normal Populations

RTM is a phenomenon in statistical analysis that refers to the tendency of extreme observations in a data set to move closer to the population mean upon repeated assessments. This effect is popular in natural systems and is well-known when working with data with a normal distribution. But in real life, when we deal with population datasets, sometimes the data do not follow a bivariate normal distribution; that data set is called non-normal data. The assumption of normality is violated. In these situations, specialized statistical techniques are needed to account for the non-normality of the data. There are numerous strategies and procedures for transforming non-normal data into a normal distribution. The Box-Cox transformation within the framework of RTM is the strategy we used to tackle this problem. It is interesting to note that the Box-Cox transformation has not yet been used in RTM to address this specific issue.

3.2 Box-Cox Transformation Application

The Box-Cox transformation is a statistical method for converting non-normal data to normal data. It can improve the accuracy of predictions made using linear regression. In this thesis, we use this technique to overcome the problem of non-normal data to normal data while estimating RTM. Box and Cox (1964) suggested a parametric power transformation approach to eliminate non-normality and heteroscedasticity. The Box-Cox transformation is a statistical approach commonly used for data normalization and variance stabilization in various applications. Its applications include data analysis, regression modeling, time series forecasting, machine learning, and many more.

For the estimation of RTM from non-normal data, we are using this transformation to tackle this problem and see how it would work in the estimation of RTM. In RTM estimation, the Box-Cox transformation is often used to solve the problem of non-normal data with accurate power transformation of lambda value. Using the Box-Cox transformation, the assumption of normality in the data set could be satisfied and the methods developed under the assumption of normality could be used to estimate RTM by selecting an accurate value of lambda using different parameters of different distributions. With the help of it, we draw a better and more accurate conclusion by accounting for RTM. The Box-Cox transformation is defined as follows:

$$w(\lambda) = \begin{cases} \frac{w^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0 \\ \ln(w), & \text{if } \lambda = 0 \end{cases}$$

The application of Box-Cox transformation is used in many other fields such as time series, forecasting, machine learning, and data analysis. In time series Box-Cox transformation plays a very significant role, especially when we deal with the real data set in the real world when the assumption of normality and non-constant variance are violated. Time series mostly contain more complex patterns like seasonality, trends, and irregularity, which makes the variance non-stationary over time. By applying the Box-Cox transformation with an accurate value of the parameter lambda, the variance can be reduced, making the data more suitable for modeling and analysis with approaches such as ARIMA or SARIMA. It can solve the non-normality issue. In conclusion, the Box-Cox transformation is a vital pre-processing step in time series analysis, helping to improve data quality and the performance of time series models.

The Box-Cox transformation is an important pre-processing technique that improves model performance when dealing with skewed or non-normal data. Batter pattern detection and resilience are made possible by normalizing data, stabilizing variance, and ensuring normality. It also reduces the impact of outliers on model performance. Overall, it is an important tool in the machine-learning process because it improves model generalization and prediction accuracy with non-normal or skewed data.

In conclusion, BoxCox transformation is valuable in transforming non-normal data into normal data in many applications. It improves the quality and reliability of statistical studies by normalizing the data, resulting in more robust and useful insights for decision-making and predictive modeling.

3.3 Idea of Percentile Change

In the context of RTM, first the subjects with initial measurements x_{1i} for $i = 1, 2, \cdots, n$ greater than a cut-off point say x_0 are selected. Upon re-measurements x_{2i}, the mean $\overline{x}_2$ of the subjects is found closer to the mean. The difference between the cumulative probabilities at point $\overline{x}_1$ and the new mean $\overline{x}_2$ is defined as the probability of percentile change. Mathematically,

$$PC = Pr(x < \overline{x}_1) - Pr(x < \overline{x}_2). \tag{3.1}$$

From Figure 3.1, the difference of the two quantile points is the RTM effect as

$$RTM = \overline{x}_1 - \overline{x}_2. \tag{3.2}$$

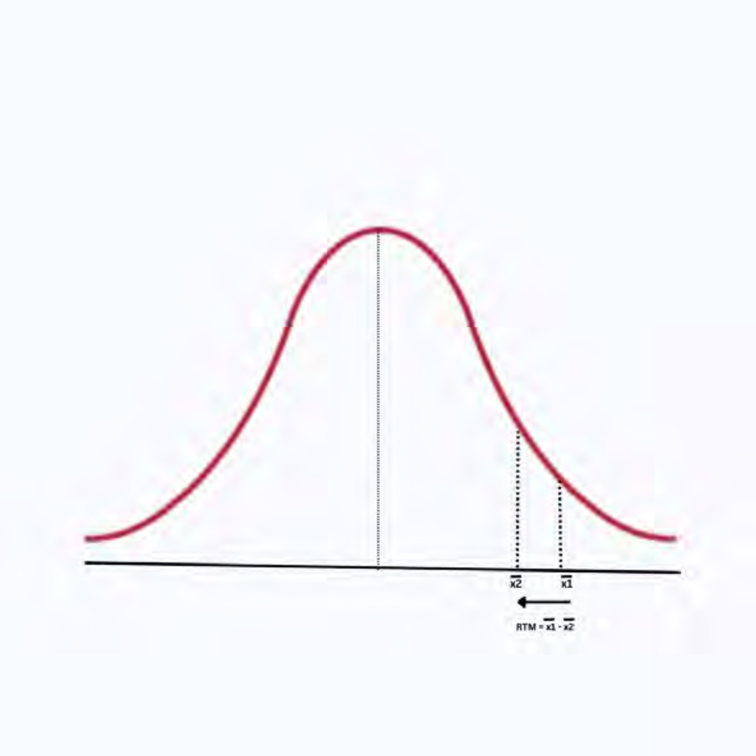

Figure 3.1: Visual display of the RTM effect after Box-Cox transformation

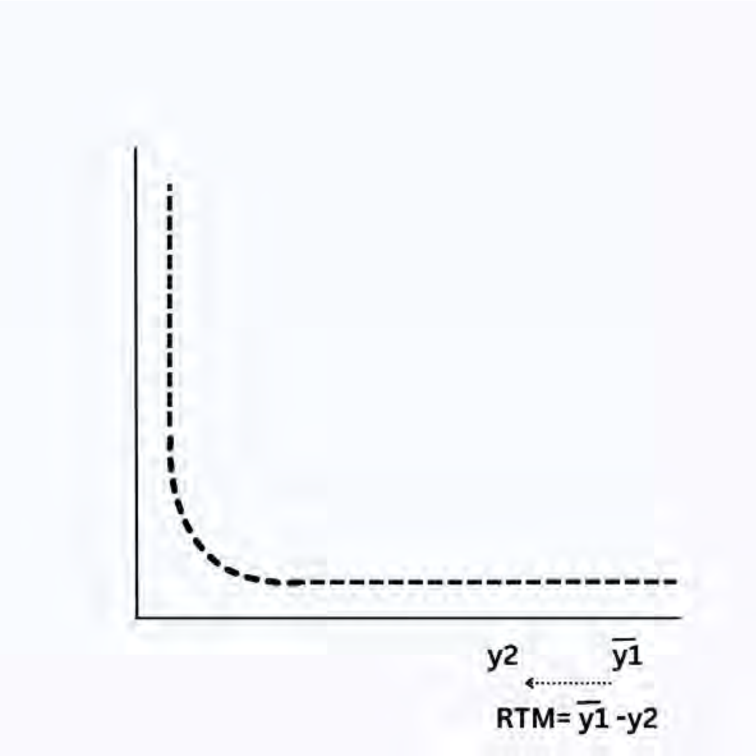

Figure 3.2: Visual display of the RTM in the original non-normal distribution using the percentile change

Let's assume data from a non-normal distribution x_i such that $x_i > 0$ for $i = 1, 2..., n$. After applying the Box-Cox transformation, the non-normal data will become normal and let it be denoted by $y_1, y_2, ..., y_n$. The position of each datum is the same, however, the values of the data will change after the application of the Box-Cox transformation technique. After applying the Box-Cox technique we used the normality check using the Shapiro-Wilk Test. The method developed under the assumption of normality could be used to quantify the RTM effect, and the relevant percentile change can be determined. The same percentage of relevant change could be used to identify the quantile points for the non-normal data. The difference in the non-normal population's quantile points gives the RTM effect. Mathematically,

$$PC = Pr(y < \overline{y_1}) - Pr(y < y_2). \tag{3.3}$$

Substituting the PC in the above equation and using the known value of $\overline{y}_1$, the value y_2 can be determined. The difference between $\overline{y_1}$ and y_2 gives the RTM effect for the non-normal distribution as

$$RTM_{non-normal} = \overline{y}_1 - y_2. \tag{3.4}$$

Visually, the RTM is depicted in Figure 3.2.

To find RTM using the proposed method, first, we will transform the data from a non-normal distribution to a normal distribution by Box-Cox transformation, and then, with the help of the quantile points, we will find the RTM.

3.4 Inverse Box-Cox Transformation

Firstly, we proposed RTM by using Box-Cox transformations to convert the non-normal data to normal, and then we also used different quantile points. Now we are interested in finding the original data. With the help of inverse transformation, we transform the data back to its original scale in RTM. Then we select the method of inverse Box-Cox transformation to fulfill our interest. Also, check if the inverse Box-Cox transformation will work in it or not.

The inverse Box-Cox transformation is a statistical technique to transform the data back to its original scale. It's a reverse transformation of the Box-Cox transformation. Inverse Box-Cox transformation is crucial to data analysis, regression modeling, and other areas where the statistical hypotheses of normality and constant variance are significant. The formula that we use in the inverse Box-Cox transformation is as follows:

$$w^{-1}(\lambda) = \begin{cases} (w\lambda + 1)^{\frac{1}{\lambda}}, & \text{if}\lambda \neq 0 \\ otherwise, & \text{if}\lambda = 0 \end{cases}$$

Now we apply the two proposed techniques to different distributions and see how closely is the RTM effect estimated by the two methods. If they work, how much good will it be on a different distribution, and if it does not, how bad will it be on a different distribution? For this purpose, we have selected continuous skewed, continuous symmetric, and discrete distributions.

3.4.1 Exponential Distribution

The exponential distribution is a continuous probability distribution that models the time between events in a Poisson process in which events occur randomly and independently over a given period. It has only one parameter called λ, which represents the rate of event occurrence. The range of this distribution is $y > 0$. The PDF of the exponential distribution is

$$f_y(y|\lambda) = \begin{cases} \lambda e^{-\lambda y}, & \text{ify}> 0 \\ 0, & \text{ify}<= 0 \end{cases}$$

Algorithm: To check the performance of the proposed methods, we used the RStudio software. Initially, Let X_{11}, X_{21}, X_{12}, X_{22},:::, X_{1n}, X_{2n} be an exponential bivariate random sample of size n from a truncated bivariate distribution, and let the corresponding observed values be $x_{11}, x_{21,x12}, x_{22}, \cdots, x_{1n}, x_{2n}$. James (1973), utilized the method of moments to calculate μ, ρ, and σ_2. The percentage of the population in the trimmed portion x_0 was supposed to be known. We generated 5000 random variables from an exponential distribution using distinct sample sizes of 50, 100, 200, 300, 500, and 700. These random variables, denoted as $x0$, $x1$, and $x2$, were generated with parameter values of λ set to 0.5, 0.2, and 0.2, respectively. Subsequently, $x1$ and $x2$ were correlated with $x0$ to create a bivariate exponential distribution. Upon generating the bivariate random variables, we employed the

MASS library to apply the Box-Cox transformation. This transformation aimed to convert the non-normal data into a normal distribution, using the λ values used in the initial random variable generation. The two new random variables have been constructed using the Box-Cox transformation with the characteristics of normal data. Now we also check, after applying Box-Cox, that the data approaches normal with the help of histograms and Shapiro-Wilk Test. Both confirmed that the data should be normally distributed. After this, we need a cut point to find the RTM. With the help of quantile 0.85 against the 85 percentiles, we select the truncated point, which is 12. A truncated point is a value that decides which subset of the data will be selected for an intervention. Then we find all the values that are used in True RTM using the existing formulas of regression to the mean: estimation and adjustment under the bivariate normal distribution by Khan and Olivier (2022).

Furthermore, we proceed with determining the proposed RTM percentile values by employing quantile points and applying them to any normally distributed data. In finding the proposed method1, we used another proposed method to convert the data into its original data, the method name is Inverse Box-Cox Transformation. With the help of this, we also want to check whether the RTM returns to its original state or not. Table 3.1 shows the different values of the estimated RTM for different sample sizes. ARTM represents the RTM estimated using the normal distribution for different sample sizes, ARTMT represents the true RTM values; ARTMC represents the estimated RTM using the quantile points; and ARTMTR represents the second proposed RTM values using the inverse Box-Cox transformation.

Table 3.1: Assessing RTM through a transformation using exponential distribution when λ_0 =0.5, λ_1 =0.2, λ_2=0.2, x_0=12

Sample Size	ARTM	ARTMC	ARTMT	ARTMTR
50	0.6051065	9.190113	8.850660	2.072915
100	0.5891113	9.402414	8.849798	1.938875
200	0.4905445	9.570617	8.843195	1.672131
300	0.443658	9.536601	8.832506	1.55624
500	0.4251441	9.422772	8.844107	1.513019
700	0.4145126	9.406074	8.837298	1.493079

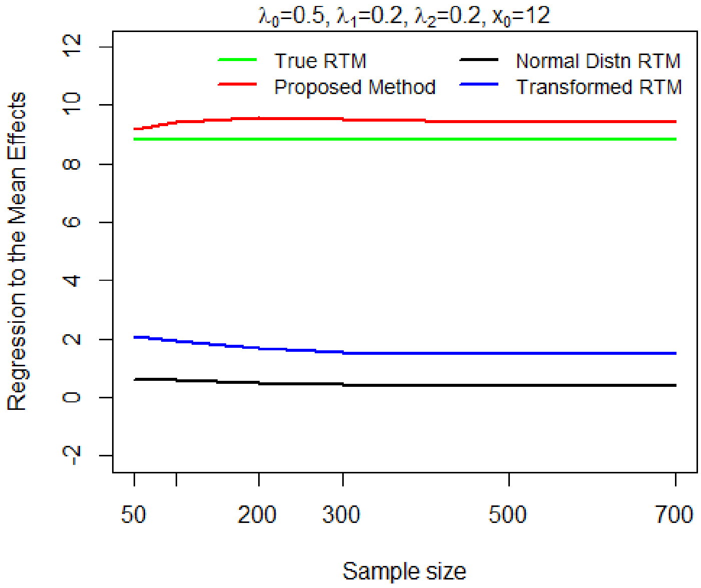

Figure 3.3: Assessing RTM through a transformation using exponential distribution

The above graph presents the behavior of the RTM methods. We investigate numerous sample values along the x-axis, such as 50, 100, 200, 300, 500, and 700. The graph illustrates that our proposed method closely resembles and overestimates the true RTM line. In comparison, the inverse Box-Cox transformation deviates greatly from the True RTM line. The proposed method demonstrates superior performance compared to inverse transformation, as it closely aligns with the True RTM line.

Similarly, using the same procedure, and same parameter values but with a different quantile point inside the same distribution, the results are shown below.

Table 3.2: Assessing RTM through a transformation using exponential distribution when λ_0 =0.5, λ_1 =0.2, λ_2=0.2, x_0=10

Sample Size	**ARTM**	**ARTMC**	**ARTMT**	**ARTMTR**
50	0.5747793	7.884843	7.465132	2.033031
100	0.5003955	8.068392	7.460439	1.740783
200	0.4273357	8.162271	7.448379	1.553180
300	0.4047200	8.162014	7.445641	1.495055
500	0.3863423	8.060589	7.456404	1.456606
700	0.3772540	8.048028	7.451085	1.440499

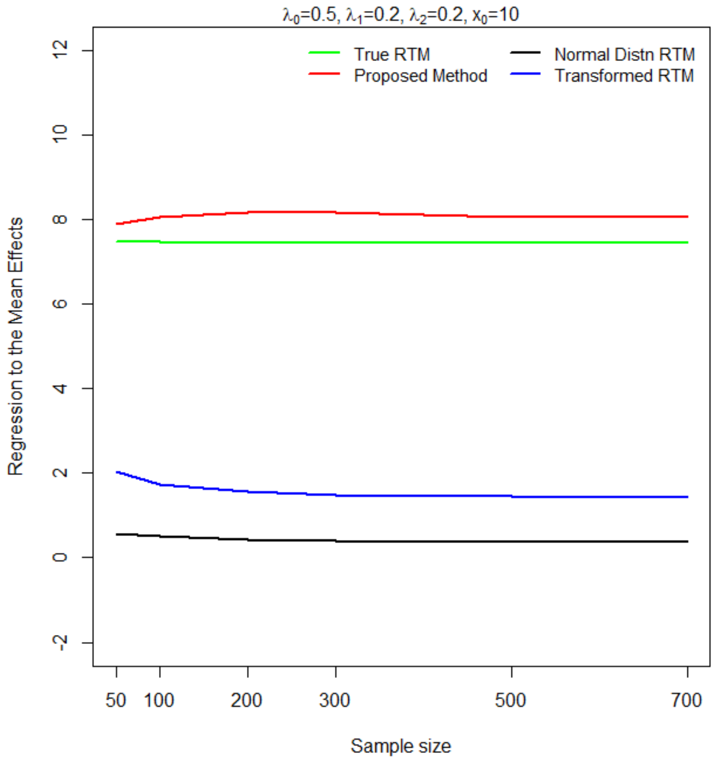

Figure 3.4: Assessing RTM through a transformation using exponential distribution

The above graph presents the behavior of the RTM methods for different sample sizes. The graph illustrates that our proposed method closely resembles the true RTM line. In comparison, the inverse Box-Cox transformation deviates greatly from the True RTM line. So overall, our proposed method demonstrates superior performance compared to inverse transformation, as it closely aligns with the True RTM line.

Similarly, another example involves applying the same procedure but with different parameter values inside the same distribution, the results are shown below.

Table 3.3: Assessing RTM through a transformation using exponential distribution when λ_0 =0.8, λ_1 =0.5, λ_2=0.5, x_0=5

Sample Size	ARTM	ARTMC	ARTMT	ARTMTR
50	0.4510861	3.469634	3.194372	1.697973
100	0.4387682	3.573141	3.192464	1.614436
200	0.4016658	3.615905	3.190236	1.526212
300	0.3498763	3.629690	3.183675	1.417099
500	0.3355998	3.616957	3.192554	1.387105
700	0.3290473	3.608417	3.190693	1.375249

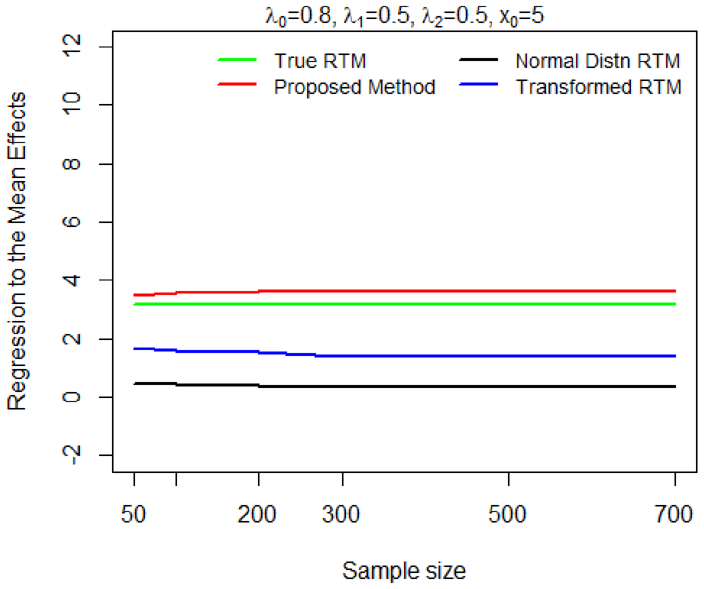

Figure 3.5: Assessing RTM through a transformation using exponential distribution

The graph provided illustrates how different sample sizes $(50, 100, 200, 300, 500, 700)$ impact the performance of RTM methods. The graph indicates that our innovative approach closely corresponds to the True RTM line, unlike alternative techniques. Conversely, the method involving inverse transformation significantly diverges from the True RTM line. The suggested method stands out as superior in comparison to the alternative, as it maintains close proximity to the True RTM line. Especially for sample sizes of 50 and 100, both our suggested method and the true RTM method display closely converging values. As a result, on the whole, our proposed method surpasses the effectiveness of the inverse Box-Cox transformation in the current context. It's important to note that the inverse transformation also showcases commendable performance, although it doesn't fully meet expectations. Among the three instances analyzed, this particular example demonstrates the most adept performance. In this specific scenario, the proposed method shows a stronger resemblance to the True RTM line.

3.4.2 T-Distribution

The t-distribution commonly known as the Student t-distribution, is a kind of probability distribution with a bell-shaped pattern that resembles the normal distribution but with thicker tails. It has one parameter called the degree of freedom, denoted by df.

Algorithm: To check the performance of the proposed methods we used the RStudio software. First, a random sample of size 5000 was generated from the t-distribution. Different sizes of the truncated sample like $(50, 100, 200, 300, 500, 700)$ were generated. The pair of observations (y_1, y_2) were produced using the relation $y_1 = x0 + x1$, and $y_2 = x0 + x2$ with choosing the degree of freedom λ values 100, 70 and 70. The Box-Cox transformation was used to make the data normally distributed. Thus, the two new random variables have been constructed using the Box-Cox transformation with the characteristics of normal data. Now we also check the normality of the transformed data with the help of histogram and Shapiro-Wilk test. After this, we need a cut point used to find the RTM effect. Points above the 85the percentile were considered as the pre observations which are observations above a cut-off point 2.5. The method developed by Khan and Olivier (2022) was used to estimate the RTM effect.

Furthermore, the methods of percentile change and the inverse Box-Cox transformation were used to estimate the RTM effect. Table 3.4 shows the estimated RTM for different different sample sizes.

Table 3.4: Assessing RTM through a transformation using t-distribution when λ_0 =100, λ_1 =70, λ_2=70, x_0=2.5

Sample Size	**ARTM**	**ARTMC**	**ARTMT**	**ARTMTR**
50	0.2400115	0.9134459	0.7607717	1.287333
100	0.2316016	0.8985153	0.7585636	1.262547
200	0.19189553	0.8924755	0.7571582	1.214506
300	0.1863799	0.9082511	0.7601208	1.198788
500	0.1797155	0.8971447	0.7575833	1.189430
700	0.1762163	0.8998460	0.7582616	1.185212

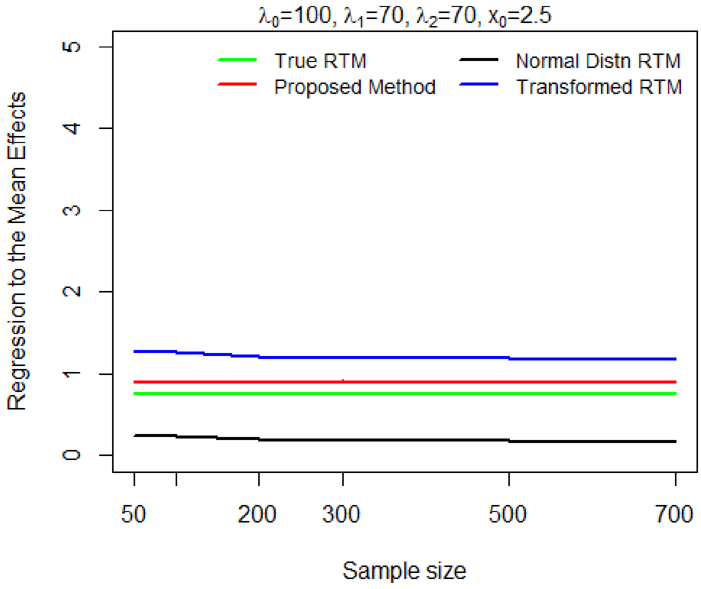

Figure 3.6: Assessing RTM through a transformation using t-distribution

The depicted graph illustrates the RTM's behavior across various techniques. The graph indicates that both our suggested approaches are overestimated but the proposed method aligns overestimate but most closely with the true RTM trajectory. Conversely, the inverse transformation technique overestimates the RTM much more than the proposed technique when compared to the RTM trend. In comparison to alternative methods, our proposed approach exhibits superiority by maintaining proximity to the true RTM. Consequently, in this context, our proposed technique outperforms the inverse Box-Cox transformation, which tends to overestimate the outcome.

Similarly, another example involves applying the same procedure but with different parametric values using the same distribution, the results are shown below.

Table 3.5: Assessing RTM through a transformation using t-distribution when λ_0 =100, λ_1 =50, λ_2=50, x_0=3

Sample Size	ARTM	ARTMC	ARTMT	ARTMTR
50	0.2819583	0.9102047	0.7716964	1.344405
100	0.2360769	0.8985836	0.7719679	1.268804
200	0.1988688	0.9110188	0.7729954	1.216088
300	0.1950896	0.9154427	0.7716701	1.208940
500	0.1896344	0.9087024	0.7733252	1.201261
700	0.1905470	0.9097247	0.7743032	1.201293

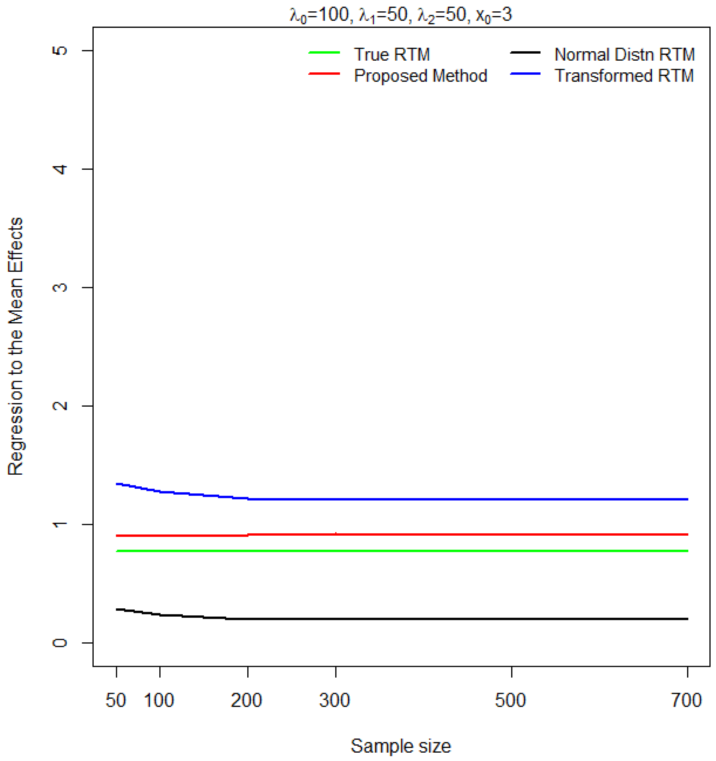

Figure 3.7: Assessing RTM through a transformation using t-distribution

Figure 3.7 expresses the behavior of the RTM methods. We take different sample values on the x-axis such as 50, 100, 200, 300, 500, and 700. Additionally, our proposed method1 overestimates and nearly follows the true RTM line. The inverse transform method is deviating and overestimated compared to the true RTM. It is close to true RTM when the sample size increases. The proposed method is comparatively better than other methods as it is close to the true RTM line. So, overall, our proposed method1 is considered to perform better than the inverse Box-Cox transformation in the current situation.

Similarly, another example involves applying the same procedure but with different permutations of the parameters of the same distribution. The results are shown in Table 3.6.

Table 3.6: Assessing RTM through a transformation using t-distribution when λ_0 =40, λ_1 =40, λ_2=40, x_0=3

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.2396382	0.8952023	0.7687866	1.287693
100	0.2224756	0.9003527	0.7699428	1.251571
200	0.1939674	0.8963792	0.7704535	1.21076
300	0.1835125	0.8945976	0.7685723	1.195458
500	0.1809682	0.8948735	0.7708524	1.19151
700	0.1768102	0.8985692	0.7714213	1.186134

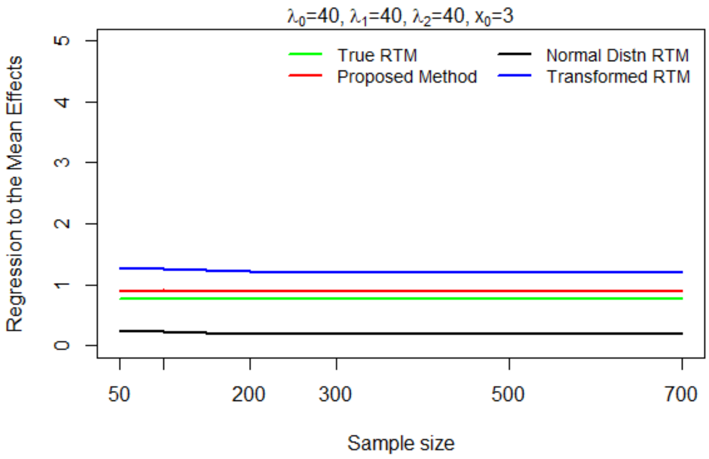

Figure 3.8: Assessing RTM through a transformation using t-distribution

The outcomes are also portrayed in Figure 3.8. The graph illustrates that our suggested technique closely aligns with the true RTM pattern. In comparison to the RTM trend, the inverse transformation method highly overestimates the RTM. In contrast to alternate methods, our proposed method1 demonstrates its superiority by maintaining proximity to the true RTM line. Consequently, considering the entire context, our method1 is deemed more effective than the inverse Box-Cox transformation in this scenario.

Table 3.7: Assessing RTM through a transformation using t-distribution when λ_0 =200, λ_1 =200, λ_2=200, $x0$=2

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.2517502	0.8837896	0.7432032	1.302467
100	0.2053486	0.8948849	0.7439597	1.229024
200	0.1891601	0.8883394	0.7417736	1.203985
300	0.179559	0.8851499	0.7424031	1.191034
500	0.1656299	0.8964062	0.747207	1.17395
700	0.1694763	0.8868447	0.7456849	1.177735

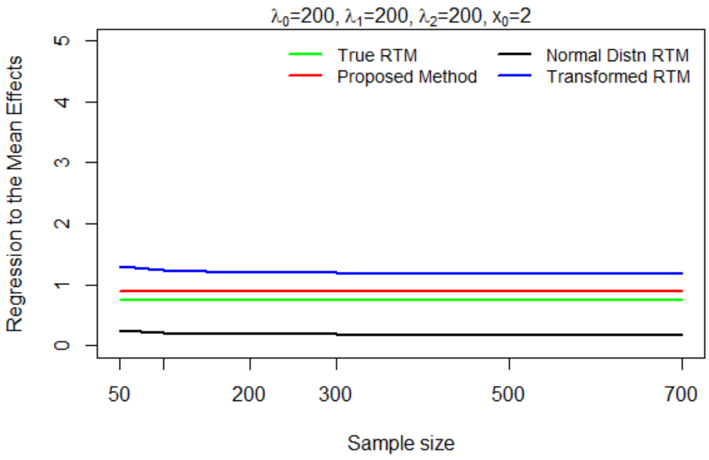

Figure 3.9: Assessing RTM through a transformation using t-distribution

The above graph presents the behavior of the estimated RTM for different methods. The percentile change method for sample sizes 50, 100, 200, 500, and 700 closely estimates the true RTM. The inverse transform method vastly overestimated the true RTM by more than half of the true value of RTM. The proposed method is comparatively better than other methods as it is closer to the true RTM line. Once again, for the new choices of the parameters λ_0 =200, λ_1 =200, λ_2 =200, $x0$=2, the percentile change method performs better than the inverse Box-Cox transformation in the current situation.

Table 3.8: Assessing RTM through a transformation using t-distribution when λ_0 =100, λ_1 =60, λ_2=60, x_0=2

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.2705851	0.9144907	0.764706	1.330423
100	0.2394121	0.9107902	0.7654308	1.275675
200	0.2063942	0.9020308	0.7648465	1.224454
300	0.1901289	0.8980919	0.764812	1.202768
500	0.1794948	0.9017274	0.7652762	1.189271
700	0.1809268	0.8962105	0.7645745	1.190531

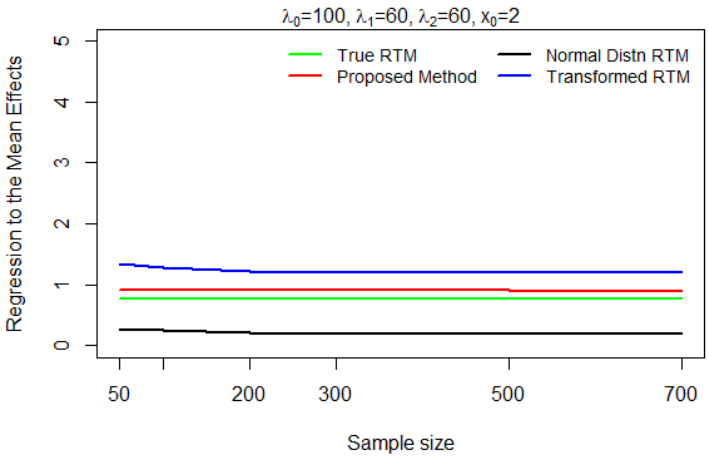

Figure 3.10: Assessing RTM through a transformation using t-distribution

The above graph presents the behavior of the RTM methods. We take different sample values on the x-axis. The graph shows that our proposed method is closest to the true RTM line. After 500, the proposed method is closer to the true RTM when the sample size increases, the proposed RTM is moving closer to the true RTM and stabilizes for larger sample sizes. The inverse transform method again overestimated the true RTM effect. The proposed method is comparatively better than other methods as it is closer to the true RTM line. So, overall, our proposed method1 is considered to perform better than the inverse Box-Cox transformation in the current situation.

3.4.3 Gamma Distribution

The gamma distribution is a continuous distribution that is commonly used in business, science, and engineering. It is a positively skewed distribution. It has two parameters i.e. scale and shape parameter. The gamma function is denoted by Γ. The PDF of the gamma function is

$$f(x;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(k)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

Algorithm: where $\alpha > 0$, $\beta > 0$ and $x > 0$. To check the performance of the proposed methods we used the RStudio software. First, we will generate 5000 random variables from gamma distribution using shape and scale parameters with a different samples 50, 100,200,300,500, and 700. First $x0$, $x1$, and $x2$ were generated from a gamma distribution with the shape parameter as λ values 7, 7 and 7, then correlated pairs of observations were produced using the relation $y_1 = x_0 + x1$ and $y_2 = x_0 + x2$. To convert the non-normal data to normal data, the Box-Cox transformation was used. Now we also checked the assumption of normality with the help of histogram and Shapiro-Wilk test. Both represent the data should be normal. The 85th percentile of the generated data was used to have bivariate truncated data which is equivalent to 6 on average over different samples. The methods developed by Khan and Olivier (2022) were used to estimate the RTM effect. The percentile change and inverse transformation techniques were used to estimate the RTM effect for the non-normal population.

Table 3.9 shows the estimated RTM for sample sizes and methods. ARTM represents the normal distribution of the values for different samples, ARTMT represents the True RTM values, ARTMC represents the proposed RTM values using quantiles and ARTMTR represents the second proposed RTM values using inverse Box-Cox transformation.

Table 3.9: Assessing RTM through a transformation using gamma distribution when λ_0 =7 λ_1 =7, λ_2=7, x_0=4

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.13737706	0.6955125	0.7878318	1.158826
100	0.1252764	0.5658763	0.7869629	1.140318
200	0.10734913	0.6009333	0.7881772	1.115027
300	0.09921443	0.6523553	0.7881804	1.104691
500	0.09521137	0.6395224	0.7874632	1.099521
700	0.09358064	0.6762762	0.7854989	1.097465

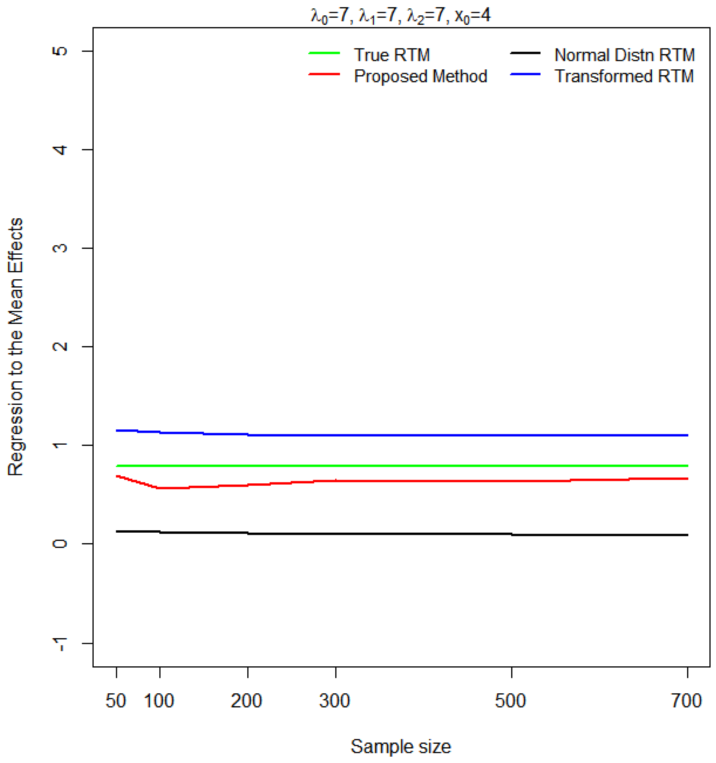

Figure 3.11: Assessing RTM through a transformation using gamma distribution

Figure 3.11 also illustrates the performance of the methods for estimation of the RTM effect. The graph indicates that our suggested approach aligns most closely with the actual RTM line. However, the outcomes from the inverse transform method overestimate and resemble the true RTM line. Within the context of the gamma distribution, our suggested approach has even closer results compared to the inverse transform yields. Overall, the suggested technique stands out as superior among the various methods.

For another permutation of the parameters in gamma distribution, the results of estimated RTM using different techniques are given in Table 3.10.

Table 3.10: Assessing RTM through a transformation using gamma distribution when λ_0 =5 λ_1 =5, λ_2=5, x_0=4

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.1517956	0.795597	0.8996485	1.178645
100	0.1307754	0.8466264	0.8959284	1.145108
200	0.120365	0.76453	0.8981219	1.130132
300	0.1111027	0.8210398	0.8988831	1.117558
500	0.1041235	0.8138073	0.8982093	1.109136
700	0.1038676	0.8172235	0.8986351	1.108675

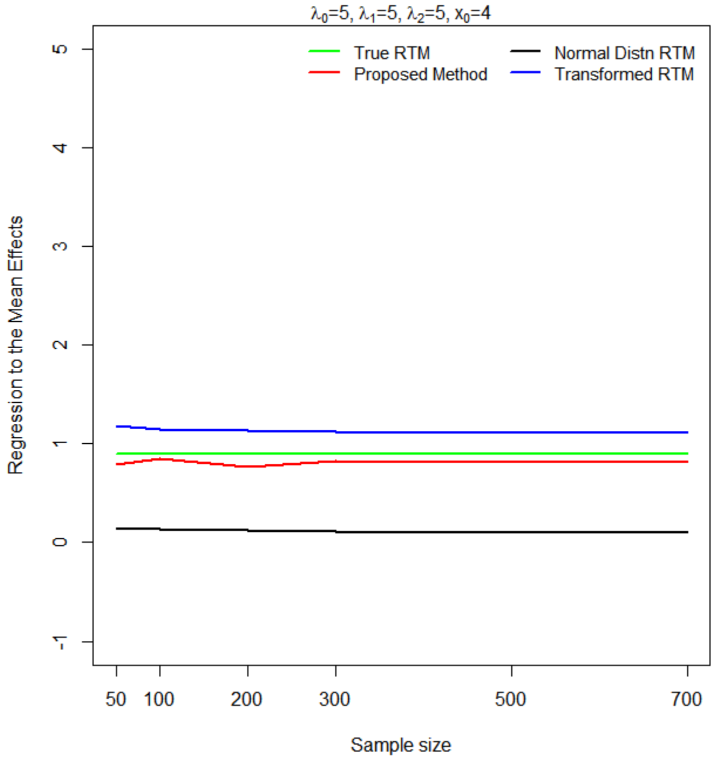

Figure 3.12: Assessing RTM through a transformation using gamma distribution

The graph above depicts the observed behavior of the estimated RTM using various techniques. Initially, with sample sizes such as 100, 300, and above, the proposed method closely aligns with the actual RTM. However, as the sample size for the proposed RTM closely resembles the true RTM. Conversely, the inverse transformation method tends to overestimate

RTM in this scenario. In comparison, the suggested percentile change method stands out as superior among the alternatives, given its proximity to the true RTM trend. In conclusion, our proposed method1 outperforms the inverse Box-Cox transformation.

In another permutation of the parameters, the estimated RTM for different sample sizes and methods is presented in Table 3.11 below.

Table 3.11: Assessing RTM through a transformation using gamma distribution when λ_0 =4 λ_1 =4, λ_2=4, x_0=4

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.1388959	0.7902546	0.8111613	1.161058
100	0.1403572	0.8255786	0.8107353	1.157584
200	0.1201642	0.7855641	0.8109963	1.129271
300	0.1098971	0.8036758	0.8117063	1.116177
500	0.1074623	0.8507712	0.8104843	1.112694
700	0.1000879	0.8015176	0.8120511	1.104396

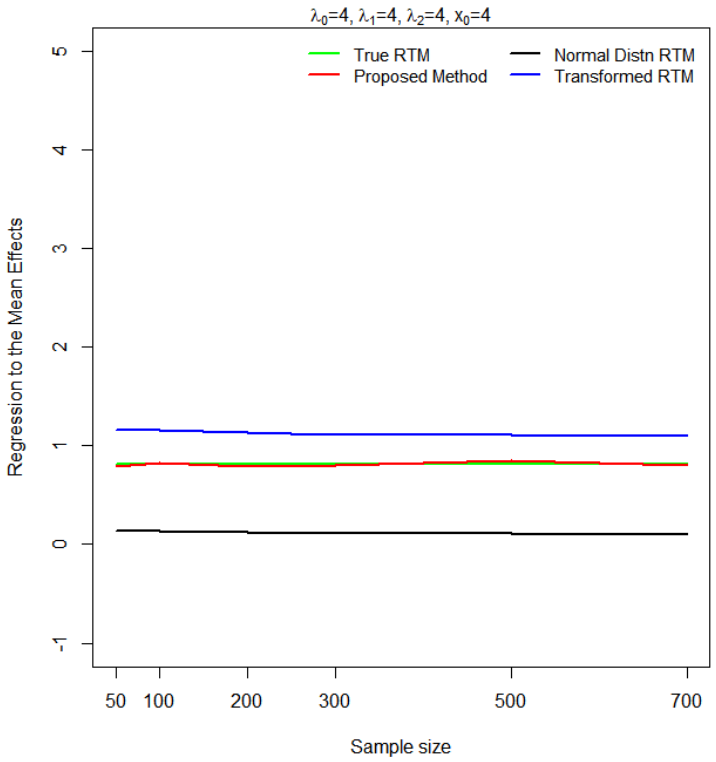

Figure 3.13: Assessing RTM through a transformation using gamma distribution

The depicted graph illustrates the behavior of the RTM techniques. Evidently, our suggested approach closely tracks the True RTM trend. our suggested technique is best performed in this case. Conversely, the inverse transformed method tends to overestimate in this context. In comparison, the proposed method stands out as superior compared to other techniques due to its proximity to the True RTM line. Consequently, our proposed method1 is generally regarded as exhibiting superior performance in contrast to the inverse Box-Cox transformation.

Table 3.12: Assessing RTM through a transformation using gamma distribution when λ_0 =15 λ_1 =12, λ_2=12, x_0=3

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.08549664	0.281832	0.3807379	1.094176
100	0.07252033	0.2182242	0.3808462	1.077608
200	0.06803098	0.2208731	0.3817402	1.071389
300	0.06252937	0.2431511	0.381205	1.0652
500	0.05799217	0.2171518	0.3809285	1.059823
700	0.05839066	0.2702581	0.3824145	1.06016

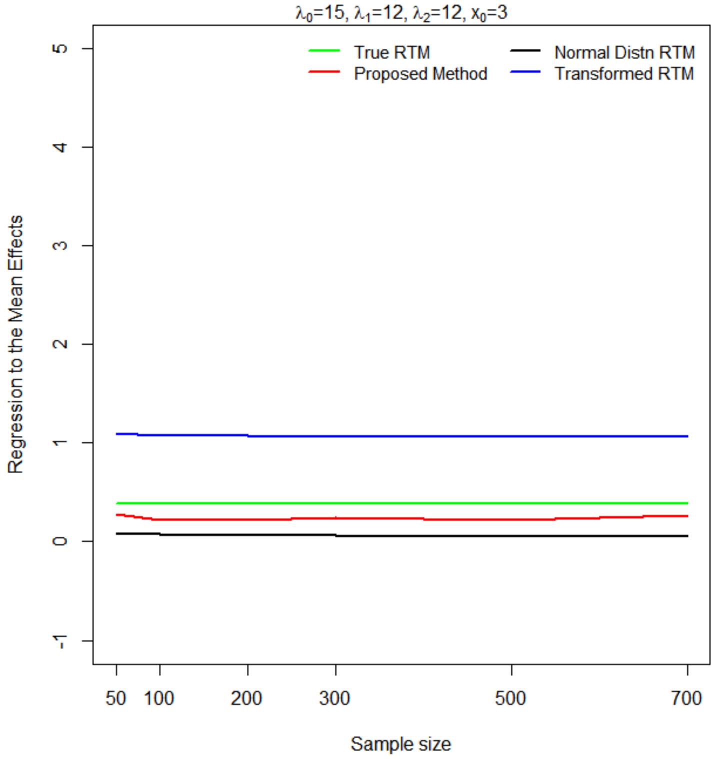

Figure 3.14: Assessing RTM through a transformation using gamma distribution

The values given in Table 3.12 are also depicted in Figure 3.14.

Chapter 4

Assessing RTM for Discrete Distributions

4.1 Poisson Distribution

Poisson distribution is used to model count data. It has only one parameter called the average rate of occurrance denoted by λ. The probability mass function of the Poisson distribution is

$$P(Y = y; \lambda) = \frac{e^{-\lambda}\lambda^y}{y!}$$

where $\lambda > 0$ and $y = 0, 1, 2, \cdots$.

Algorithm: To check the performance of the proposed methods we used the RStudio software. Firstly, a large sample of size 5000 was generated from the Poisson distribution with λ equal to 5, 2 and 2 and the truncated samples of different sizes $50, 100, 200, 300, 500, 700$. The generated variables $x0$, $x1$, and $x2$ were linked by the equation $y_i = x_0 + x_i$ for $i = 1, 2$ to produce the bivariate count data (y_1, y_2). To convert the non-normal data to normal data, the Box-Cox transformation was used. A positive number was added to avoid the computational complication, e.g, $\log(0)$. The cut-off point against the 85 percentages was found to be 11. RTM was estimated using the methods developed under the bivariate normal distribution (Khan and Olivier, 2022).

The inverse Box-Cox transformation, and the percentile change approach were used to estimated the RTM effect. Table 4.1 shows the estimated RTM against different sample sizes and methods.

Table 4.1: Assessing RTM through a transformation using poisson distribution when λ_0 =5,λ_1 =2, λ_2=2, x_0=11

Sample Size	ARTM	ARTMC	ARTMT	ARTMTR
50	0.03847092	1.399420	1.441414	1.040800
100	0.03405013	0.322383	1.446547	1.034539
200	0.03256721	1.468332	1.445569	1.032909
300	0.03089452	0.140970	1.438509	1.031166
500	0.02946150	1.241715	1.438775	1.029684
700	0.02963760	1.399923	1.437653	1.029855

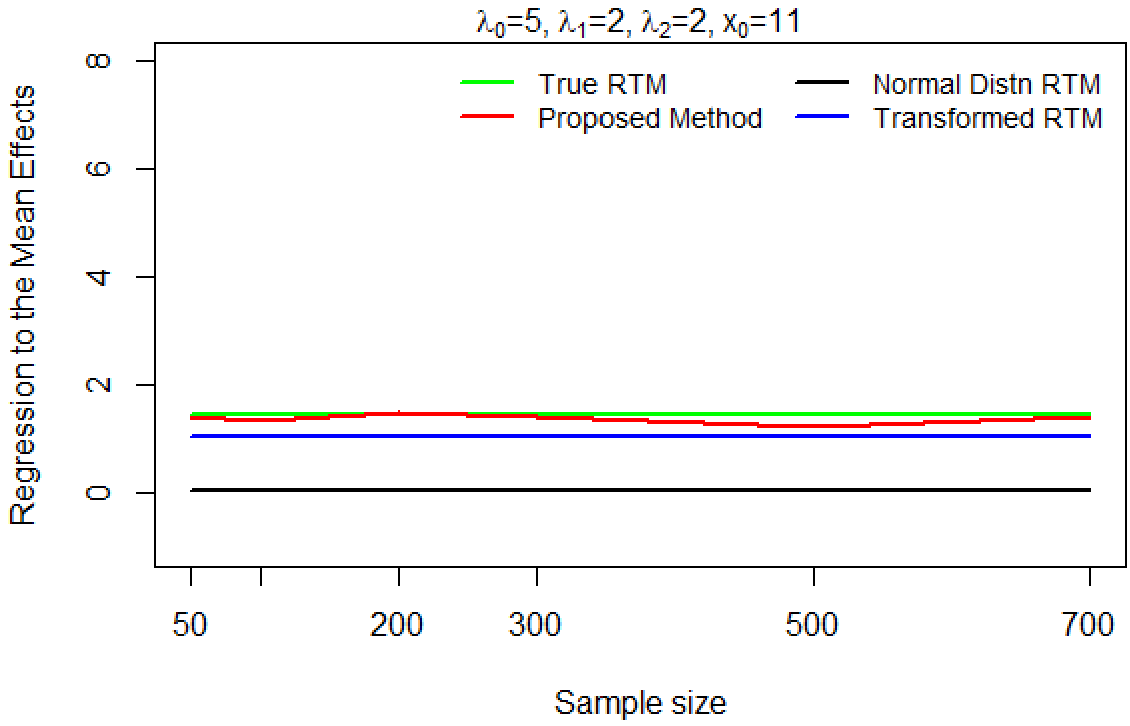

Figure 4.1: Assessing RTM through a transformation using Poisson distribution

Graphically, the behavior of the estimated RTM was also depicted. The graph shows that percentile change method very closely estimates the true RTM for all sample sizes. The inverse transformation is also close to the true RTM line. The percentile change method performs well for discrete distribution as well.

As a different permutation of the parametric values, small values of the average rate of occurrence were considered, and the resulting estimated RTM values are given in Table 4.1 .

Here both the suggested methods are performing well and as the sample size increases, the true RTM and the estimated RTM approaches each other.

Table 4.2: Assessing RTM through a transformation using poisson distribution when λ_0 =2,λ_1 =1, λ_2=1, x_0=6

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.01333147	1.213055	1.20003	1.01399
100	0.01100145	1.206735	1.204576	1.011054
200	0.01162699	1.208547	1.204914	1.011673
300	0.01127606	1.201643	1.199155	1.011315
500	0.01173736	1.133012	1.201077	1.011777
700	0.01149682	1.071588	1.19988	1.011534

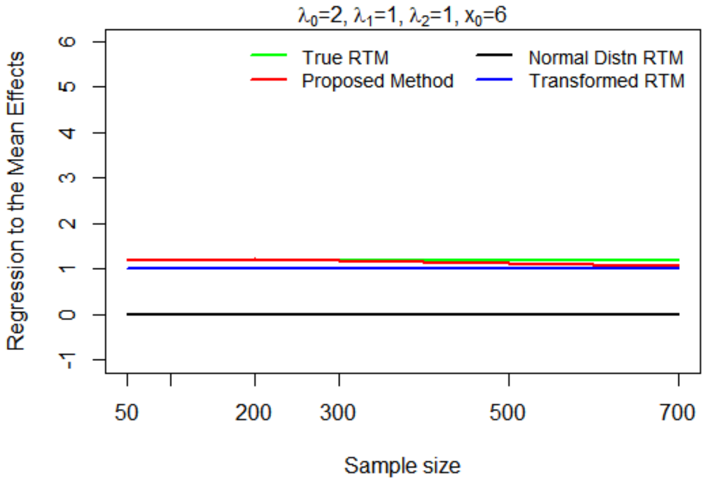

Figure 4.2: Assessing RTM through a transformation using Poisson distribution

The values given in Table 4.2 are also depicted in Figure 4.2.

Similarly, data were generated for different choices of the parameters of the Poisson distribution. The estimated RTM using the two different methods is presented in Table 4.3. For larger sample sizes the difference between the true values of RTM and those estimated by the percentile change decreases. The inverse transformation method remains biased for all the sample sizes.

Table 4.3: Assessing RTM through a transformation using poisson distribution when λ_0 =7 λ_1 =4, λ_2=4, x_0=15

Sample Size	aRTM	aRTMc	aRTMt	aRTMtr
50	0.09044454	1.693271	2.057195	1.098766
100	0.07679414	1.569832	2.065759	1.078966
200	0.07347344	1.924171	2.05975	1.075061
300	0.06835397	1.711881	2.057873	1.069574
500	0.06690417	1.657408	2.0586	1.067961
700	0.06789422	1.856034	2.049104	1.068949

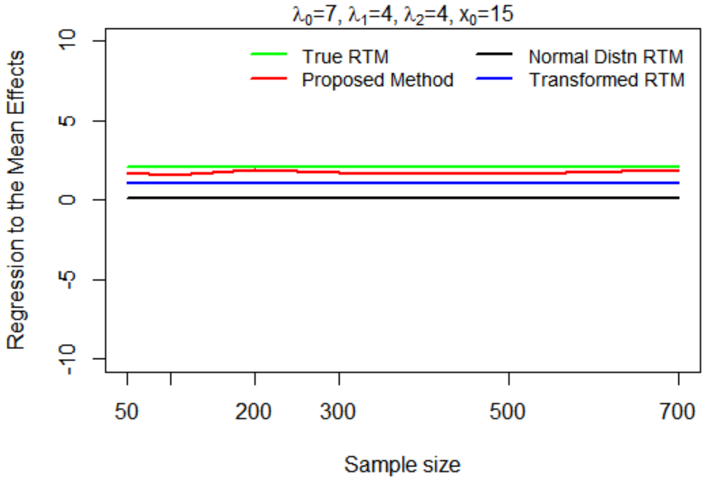

Figure 4.3: Assessing RTM through a transformation using Poisson distribution

The same given in Table 4.3 have been portrayed in Figure 4.3.

4.2 Negative Binomial Distribution

In a negative binomial experiment, a discrete probability distribution known as a negative binomial distribution is used to distribute random variables. The distribution is almost the same as the binomial experiment with only one difference. In the binomial experiment, there is a fixed number of trails. However, in a negative binomial experiment, there is a fixed number of successes. The PDF of negative binomial e

$$P(X = k; r, p) = \binom{k + r - 1}{k} p^r (1 - p)^k$$

Algorithm: To check the performance of the proposed methods we used the RStudio software. First, we will generate a 5000 random variable from a Negative binomial distribution with a different sample like $50, 100, 200, 300, 500, 700$. $x0$, $x1$, and $x2$ with λ parameter 7, 5, and 5, then correlated $x1$ and $x2$ with $x0$ using the formulae $y_1 = x_0 + x_1$ and $y_2 = x_0 + x_2$ as a bivariate count data from the bivariate negative binomial distribution. After generating the bivariate random variables, we apply the Box-Cox transformation. To convert the non-normal data to normal data by using the parameter values i.e. number of successes as a λ value that we use in generating random variables. The two new random variables have been constructed using the Box-Cox transformation with the characteristics of normal data. After this, we need a cut point used to find the RTM. Against the 85 percentile, the average truncated point was found to be 12. Then we find all the value that is used in True RTM using the existing formulas of Regression to the mean: Estimation and adjustment under the bivariate normal distributionKhan and Olivier (2022).

Furthermore, we find proposed RTM percentile points using the quantile points. Finding the proposed method1, we used another suggested method to convert the data into its original data, the method name is the inverse Box-Cox transformation, With the help of this we also want to check whether the RTM returns to its original state or not. The below Table 4.4 shows the different values of different sample sizes.

Table 4.4: Assessing RTM through a transformation using negative binomial distribution distribution when λ_0 =7, λ_1 =5, λ_2 =5, x_0=12, prob =0.6

Sample Size	**aRTM**	**aRTMc**	**aRTMt**	**aRTMtr**
50	0.1866011	3.231867	2.873298	1.238246
100	0.1338608	3.142283	2.886417	1.14569
200	0.1175479	3.021516	2.875532	1.121764
300	0.1194922	3.149227	2.884187	1.123424
500	0.1121149	3.109209	2.870332	1.115302
700	0.115745	3.091725	2.88152	1.11898

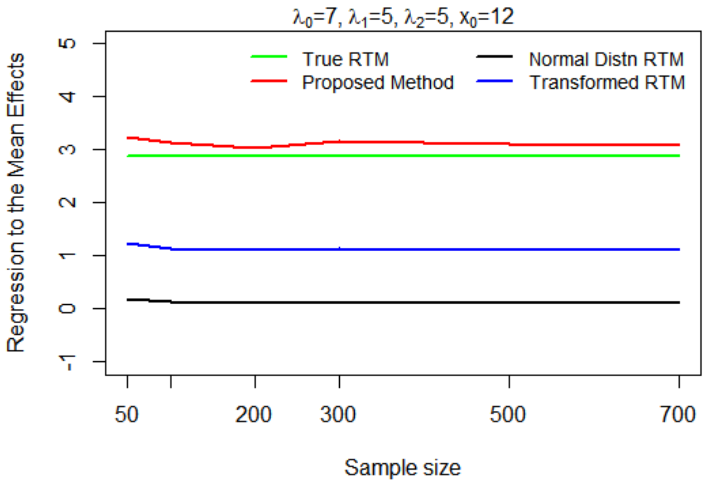

Figure 4.4: Assessing RTM through a transformation using negative binomial distribution distribution

Figure 4.4 presents the behavior of the RTM methods. We take different sample values on the x-axis. The graph shows that our proposed methods are closely aligned with the actual RTM. In comparison, the inverse technique is too far away from the true RTM. Overall over proposed technique performs well as compared to other techniques.

Table 4.5: Assessing RTM through a transformation using negative binomial distribution distribution when λ_0 =3, λ_1 =3, λ_2=3, x_0=6

Sample Size	**ARTM**	**ARTMC**	**ARTMT**	**ARTMTR**
50	0.10050249	2.442852	2.650404	1.113044
100	0.07983249	2.472135	2.658698	1.083011
200	0.07585102	2.482147	2.664030	1.077761
300	0.07326597	2.471725	2.658566	1.074820
500	0.07152363	2.453475	2.666679	1.072845
700	0.07102765	2.425575	2.651657	1.072292

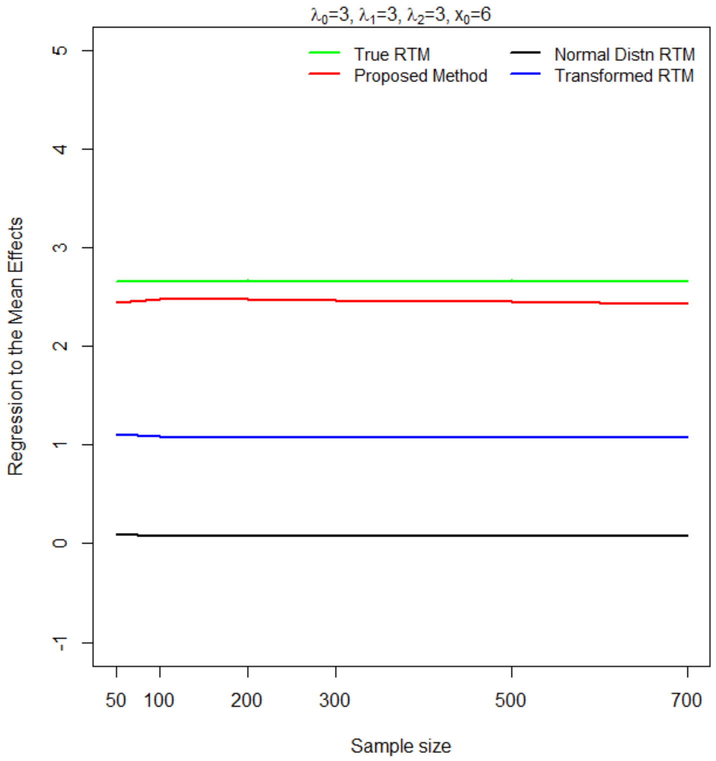

Figure 4.5: Assessing RTM through a transformation using negative binomial distribution distribution

The above graph presents the behavior of the RTM methods. We take different sample values on the x-axis and regression effect on the y-axis. The graph shows that our proposed methods perform well as compared to the inverse transformation. Because our proposed method is close to the true RTM and inverse deviates significantly from the true RTM. In a nutshell, the proposed performs better than the inverse technique.

Table 4.6: Assessing RTM through a transformation using negative binomial distribution distribution when λ_0 =10, λ_1 =10, λ_2=10, x_0=33, prob =0.5

Sample Size	**aRTM**	**aRTMc**	**aRTMt**	**aRTMtr**
50	0.8169171	5.853239	8.385966	2.389934
100	0.735523	5.310843	8.366834	2.094927
200	0.5982958	6.270405	8.378128	1.769087
300	0.5444862	6.09316	8.3685	1.658782
500	0.5390819	6.730671	8.366261	1.634028
700	0.5099215	6.100299	8.361804	1.592232

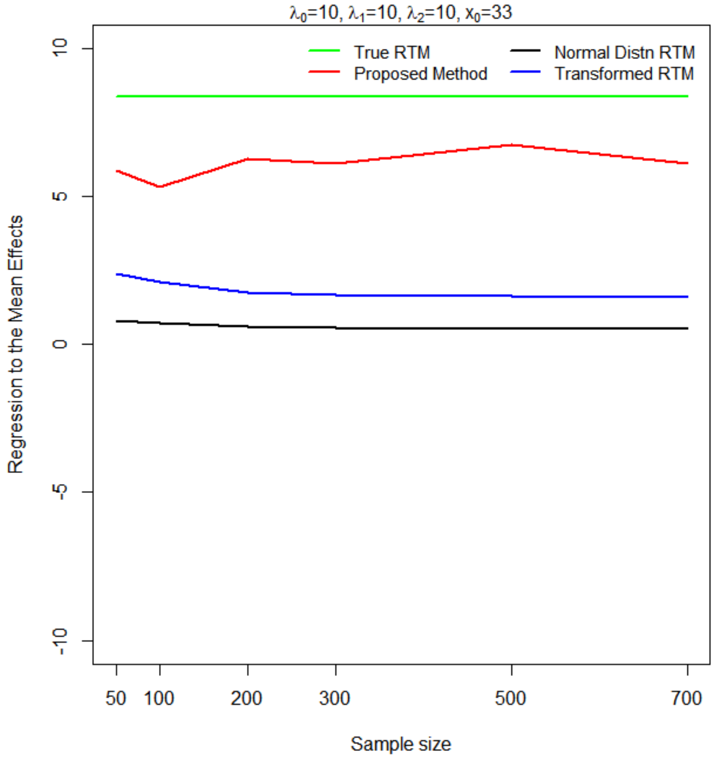

Figure 4.6: Assessing RTM through a transformation using negative binomial distribution

The above graph presents the behavior of the RTM methods. We take different sample values on the x-axis like 50, 100, 200, 300, 500, and 700. The graph shows that our suggested technique is close to the true RTM line but sometimes it is close to RTM and at some points, it is far way to the RTM line. In this case, our suggested method does not show an appropriate result. The inverse approach is diametrically opposed to the true RTM. So, in general, the suggested strategy outperforms the other.

Chapter 5

Conclusion and Future Work

This chapter discusses the conclusion of the study and the possible future work. In intervention studies, the subjects are selected on specific criteria for an intervention, thereby resulting in data that come from either the left or right tails of a distribution. Regression to mean (RTM) occurs when the initial observations are extreme, i.e., selected in the tail of a distribution. When the intervention/treatment is applied to such subjects, the total effect does not only include the treatment effect but also the RTM effect, thereby leading to incorrect conclusion about the effectiveness of an intervention effect. Thus, estimating and accounting for RTM is very essential to accurately estimate the intervention/treatment effect in research areas like epidemiology, health, clinical trials, sports, economics, etc.

Past research has mainly focused on developing methods under the well known distributions like the bivariate truncated normal, Poisson, binomial distributions. However, not all data follow these distribution, and warrants to investigate estimation of the RTM effect for other non-normal distributions. The main purpose of this thesis is to assess estimation of RTM for non-normal populations through the Box-Cox transformation and the newly defined percentile change method.

In this procedure, the non-normal population is transformed into a normal population using the Box-Cox transformation. The readily available method developed under the bivariate normal distribution is used to estimate the RTM effect. The quantile points are identified whose difference gives the RTM effect. The percentile change probability is used to identify the quantile points in the non-normal distribution and the RTM effect thereafter.

The methods developed were used to estimate the RTM effect for skewed distribution (exponential and gamma) using different permutations of the parameters and different sample sizes. The percentile change method closely estimated the true RTM while the inverse transformation method overestimated the true RTM effect. Moreover, the symmetric t-distribution was also used to generate data and RTM was estimated. Similar results were found for the non-normal symmetric distribution.

Data were also generated from two discrete distributions for different permutation of the parameters. Moreover, different sample sizes were also used for the estimation of RTM. The proposed methods closely estimated the true RTM. Moreover, the percentile change method produced very close results as compared to the inverse transformation method.

In conclusion, the proposed percentile change method closely estimate the RTM effect when data are generated from a non-normal population. Thus, to avoid erroneous conclusion about the effectiveness of an intervention effect and accurately estimate it, the method developed in thesis should be used when the data follow non-normal distributions.

5.1 Future Work

The work done in this thesis can be expanded by

1. using different transformation techniques for transforming non-normal data to normal,
2. and using non-stationary distributions to know when both RTM and treatment effects are part of the observed change.

References

Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J. (2005). Regression to the mean: what it is and how to deal with it. *International Journal of Epidemiology*, 34(1):215–220.

Beath, K. J. and Dobson, A. J. (1991). Regression to the mean for nonnormal populations. *Biometrika*, 78(2):431–435.

Biarnés, M. and Monés, J. (2020). Regression to the mean in measurements of growth rates in geographic atrophy. *Ophthalmic Research*, 63(5):460–465.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B(Methodological)*, 26(2):211–243.

Browne, S., Halligan, P., Wade, D., and Taggart, D. (1999). Cognitive performance after cardiac operation: implications of regression toward the mean. *The Journal of Thoracic and Cardiovascular Surgery*, 117(3):481–485.

Bush, H. F., Canning, M. D., et al. (2006). Regression towards the mean versus efficient market hypothesis: An empirical study. *Journal of Business and Economics Research (JBER)*, 4(12).

Chinn, S. and Heller, R. F. (1981). Some further results concerning regression to the mean. *American Journal of Epidemiology*, 114(6):902–905.

Cochrane, K., Williams, B., Fischer, J., Samson, K., Pei, L., and Karakochuk, C. (2020). Regression to the mean: A statistical phenomenon of worthy consideration in anemia research. *Current Developments in Nutrition*, 4(10):nzaa152.

Daimon, T. (2011). Box-cox transformation. In *International Encyclopedia of Statistical Science*.

Davis, C. (1976). The effect of regression to the mean in epidemiologic and clinical studies. *American Journal of Epidemiology*, 104(5):493–498.

Galton, F. (1886). Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263.

Gardner, M. and Heady, J. (1973). Some effects of within-person variability in epidemiological studies. *Journal of Chronic Diseases*, 26(12):781–795.

Gaudry, M. J. I. and Laferriere, R. R. (1988). The box-cox transformation: Power invariance and a new interpretation. *Economics Letters*, 30:27–29.

Healy, M. and Goldstein, H. (1978). Regression to the mean. *Annals of Human Biology*, 5(3):277–280.

Hossain, M. Z. (2011). The use of box-cox transformation technique in economic and statistical analyses. *Journal of Emerging Trends in Economics and Management Sciences*, 2:32–39.

James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics*, pages 121–130.

John, M. and Jawad, A. (2010). Assessing the regression to the mean for non-normal populations via kernel estimators. *North American Journal of Medical Sciences*, 2(7):288 – 292.

Johnson, W. D. and George, V. T. (1991). Effect of regression to the mean in the presence of within-subject variability. *Statistics in Medicine*, 10(8):1295–1302.

Khan, M. and Olivier, J. (2018). Quantifying the regression to the mean effect in poisson processes. *Statistics in Medicine*, 37(26):3832–3848.

Khan, M. and Olivier, J. (2019). Regression to the mean for the bivariate binomial distribution. *Statistics in Medicine*, 38(13):2391–2412.

Khan, M. and Olivier, J. (2022). Regression to the mean: Estimation and adjustment under the bivariate normal distribution. *Communications in Statistics-Theory and Methods*, pages 1–19.

Kypri, K. (2020). Interpretation of within-group change in randomised trials. *BMC Psychiatry*, 20(1):1–2.

Moore, M., Atkins, E., Abdul Salam, M., Callisaya, M., Hare, J., Marwick, T., Nelson, M., Wright, L., Sharman, J., and Rodgers, A. (2019). Regression to the mean of repeated ambulatory blood pressure monitoring in five studies. *Journal of Hypertension*, 37(1):24–29.

Müller, H., Abramson, I., and Azari, R. (2003). Nonparametric regression to the mean. *Proceedings of the Nationa Academy of Sciences of the United States of America*, 100(17):9715 – 9720.

Osborne, J. W. (2010). Improving your data transformations: Applying the box-cox transformation. *Practical Assessment, Research and Evaluation*, 15:12.

Ozgur Asar, O. I. and Dag, O. (2014). Estimating box-cox power transformation parameter via goodness- of-fit tests. *Communications in Statistics - Simulation and Computation*, 46:105 – 91.

Pritchett, L. and Summers, L. H. (2014). Asiaphoria meets regression to the mean. Technical report, National Bureau of Economic Research.

Proietti, T. and Lütkepohl, H. (2013). Does the box-cox transformation help in forecasting macroeconomic time series? *International Journal of Forecasting*, 29:88–99.

Rahman, M. (1999). Estimating the box-cox transformation via shapiro-wilk w statistic. *Communications in Statistics - Simulation and Computation*, 28:223–241.

Sakia, R. M. (1992a). The box-cox transformation technique: a review. *Journal of the Royal Statistical Society Series D: The Statistician*, 41(2):169–178.

Sakia, R. M. (1992b). The box-cox transformation technique: a review. *The Statistician*, 41:169–178.

Sarkar, N. (1985). Box-cox transformation and the problem of heteroscedasticity. *Communications in Statistics-Theory and Methods*, 14:363–379.

Stigler, S. M. (1997). Regression towards the mean, historically considered. *Statistical Methods in Medical Research*, 6(2):103–114.

Tommaso, P. and Helmut, L. (2011). Does the box-cox transformation help in forecasting macroeconomic time series?

Wang, N., Atkins, E. R., Salam, A., Moore, M. N., Sharman, J. E., and Rodgers, A. (2020). Regression to the mean in home blood pressure: Analyses of the bp guide study. *The Journal of Clinical Hypertension*, 22(7):1184–1191.

Yang, Z. (1996). Some asymptotic results on box-cox transformation methodology. *Communications in Statistics-Theory and Methods*, 25:403–414.

Yudkin, P. and Stratton, I. (1996). How to deal with regression to the mean in intervention studies. *The Lancet*, 347(8996):241–244.