Estimation of Regression to the mean for the Bivariate Generalized Poisson Lindley Distribution

ومن يؤت الحكمة
فقد اوتی خيرًا كثيرًا

QUAID-I-AZAM UNIVERSITY
ISLAMABAD

By

Abdul Malik

Department of Statistics

Faculty of Natural Sciences

Quaid-i-Azam University, Islamabad

2024

بسم الله الرحمن الرحيم

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

Estimation of Regression to the mean for the Bivariate Generalized Poisson Lindley Distribution

By

Abdul Malik

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

2024

Declaration

I "Abdul Malik" hereby solemnly declare that this thesis titled, "Estimation of Regression to the mean for the Bivariate Generalized Poisson Lindley Distribution ".

- 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 of such quotations, this thesis is entirely my own 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:________________

CERTIFICATE

Estimation of Regression to the mean for the Bivariate Generalized Poisson Lindley Distribution

By

Abdul Malik

(Reg. No. 02222211013)

A THESIS SUBMITTED IN THE PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF M.PHIL. IN

STATISTICS

We accept this thesis as conforming to the required standards

1. For

Dr. Manzoor Khan
(Supervisor)

2.

Prof. Dr. Tahir Mehmood
(External Examiner)

3 30/04

Prof. Dr. Ijaz Hussain
(Chairman)

DEPARTMENT OF STATISTICS
QUAID-I-AZAM UNIVERSITY
ISLAMABAD, PAKISTAN
2023

Dedication

I am feeling great honor and pleasure to dedicate this research work to

My Parents and Family

Whose endless affection, prayers and wishes have been a great source of comfort for me during my whole education period and my life

Acknowledgments

This thesis would not have been possible without the unwavering support of my advisor, Dr. Manzoor Khan, whose office door was always open whenever I faced challenges or had questions about my work. I extend my gratitude to the Almighty Allah for granting me the opportunity, determination, and strength to undertake this research. His continuous grace and mercy accompanied me throughout my life and during my research journey.

I would like to convey my sincere appreciation to the Chairman of the Department of Statistics, Dr. Ijaz Hussain, and to all the teachers, particularly Dr. Abdul Haq, Dr. Ismail Shah, and Dr. Sajid Ali. Their guidance and teachings have illuminated my knowledge throughout my academic career and provided invaluable support during my research endeavors.

Abstract

Regression to the mean (RTM) occurs when measurement/observations tends toward the mean of the population upon re-measurement. In pre-post studies interventions are applied to subjects based on some cut-off points or baseline criteria. The change in the difference of the pre-post means after the application of an intervention is known as the total effect which is the sum of RTM and intervention effects. The total effect needs to be accounted for the RTM effect to unbiasedly estimate the intervention effect. In this work, we have derived the RTM effect for bivariate generalized Poisson lindley distribution with a particular emphasis on positive correlated count variables. Expressions for the total effect are derived for a model and then partitioned into RTM and intervention effect. Maximum likelihood estimators are derived and its properties are verified via simulations. Finally, using the bivariate accident count data of 122 shunters, RTM and intervention effect are estimated.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Regression toward the mean (RTM) is a statistical phenomenon observed when multiple measurements are taken from the same subject at different times and the calculated observations are found closed to the mean. In such cases, if an initial observation deviates significantly from the true mean, it tends to shift closer to the true mean when a second measurement is made. In the year 1886, the concept of "regression toward the mean" (RTM) was initially introduced by Sir Francis (Galton, 1886) Galton's observation illustrated that parents who had heights significantly taller than the average population height tended to have children with heights closer to the population average. Conversely, parents with heights notably shorter than the population mean tended to have children whose heights were also closer to the population average.

A simple illustration of RTM is shown in figure 1.1, which uses a fictitious but realistic distribution of high-density lipoprotein (HDL) levels in a single person. A normal distribution of the subject's observations is shown in the first panel. Practically speaking, we don't know the true mean value for this subject, which is shown as 50 mg/dL here, and we presume it stays constant throughout time.We assume that the oscillations are due only to random error, which could be caused by differences in the HDL cholesterol readings or dietary decisions made by the subjects.

We display the observed HDL cholesterol value for this person in the second panel, which is 30 mg/dL, which is on the low side. It is more likely that the number, as indicated in the third panel, will be more than 30 mg/dL if we were to take another reading from the same person. Alternatively put, as the third panel shows, the subsequent observed value should be more in line with the 50 mg/dL mean.

Generally, we find that highly extreme (very high or very low) observations tend to be followed by less extreme measurements that are closer to the genuine mean or actual average of the topic when we examine many measurements obtained from the same subject. RTM presents a practical challenge in that it must be possible to discern between an actual change and the expected change resulting from natural variation. For instance, looking at the third panel of Figure 1.1, we might assume that the subject's HDL cholesterol has gone up, but in reality, the subject's true mean HDL cholesterol has stayed the same and the first measurement was just abnormally low.

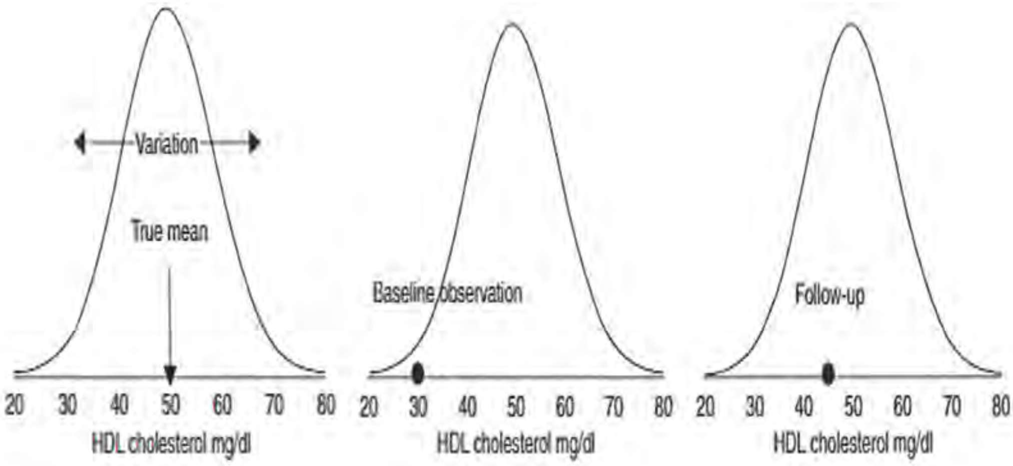

Figure 1.1: RTM effect on baseline and follow-up measurements of high density lipoprotein (HDL) cholesterol with true mean and variation.

Regression to the Mean (RTM) is a concept that originates from random measurement errors that happen when multiple observations are made on the same subject at different times (Barnett et al., 2005). The pre-post variables will have perfect correlation and the RTM effect won't exist in the absence of random error. RTM is a common phenomenon because data are rarely observed error-free and generally contain some sort of error.

James (1973) examined the effects of RTM in uncontrol clinical trials without a control group, in which observations were made both before and after a treatment was administered. He came to the conclusion that there were regression effects, two of which were caused by measurement variation and biological variation over time. The author made a compelling case for randomized clinical trials and emphasized the value of the control group. To prevent drawing incorrect conclusions, it is crucial to isolate the RTM and treatment effects.

Prior et al. (2005) addressed the challenge of mitigating the regression to the mean (RTM) effect while working on the reduction of cardiovascular risk factors through health programs involving the implementation of minor interventions in the field of medicine. Notably, there was no control group in their study. The researchers concluded that the intervention had a significant impact and resulted in a decrease in cardiovascular risk. This reduction was achieved through improvements in diastolic blood pressure, lower smoking rates, and lower levels of hypercholesterolemia in high-risk individuals. Nonetheless, RTM was the reason for the low-risk patients' decreased systolic blood pressure.

Using the Alcohol Use Disorders Identification Test (AUDIT) results, (McCambridge et al., 2014) looked into a cohort study with 967 alcoholic participants.After six months of intervention follow-up, the study found that participants with lower baseline scores tended to have higher AUDIT scores. After the intervention's six-month follow-up, participants with high baseline scores tended to have lower AUDIT scores, indicating that the RTM effect is probably the reason for the decline.

The amount of RTM in evaluating the placebo response in clinical trials for Raynaud's Phenomenon (RP) was examined by (Roustit et al., 2022). By calculating the difference between intake at baseline and after treatment, the authors were able to estimate the placebo effect. The outcome was significant with a placebo response at baseline, according to the authors, suggesting the presence of RTM, which was later confirmed by Galton squeeze plots on individual data.

In a free-living cohort of adults with newly diagnosed diabetes and intermediate hyperglycemia, (Schmidt et al., 2021) estimated the glycemic regression both before and after adjusting the RTM. The study found that accounting for the RTM effect decreased the number of diabetes cases from 526 to 94, the IH defined by the American Diabetes Association (ADA) from 6,182 to 5,711, and the IH defined by the World Health Organization (WHO) from 3,118 to 1,986.

The RTM effect is not limited to the medical field; it also exists in economics and the determination of market capital efficiency (Bush et al., 2006),economic forecasting (Pritchett and Summers, 2014),and measurements of geographic atrophy growth rate (Biarnés and Monés, 2020). Among the other study areas include sports (Lee and Smith, 2002), road accidents (Retting et al., 2003), birth weight (Wilcox et al., 1996), anemia research of hemoglobin (Cochrane et al., 2020),blood pressure (Kario et al., 2000),cholesterol measurement (Schectman and Hoffmann, 1988) etc. Since the RTM was discovered to be the likely cause of the treatment effect's effectiveness, it should be taken into account to prevent drawing incorrect conclusions.

1.1 Some Existing methods

Initially, strategies for dealing with (RTM) were developed under the assumption that the data followed a normal distribution. To estimate the regression to mean effect for bivariate normal distribution in uncontrolled clinical studies, (James, 1973) and (Gardner and Heady, 1973) used pre-post variables that had stationary mean and variance and were positively correlated. (Davis, 1976) developed a strategy to reduce the Regression to the Mean (RTM) effect. This method entailed measuring subjects multiple times before administering the treatment. Davis expanded on this approach by estimating RTM in scenarios where multiple measurements were taken prior to treatment application, providing a more complete understanding of the RTM phenomenon. (Shahane et al., 1995) expanded on previous research by addressing situations in which researchers may be interested in more than one variate both before and after treatment application. They developed a formula to estimate the regression to the mean (RTM) effect, used the same model as (Gardner and Heady, 1973) and (Johnson and George, 1991).

Not all data set follows a normal distribution . In response to this situation, (Das and Mulder, 1983), (John and Jawad, 2010), (Müller et al., 2003), and (Beath and Dobson, 1991) developed estimation methods for RTM. (Das and Mulder, 1983) developed a simple formula for regressing toward the mode in the case of arbitrary continuous measurements based on the assumption of pre-post variable stationarity. Their findings suggest that when data have a uni-modal distribution, the regression tends to align with the mode. In the context of non-normal populations, (Beath and Dobson, 1991) investigated two approximation methods (Edgeworth and saddlepoint) for estimating the regression to the mean (RTM). However, the Edgeworth approximation has limitations in that it is not applicable for all values of skewness and kurtosis, and it has the potential to produce negative or multi-modal results (Barton and Dennis, 1952). As a result, the Saddlepoint approximation was chosen due to its property of always being positive, but it is more computationally difficult. To adapt (Das and Mulder, 1983) method to empirical distributions, (John and Jawad, 2010) investigated density kernel estimation approaches while maintaining the same error component assumptions.

Not all variables in a study have continuous characteristics in practice; some are discrete, representing count data or binary responses governed by discrete probability distributions. (Khan and Olivier, 2018, 2019) addressed this issue by developing formulas to estimate the RTM effect for such discrete data types. Their interest extends to scenarios in which variables represent counts or events. They proposed methods for dealing with the RTM effect in these discrete contexts, based on bivariate Poisson and binomial distributions.

1.2 Objective of the thesis

RTM occurs in all disciplines and must be accounted for to unbiasedly estimate an intervention effect. The aim of this thesis is to propose and estimate RTM effect within count models, particularly addressing positive correlations between pre-post count variables.

1.3 Structure of thesis

In order to fulfill the aforementioned objectives, the structure of this thesis is outlined as follows. In chapter 2 the literature of existing methods for quantifying the RTM effect is discussed. Chapter 3 presents the derivation of the regression to mean (RTM) formula for bivariate generalized Poisson-Lindley (BGPL) distribution as introduced by (Aryuyuen and Bodhisuwan, 2023). This formulation accounts for both positive and negative correlations within the distribution. Chapter 4 contain the conclusion.

Chapter 2

Literature Review

Formulas for Regression to the Mean (RTM) have been developed by researchers to estimate its effects under various situations. While conducting pre and post studies, some researchers employ various methods to mitigate the RTM effect. RTM is observed when measurements are taken on the same subject at multiple points in time (Barnett et al., 2005).The details of existing literature is discussed below.

2.1 RTM Effect Under Bivariate Normal Disribution

Originally, approaches to address RTM were formulated under the assumption of a normal distribution. They are briefly outlined below.

2.1.1 James's Method

In clinical studies, (James, 1973) conducted early research on the RTM effect in the context of bivariate normal distribution. He claimed that the observed variable is made up of true and random error components. The author proposed this relationship by denoting X_i as the effect size on the i-th measurement of the same subject, X_0 as the true measurement and e_i as the error.

$$X_i = X_0 + e_i$$

The true component X_0 follows a normal distribution with a mean μ and a variance σ_0^2. The error components are independent, identically distributed, and have a normal distribution with a mean of 0 and a variance of σ_e^2. Both X_0 and the error components are mutually independent. Consequently, the observed variable X_i is normally distributed with a mean μ and a variance σ^2, where $\sigma^2 = \sigma_0^2 + \sigma_e^2$.

The correlation between X_i and X_j is denoted as ρ, with a value of σ_0^2/σ^2, and it is greater than 0.

For a right cut-off x_0 for both below and above measurements, the joint probability density function (PDF) of X_1 and X_2 is given by:

$$f(X_1, X_2 \mid X_1 > X_0) = \frac{\exp\left(-\frac{1}{1-\rho^2}\left[\frac{(X_1-\mu)^2}{\sigma^2} + \frac{(X_2-\mu)^2}{\sigma^2} - 2\rho\left(\frac{X_1-\mu}{\sigma}\right)\left(\frac{X_2-\mu}{\sigma}\right)\right]\right)}{(1-\Phi(z_0))\sigma^2\sqrt{1-\rho^2}}$$

Consider random variables X_1 and X_2, where $X_0 < X_1 < \infty$ and $-\infty < X_2 < \infty$. where $\Phi(\cdot)$ denote the cumulative distribution function (CDF) of the standard normal distribution. Define $z_0 = (X_0 - \mu)/\sigma$, where μ is the mean and σ is the standard deviation. The mean and variance of the distribution of the difference $d = E(X_2 - X_1 \mid X_1 > X_0)$, truncated such that $x_1 \geq x_0$, were derived as follows:

$$\mu_d = \frac{\phi(z_0)}{1-\Phi(z_0)}\sigma(\rho - 1)$$

and

$$\sigma_d = (1-\rho)\left[\frac{\phi(z_0)}{1-\Phi(z_0)}(z_0 - \frac{\phi(z_0)}{1-\Phi(z_0)})(1-\rho) + 2\right]\sigma^2$$

James (1973) developed the RTM effect under bivariate normality, which is defined as the difference between the conditional means of two identically distributed variables, x_1 and x_2.

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

Here, $\phi(z_0)$ represents the probability density function (PDF) of the standard normal distribution. The total proportion reduction (TPR) in the mean of X_1, which is attributed to both (RTM) and the treatment effect, is defined as,

$$TPR = \frac{x_1 - \gamma\rho x_1}{x_1} = 1 - \gamma\rho$$

The proportional reduction observed in the mean difference is solely due to regression to the mean (RTM) and is expressed as,

$$\text{Proportion of Reduction due to RTM} = 1 - \rho$$

James (1973) used the moments method of estimation to compute parameter estimates from sample data.

$$\hat{\mu} = \bar{x}_1 - l_0\hat{\sigma}$$
$$S_{x1}^2 = \hat{\sigma}(l_0(x_0 - l_0) + 1)$$
$$S_{x2}^2 = \hat{\sigma}\left(\hat{\gamma}^2\hat{\rho}^2 l_0(x_0 - l_0) + \hat{\gamma}^2\hat{\rho}^2 + (1 - \hat{\rho}^2)\right)$$

$$\hat{\rho} = \left(b^2(l_0(x_0 - l_0) + 1) - \frac{S_{x1}^2}{\hat{\sigma}^2} + 1\right)^{\frac{1}{2}}$$

where

$$l_0 = \frac{\phi(z_0)}{1 - \Phi(z_0)}$$
$$\hat{\gamma} = \frac{b}{\hat{\rho}}$$

James (1973) demonstrated that when there is a weak correlation between pre and post treatment measurements, the magnitude of the RTM effect increases, emphasizing the importance of the control group. Furthermore, when dealing with non-normally distributed data, the robustness of the estimate may be compromised, necessitating a recommendation for further investigation.

2.1.2 Gardner's Method

Like (James, 1973), (Gardner and Heady, 1973), who focused on deriving the RTM effect, the authors extended their exploration beyond bivariate measures to investigate the impact of multiple measurements on RTM. The authors made the assumption that the pre-post variables follow a normal distribution with a correlation coefficient (ρ) equal to $\sigma_0/\sqrt{\sigma_0^2 + \sigma_1^2}$. The subjects, chosen based on the right cut-off point, i.e., $X_i > x_0$, exhibit a univariate truncated normal distribution with a mean;

$$E(X_i|X_i > x_0) = \mu + \sigma v\left(\frac{x_0 - \mu}{\sigma}\right)$$

Where $\nu(z_0) = \phi(\cdot)/1 - \Phi(\cdot)$. The truncated distribution and conditional expectation of X_i given $X_i > x_0$ is,

$$f(X_i|X_i > x_0) = \frac{\int_{x_0}^{\infty} f(x_i, x_0)\, dx_i}{\int_{x_0}^{\infty} f(x_i)\, dx_i}$$

$$E(X_0|X_i > x_0) = \mu + \frac{\sigma_0^2}{\sigma^2} \cdot \nu(z_0)$$

Comparing $E(X_i|X_i > x_0)$ and $E(X_0|X_i > x_0)$, it becomes evident that the mean of observed values will consistently exceed the mean of true values unless the variance of the error term equals zero. For multiple observations on the same subject, (Gardner and Heady, 1973) derived the formula of RTM effect as,

$$R(x_0) = \sqrt{\sigma_0^2 + \frac{\sigma_e^2}{n}} - \frac{\sigma_0^2}{\sqrt{\sigma_0^2 + \frac{\sigma_e^2}{n}}} \cdot \nu(z_{0n})$$

$$= \frac{\frac{\sigma_e^2}{n}}{\sqrt{\sigma_0^2 + \frac{\sigma_e^2}{n}}} \cdot \nu(z_{0n}).$$

and the formula for $n = 1$ reduces to the RTM equation derived by (James, 1973).

2.1.3 Johnson's Method

Johnson and George (1991) adopted the model initially proposed by (Gardner and Heady, 1973), wherein the variables x_1 and x_2 represent the pre and post observations, respectively, characterized by a true value X_0 and an error component e . Gardner and Heady (1973) posited the assumption that variability arises solely from independent measurement errors. However, in practical situations, this assumption may not hold true, and additional sources of variability could be present in the measurements. In a real-world context, the measurement of individuals' blood pressure, which exhibits constant fluctuations over time, is influenced by various factors such as the individual's emotional state and other variables. This introduces within-subject variability. Johnson and George (1991) integrated this factor into the model by,

$$X_i = X_0 + S_i + D_i \quad \text{for } i = 1, 2, \ldots, m$$

where X_i and X_0 represent the observed and true values, respectively. S_i denotes the subject effect, which has a normal distribution with a mean of zero and a variance of σ_s^2. Importantly, S_i is unaffected by either X_0 or D_i. The correlation between S_i and S_j is denoted as ρ_s and is equal to a positive value, where $i \neq j$. Then, the correlation and RTM formulae are

$$R(x_0) = \frac{\sigma_0^2 + (1 - \rho_s)\sigma_s^2}{\sqrt{\sigma_0^2 + \sigma_s^2 + \sigma_e^2}} \cdot \nu(z_1)$$

and

$$\mathrm{cor}(X_1, X_2) = \frac{\sigma_0^2 + \rho_s\sigma_s^2}{\sigma_0^2 + \sigma_s^2 + \sigma_e^2}$$

where $\nu(z_1) = \phi(z_1)/1 - \Phi(z_1)$, and $z_1 = x_0 - \mu/\sqrt{\sigma_0^2 + \sigma_s^2 + \sigma_e^2}$.

Suppose repeated measurements of n replicates at m different times are taken, then,

$$X_{ij} = X_0 + S_i + D_{ij}, \quad \text{for } i = 1, 2, \ldots, m, \quad j = 1, 2, \ldots, n$$

where,

$$X_0 \sim \mathcal{N}(\mu, \sigma_0^2), \quad (S_1, \ldots, S_m) \sim \mathcal{N}(\mathbf{0}, \Sigma), \quad \Sigma \text{ is a symmetric matrix}$$

Let $\bar{X}$ be the sample mean calculated as $\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} x_{ij}$, representing the mean of individual observations X_{ij} from an experiment involving m subjects, each contributing n observations. Assume this mean is calculated for values exceeding a truncated point x_0. The variance of the sample mean, denoted as $\mathrm{Var}(\bar{X})$, is determined by the formula:

$$\mathrm{Var}(\bar{X}) = \sigma_0^2 + \frac{\sigma_s^2}{m}(1 + (m-1)\rho_s) + \frac{\sigma_e^2}{mn}$$

If a new observation X^* is taken after the treatment, the correlation between this new observation and the sample mean can be expressed as follows:

$$\mathrm{cor}(\bar{X}, X^*) = \frac{\sigma_0^2 + \rho_s\sigma_s^2}{\left(\sigma_0^2 + \frac{\sigma_s^2}{m}(1 + (m-1)\rho_s) + \frac{\sigma_e^2}{mn}\right)^{1/2}} \cdot \frac{1}{(\sigma_0^2 + \sigma_s^2 + \sigma_e^2)^{1/2}}$$

and the total RTM effect is expressed as follows

$$R_T(x_0) = \frac{(1-\rho_s)\sigma_s^2 + \sigma_e^2/n}{m \cdot \mathrm{var}(\bar{X})} \cdot \upsilon(z_2)$$

The function $\upsilon(z_2)$ is defined as $\phi(z_2)/1 - \Phi(z_2)$, where $z_2 = x_0 - \mu/\mathrm{var}(\bar{X})$, and "T" in the subscript denotes total. When decomposing the total RTM effect, it is equal to the sum of RTM due to subject and measurement error.

$$R_T(x_0) = R_s(x_0) + R_e(x_0)$$

$$R_s(x_0) = (1-\rho_s)\frac{\sigma_s^2}{\mathrm{var}(\bar{X})} \cdot \nu(z_2)$$

$$R_e(x_0) = \frac{\sigma_e^2}{mn \cdot \mathrm{var}(\bar{X})} \cdot \nu(z_2)$$

Because of the independent error, raising the number of repeated measurements and/or replications could help to reduce the RTM effect. Furthermore, by increasing the number of measurements at different time points, the RTM of subject variability can be reduced, and the RTM of measurement error can be reduced by replicating the measurement at a given time point.

2.1.4 Shahane's Approach

In the preceding literature, the authors examined the RTM effect for a single variable that had been truncated at a specific cut-off point. (Shahane et al., 1995) examined the scenario in which truncation occurred for two variables. For example, a researcher may be interested in studying the IQ level of second generation immigrants by drawing a sample from first generation parents whose IQ exceeds the truncated point. Let X = (X_1, X_2, Y_1, Y_2) be the random vector from the multivariate normal distribution (MND). Subsequently, the mean, variances, and correlations can be expressed as follows:

$$\mu = (\mu_1, \mu_1 - \delta_1, \mu_2, \mu_2 - \delta_2), \sigma_x^2 = \text{var}(X_i), \sigma_y^2 = \text{var}(Y_i)$$

$$\rho_{XX} = \text{corr}(X_1, X_2), \rho_{yy} = \text{corr}(y_1, y_2), \rho_{xY} = \text{corr}(X_i, y_i), \text{ for } i, j = 1, 2$$

In clinical trial studies, let (X_1, Y_1) represent the measurements before treatment application, and (X_2, Y_2) represent the measurements after treatment application. (Tallis, 1961) used the moment generating function for the bivariate normal distribution to derive expressions for expected values of truncated variables. The expression for the truncated RTM effect on M is as follows:

$$E(X_1 - X_2|M) = \delta_1 + \frac{\sigma_X}{P(M)}\left((1 - \rho_{XX})\phi(\alpha)\Phi(A)\right),$$

$$E(Y_1 - Y_2|M) = \delta_2 + \frac{\sigma_Y}{P(M)}\left((1 - \rho_{YY})\phi(\beta)\Phi(B)\right),$$

where $M = [X_1 > m_1 \cup Y_1 > m_2]$, $\Phi(w) = 1 - \int_w^\infty \phi(u)\,du$, $\alpha = m_1 - \mu_1/\sigma_X$, $\beta = m_2 - \mu_2/\sigma_Y$, $A = \beta - \rho_{XY}\alpha/\sqrt{1 - \rho_{XY}^2}$, and $B = \alpha - \rho_{XY}\beta/\sqrt{1 - \rho_{XY}^2}$.

The equations above indicate that the conditional difference can be expressed as the sum of the true treatment effect and the expected effect for each variable or variate. The regression effect is zero when the correlation is 1 and reaches its maximum when the correlation is 0. Within the framework of the (Gardner and Heady, 1973) model assumptions, the RTM effect can be expressed in a simplified form as follows:

$$R(X_2|M) = \frac{\sigma_{e_1}\phi(\alpha)\Phi(A)}{\sigma_X P(M)}$$

$$R(Y_2|M) = \frac{\sigma_{e_2}\phi(\beta)\Phi(B)}{\sigma_Y P(M)}$$

When utilizing the averages of n replicates and subsequently obtaining the $(n+1)$th measurement after the treatment, the RTM effects can be characterized as:

$$R(X_{n+1}|M_n) = \frac{\sigma_{e_1}\phi(\alpha)\Phi(A)}{n\sigma_{\bar{X}} P(M_n)},$$

$$R(Y_{n+1}|M_n) = \frac{\sigma_{e_2}\phi(\beta)\Phi(B)}{n\sigma_{\bar{Y}} P(M_n)},$$

where $M_n = [\overline{X}_n > m_1 \cup \overline{Y}_n > m_2]$, $\alpha = m_1 - \mu_1/\sigma_{\overline{X}}$, and $\beta = m_2 - \mu_2/\sigma_{\overline{Y}}$, $A = \beta - \rho_{\bar{X}\bar{Y}}\alpha/\sqrt{1-\rho^2_{\bar{X}\bar{Y}}}$, $B = \alpha - \rho_{\bar{X}\bar{Y}}\beta/\sqrt{1-\rho^2_{\bar{X}\bar{Y}}}$, $\sigma_{\bar{X}} = \sigma^2_{u1} + \sigma^2_{e1}/n$, $\sigma_{\bar{Y}} = \sigma^2_{u2} + \sigma^2_{e2}/n$ and $\rho_{\bar{X}\bar{Y}} = \rho_u\sigma_{u1}\sigma_{u2}/\sigma_{\bar{X}}\sigma_{\bar{Y}}$.

Shahane et al. (1995) demonstrated that by maximizing replicates, it is possible to minimize RTM caused by within-subject variability. Furthermore, by raising the number of replicates and the frequency of repeated measurements n, the RTM effect due to measurement error can be minimized.

2.2 RTM for bivariate discrete distributions

Recent research has focused on deriving and estimating RTM effect in the context of bivariate discrete distributions. The key findings are summarized below:

2.2.1 Bivariate Poisson

The Poisson distribution is widely used in real-world scenarios, particularly when dealing with variables that represent counts of specific characteristics of interest. Furthermore, the Poisson distribution can be used to approximate a wide range of other probability distributions. (Khan and Olivier, 2018) derived the expression for the RTM effect in the context of the bivariate Poisson distribution. They consider the successive random variables for the same subject X_1 and X_2 as

$$X_1 = Y_0 + Y_1, \quad X_2 = Y_0 + Y_2$$

If X_i follows a Poisson distribution with parameter $\alpha_0 + \alpha_i$ for $i = 1, 2$, where Y_0 represents the true number of occurrences and Y_1 and Y_2 denote the counting errors, and further, Y_0, Y_1, and Y_2 are independent with rates of occurrence α_i for $i = 0, 1, 2$, (Khan and Olivier, 2018) utilized the bivariate Poisson distribution. This distribution, originally discussed by (Campbell, 1934), is given by

$$P(X_1 = x_1, X_2 = x_2) = e^{-(\alpha_0+\alpha_1+\alpha_2)}\tfrac{\alpha_1^{x_1}}{x_1!}\tfrac{\alpha_2^{x_2}}{x_2!}\textstyle\sum_{y_0=0}^{\min(x_1,x_2)} y_0!\left(\tfrac{\alpha_0}{\alpha_1\alpha_2}\right)^{y_0}\tbinom{x_1}{x_0}\tbinom{x_2}{x_0}$$

The covariance between X_1 and X_2 is denoted as $\mathrm{cov}(X_1, X_2) = \alpha_0$, then the correlation $\mathrm{cor}(X_1, X_2)$ can be expressed as,

$$\mathrm{cor}(X_1, X_2) = \frac{\alpha_0}{\sqrt{(\alpha_0 + \alpha_1)(\alpha_0 + \alpha_2)}}$$

(Khan and Olivier, 2018) derived formulae for the total effect considering cut-off points in both the right and left tail of the distribution as,

$$\mathrm{Tr}(x_0; \alpha) = \alpha_1 \left(1 - F\left(\frac{x_0 - 1}{\alpha_0 + \alpha_1}\right) / \left(1 - F\left(\frac{x_0}{\alpha_0 + \alpha_1}\right)\right)\right) - \alpha_2$$

$$\mathrm{T}_l(x_0; \alpha) = \alpha_2 - \alpha_1 \left(F\left(\frac{x_0 - 1}{\alpha_0 + \alpha_1}\right) / F\left(\frac{x_0}{\alpha_0 + \alpha_1}\right)\right)$$

The recursive relationship proposed by (Teicher, 1954) is employed for solving systems of equations. The Maximum Likelihood Estimators (MLE) for the right and left (RTM) are provided as follows,

$$\hat{R}_r(x_0, x) = \bar{x}_1 \,|x_1 > x_0 - \bar{x}_2 \,|x_1 > x_0,$$

$$\hat{R}_l(x_0, x) = \bar{x}^2 |x_1 \leq x_0 - \bar{x}_1 |x_1 \geq x_0,$$

The average intervention or treatment effect is mathematically defined as the expected difference between successive variables.

$$\delta(\alpha) = \alpha_1 - \alpha_2$$

When considering the impact of a null intervention, the before-after observations are distributed identically. By letting $\alpha_1 = \alpha_2$, an expression for the RTM effect can be obtained.

$$\mathrm{R}_r(x_0; \alpha) = \alpha_1 \left(1 - F\left(\frac{x_0 - 1}{\alpha_0 + \alpha_1}\right) / \left(1 - F\left(\frac{x_0}{\alpha_0 + \alpha_1}\right)\right)\right) - \alpha_1$$

The total effect is the sum of the RTM and intervention/treatment effects, expressed as:

$$T(x_0; \alpha) = R_r(x_0; \alpha) + \delta(\alpha)$$

$$T(x_0; \alpha) = (\alpha_1 \left(1 - F\left(\frac{x_0 - 1}{\alpha_0 + \alpha_1}\right) / \left(1 - F\left(\frac{x_0}{\alpha_0 + \alpha_1}\right)\right)\right) - \alpha_1) + (\alpha_1 - \alpha_2)$$

(Khan and Olivier, 2018) conducted a simulation study to evaluate properties and derivation, comparing the true RTM effect with its estimated RTM. Distinct behavior was observed in both homogeneous and in-homogeneous RTM effects.

2.2.2 Bivariate Poisson Yousaf et al, (2023)

Yousaf et al, (2023) formulated an expression for the RTM effect in the context of count data, specifically Poisson data, in their study. They consider λ_1 and λ_2 represent the Poisson parameters, i.e., the mean occurrence rate, $c = 1 - e^{-1}$, and θ is a known constant ensuring $f(x_1, x_2)$ is non-negative for all $x_1, x_2 \geq 0$. Yousaf et al, (2023) employed the bivariate Poisson distribution, initially introduced by (Lakshminarayana et al., 1999) and defined as,

$$P(x_1, x_2) = \frac{e^{-\lambda_1}\lambda_1^{x_1}}{x_1!} \cdot \frac{e^{-\lambda_2}\lambda_2^{x_2}}{x_2!} \cdot \left\{1 + \theta(e^{-x_1} - e^{-\lambda_1 c})(e^{-x_2} - e^{-\lambda_2 c})\right\}$$

$$x_1, x_2 = 0, 1, 2, \ldots, \infty, \quad \lambda_1, \lambda_2 > 0$$

The correlation coefficient between x_1 and x_2 is determined by

$$\rho = \theta\sqrt{\lambda_1\lambda_2} c^2 e^{-c(\lambda_1+\lambda_2)}$$

The correlation can be positive or negative depending on the choice of θ. The θ must lie in the range

$$|\theta| \leq \frac{1}{(1 - e^{-\lambda_1 c})(1 - e^{-\lambda_2 c})}$$

In their research, Yousaf et al.(2023) derived formulas for the total effect, incorporating cut-off points in both the right and left tail of the distribution.

$$\mathrm{Tr}(x_0; \lambda) = \frac{\lambda_1\left[1 - F(x_0 - 1|\lambda_1)\right] + e^{-c\lambda_2}\lambda_2 c\theta D}{1 - F(x_0|\lambda_1)} - \lambda_2\left[1 + \theta c e^{-c(\lambda_1+\lambda_2)}\right],$$

where,

$$D = \sum_{x_1=x_0+1}^{\infty} \frac{e^{-\lambda_1}\lambda_1^{x_1}e^{-x_1}}{x_1!},$$

$$T_l(x_0; \lambda) = \lambda_2\left[1 + \theta c e^{-c(\lambda_1+\lambda_2)}\right] - \frac{e^{-c\lambda_2}\lambda_2 c\theta E - \lambda_1 F(x_0 - 1/\lambda_1)}{F(x_0/\lambda_1)},$$

where,

$$E = \sum_{x_1=0}^{x_0} \frac{e^{-\lambda_1}\lambda_1^{x_1}e^{-x_1-1}}{x_1!},$$

the expected difference between the pre and post observations or subject measurements. Its can be written as,

$$\Lambda(\lambda) = \lambda_1 - \lambda_2$$

For a null intervention effect, the measurements before and after the study follow identical distributions, with equivalent conditional means when λ_1 is set equal to

λ_2. Expressing this condition provides a formulation for the RTM effect,

$$Rr(x_0;\lambda) = \frac{\lambda_1\left[1 - F(x_0 - 1|\lambda_1)\right] + e^{-c\lambda_1}\lambda_1 c\theta D}{1 - F(x_0|\lambda_1)} - \lambda_1\left[1 + \theta c e^{-2c\lambda_1}\right],$$

For a non-null intervention effect, the pre and post measurements are not identically distributed.To achieve a non-null intervention effect, the RTM can be derived as follows:

$$Rr(x_0;\lambda) = Tr(x_0;\lambda) - \Lambda_r(\lambda)$$

$$Rr(x_0;\lambda) = \frac{\lambda_1\left[1 - F(x_0 - 1|\lambda_1)\right] + e^{-c\lambda_2}\lambda_2 c\theta D}{1 - F(x_0|\lambda_1)} - \lambda_2\left[1 + \theta c e^{-c(\lambda_1+\lambda_2)}\right] - (\lambda_1 - \lambda_2)$$

The total effect is determined by the sum of RTM and intervention/treatment effects, and it is expressed as:

$$T(x_0;\lambda) = Rr(x_0;\lambda) + \Lambda_r(\lambda)$$

$$= \frac{\lambda_1\left[1 - F(x_0 - 1|\lambda_1)\right] + e^{-c\lambda_2}\lambda_2 c\theta D}{1 - F(x_0|\lambda_1)} - \lambda_2\left[1 + \theta c e^{-c(\lambda_1+\lambda_2)}\right] - (\lambda_1 - \lambda_2) + (\lambda_1 - \lambda_2)$$

Yousaf et al, (2023) evaluated the properties and derivation through simulation study and compared the proposed and existing RTM on different sample size.

Chapter 3

RTM under The Bivariate Generalized Poisson Lindley Distribution

The bivariate Poisson distributions and processes are used to model count data. In literature, authors suggested different types of bivariate Poisson distribution. It is worth noticing that the form they reported does not account for the positive correlation. In this chapter, RTM formulae are derived for the bivariate generalized Poisson lindley model that generalizes the count variables for all range of the correlation coefficient.

3.1 Bivariate Generalized Poisson Lindley Distribution

Aryuyuen and Bodhisuwan (2023) developed a bivariate generalized Poisson lindley distribution as a product of marginal generalized Poisson lindley distribution with a multiplicative factor. The correlation of bivariate genralized Poisson lindley distribution which can be either, positive, zero or negative depending upon the value of multiplicative factor parameter. The probability mass function (PMF) of a bivariate generalized Poisson Lindley distribution is typically expressed as follows,

$$\begin{aligned} f(x_1,x_2) = &\frac{\alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}} \cdot \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}} \\ &\cdot \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right], \quad x_j = 0,1,2,\ldots \end{aligned} \tag{3.1}$$

where $\alpha_j > 0$, $\beta_j > 0$ for $j = 1, 2$, $-\infty < \theta < \infty$, and

$$m_j = \frac{\alpha_j^2(\beta_j + \alpha_j - e^{-1} + 1)}{(\alpha_j + \beta_j)(\alpha_j - e^{-1} + 1)^2} \quad j = 1, 2$$

The BGPL distribution is characterized by five parameters: β_1 and β_2 serve as shape parameters, while α_1, α_2, and θ act as scale parameters. The marginal pmf of X_1 and X_2 are the univariate generalized Poisson lindley with parameter α_j and β_j, respectively, for j=1,2 The respective mean and variance, of the univariate generalized Poisson lindley distribution are

$$E(X_j) = \frac{\alpha_j + 2\beta_j}{\alpha_j(\beta_j + \alpha_j)} = \mu_j,$$

$$Var(X_j) = \frac{2\beta_j^2(1 + \alpha_j) + \alpha_j^2(1 + \alpha_j) + \beta_j\alpha_j(4 + 3\alpha_j)}{\alpha_j^2(\beta_j + \alpha_j)^2} = \sigma_j^2,$$

The covariance of bivariate generalized Poisson lindley distribution is

$$\text{Cov}(X_1, X_2) = \theta(m_{11} - \mu_1 m_1)(m_{22} - \mu_2 m_2) = \sigma_{1,2}$$

where,

$$m_{jj} = \frac{\alpha_j^2(\alpha_j + 2\beta_j - e^{-1} + 1)e^{-1}}{(\alpha_j + \beta_j)(\alpha_j - e^{-1} + 1)^3}, \quad j = 1, 2$$

Therefore, the correlation coefficient for X_1 and X_2 is

$$\rho_{X_1,X_2} = \frac{\sigma_{1,2}}{\sqrt{\sigma_1^2\sigma_2^2}} = \frac{\theta(m_{11} - \mu_1 m_1)(m_{22} - \mu_2 m_2)}{\sigma_1\sigma_2} \tag{3.2}$$

When $\theta = 0$, random variables X_1 and X_2 are independent. For $\theta > 0$, the variables X_1 and X_2 exhibit a positive correlation, while for $\theta < 0$, they demonstrate a negative correlation.

3.2 RTM, total, and treatment effects

In medical, clinical, or intervention studies, measurements at some specified truncation or cut-off points y_0 are selected for treatment or intervention. Let Y_1 and Y_2 be the counts of a characteristic of interest before and after the application of a treatment. Then, the joint distribution of Y_1 and Y_2 at truncation point y_0 is given by,

$$f_T(Y_1, Y_2) = \frac{f(Y_1, Y_2)}{f(Y_1 > y_0)} \text{ where } y_0 < Y_1 < \infty, \quad -\infty < Y_2 < \infty$$

In the context of $f_T(Y_1, Y_2)$, where the subscript t represents truncation, the total effect $T(y_0, \theta)$ is defined as the conditional expectation of the difference between pre- post treatment variables. Mathematically, it is expressed as:

$$\begin{aligned} T(y_0, \theta) &= E[Y_1 - Y_2 | Y_1 > y_0, \theta] \\ &= \int_{y_0}^{\infty} \int_{-\infty}^{\infty} (Y_1 - Y_2) f(Y_1, Y_2 | Y_1 > y_0)\, dy_2\, dy_1 \end{aligned} \tag{3.3}$$

In equation 3.3, θ represents the vector of parameters, $T(y_0, \theta)$ can be obtained for a bivariate discrete distribution by replacing integrals with summations.

The total effect, $T(y_0, \theta)$, is either totally or partially due to RTM, depending the effectiveness of treatment effect. When both Y_1 and Y_2 are identically distributed, the treatment effect is zero, i.e., $E(Y_1 - Y_2) = 0$. In such cases, the conditional expectation of the difference between Y_1 and Y_2 is defined as the RTM effect and is given by:

$$R(y_0, \theta) = E(Y_1 - Y_2 | Y_1 > y_0, E(Y_1) = E(Y_2)) \tag{3.4}$$

The treatment effect is defined as the difference between the unconditional means of Y_1 and Y_2, and is given by

$$\delta(\lambda) = E(Y_1 - Y_2)$$

Hence, the total effect $T(y_0, \theta)$ can be expressed as:

$$T(y_0, \theta) = R(y_0, \theta) + \delta(\lambda),$$

Here, "λ" represents a function of the means of Y_1 and Y_2.

Researchers have introduced a range of expressions to estimate the effects of RTM and intervention, as discussed in of the literature. Up until 2018, these expressions were primarily grounded in normality or relied on approximate methods for non-normal data. In 2018, Khan and Olivier (2018) proposed expressions for estimating the RTM effect specifically for Poisson data. The probability mass function (pmf) utilized by Khan and Olivier (2018) is rooted in a trivariate reduction technique, which presupposes that bivariate Poisson data should be equi-dispersed, and variables should exhibit positive correlations.

It's worth noting that truncation is specifically applied to the pre-measurements of the same variable. This is done to assess the treatment and RTM effects on the post-variables, whether they are above or below the cut-off point.

3.2.1 Total effect derivation with a right cut-off point

Suppose an intervention is applied to subjects under the condition that their initial count, denoted as X_1, exceeds a specified cut-off point, say x_0, the corresponding joint truncated distribution is as follows

$$P_t(X_1 = x_1, X_2 = x_2 \mid X_1 > x_0) = \frac{1}{1 - P(X_1 \leq x_0)} \cdot \frac{\alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}} \cdot \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}} \cdot \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right] \tag{3.5}$$

To find the total effect, we start by determining the conditional expectation as given in equation 3.3, X_1 given $X_1 > x_0$.

$$E(X_1 \mid X_1 > x_0) = \frac{1}{1 - P(X_1 \leq x_0)} \sum_{x_1=x_0+1}^{\infty} \sum_{x_2=0}^{\infty} \frac{x_1 \cdot \alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}} \cdot \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}} \cdot \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right]$$

After some algebraic manipulation, the expectation simplifies to

$$\begin{aligned} E(X_1 \mid X_1 > x_0) = {} & \frac{1}{1 - F(x_0|\alpha_1, \beta_1) \cdot \alpha_1(\alpha_1+\beta_1)(1+\alpha_1)^{x_0+2}} \cdot [(\alpha_1+2\beta_1)(\alpha_1+1)^{x_0+2} \\ & - \left(-\alpha_1^2 x_0 - \alpha_1(1+\alpha_1) + \alpha_1^2(\alpha_1+1)^{x_0} + \alpha_1(1+\alpha_1)^{x_0}\right) \cdot (1+\alpha_1+\beta_1) \\ & - \beta_1\alpha_1^2 x_0^2 + (-2\alpha_1\beta_1 x_0 - \alpha_1\beta_1 - 2\beta_1) \cdot (1+\alpha_1) + \left(\alpha_1^2\beta_1 + 3\alpha_1\beta_1 + 2\beta_1\right) \\ & \cdot (1+\alpha_1)^{x_0}] \end{aligned} \tag{3.6}$$

where

$$F(x_0|\alpha_1, \beta_1) = \frac{(\beta_1+\alpha_1)(\alpha_1+1)^{x_0+2} - (2\alpha_1\beta_1 + \beta_1 + \alpha_1^2 + \alpha_1 + \alpha_1\beta_1 x_0)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_0+2}},$$

is the cumulative distribution function (CDF) of the univariate generalized Poisson lindley distribution. Now considering the conditional expectation of $X_2 \mid X_1 > x_0$ as

$$E(X_2 \mid X_1 > x_0) = \frac{1}{1 - P(X_1 \leq x_0)} \sum_{x_1=x_0+1}^{\infty} \sum_{x_2=0}^{\infty} \frac{x_2 \cdot \alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}} \cdot \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}} \cdot \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right]$$

After solving for $E(X_2 \mid X_1 > x_0)$, the expression is found to be

$$\begin{aligned} E(X_2 \mid X_1 > x_0) = \frac{1}{1 - F(x_0|\alpha_1, \beta_1)}&[(1 - F(x_0|\alpha_1, \beta_1)) \cdot \mu_2 \\ &+ \theta \cdot (C - m_1 \cdot (1 - F(x_0|\alpha_1, \beta_1))) \cdot (m_{22} - m_2 \cdot \mu_2)] \end{aligned} \tag{3.7}$$

where

$$C = \sum_{x_1=x_0+1}^{\infty} \frac{e^{-x_1} \cdot \alpha_1^2(1 + \alpha_1 + \beta_1 + \beta_1 x_1)}{(\alpha_1 + \beta_1)(1 + \alpha_1)^{x_1+2}}$$

To get the total effect for the right cut-off point under the bivariate genralized poisson lindley distribution, substituting the results of 3.6 and 3.7 in equation 3.3.

$$\begin{aligned} Tr(x_0; \alpha, \beta) = D - \frac{1}{1 - F(x_0|\alpha_1, \beta_1)}&[(1 - F(x_0|\alpha_1, \beta_1)) \cdot \mu_2 \\ &+ \theta \cdot (C - m_1 \cdot (1 - F(x_0|\alpha_1, \beta_1))) \cdot (m_{22} - m_2 \cdot \mu_2)] \end{aligned} \tag{3.8}$$

The D is defined as the conditional expectation of X_1 given $x_1 > x_0$, and this is represented by the equation 3.6. And the subscript r stands for the right cutt-off point.

3.2.2 Total effect derivation with a left cut-off point

When the subjects selected for treatment fall in the left tail or end of a distribution, meaning $X_1 \leq x_0$, the truncated probability distribution function of X_1 and X_2 becomes,

$$\begin{aligned} P(X_1 = x_1, X_2 = x_2 \mid X_1 \leq x_0) = \frac{1}{P(X_1 \leq x_0)} \cdot \frac{\alpha_1^2(1 + \alpha_1 + \beta_1 + \beta_1 x_1)}{(\alpha_1 + \beta_1)(1 + \alpha_1)^{x_1+2}} \\ \cdot \frac{\alpha_2^2(1 + \alpha_2 + \beta_2 + \beta_2 x_2)}{(\alpha_2 + \beta_2)(1 + \alpha_2)^{x_2+2}} \cdot \left[1 + \theta(e^{-x_1} - m_1)(e^{-x_2} - m_2)\right] \end{aligned} \tag{3.9}$$

The conditional expectation of X_1 given the event $X_1 \leq x_0$ is,

$$\begin{aligned} E(X_1 \mid X_1 \leq x_0) = \frac{1}{P(X_1 \leq x_0)} \sum_{x_1=0}^{x_0} \sum_{x_2=0}^{\infty} \frac{x_1 \cdot \alpha_1^2(1 + \alpha_1 + \beta_1 + \beta_1 x_1)}{(\alpha_1 + \beta_1)(1 + \alpha_1)^{x_1+2}} \\ \cdot \frac{\alpha_2^2(1 + \alpha_2 + \beta_2 + \beta_2 x_2)}{(\alpha_2 + \beta_2)(1 + \alpha_2)^{x_2+2}} \cdot \left[1 + \theta(e^{-x_1} - m_1)(e^{-x_2} - m_2)\right] \end{aligned}$$

Following similar steps as for the right cut-off point, the conditional expectation upon simplification becomes

$$
\begin{aligned}
E(X_1 \mid X_1 \leq x_0) = &\frac{1}{F(x_0|\alpha_1, \beta_1) \cdot \alpha_1(\alpha_1+\beta_1)(1+\alpha_1)^{x_0+2}} \\
&\times \left[(-\alpha_1^2 x_0 - \alpha_1(1+\alpha_1) + \alpha_1^2(\alpha_1+1)^{x_0} + \alpha_1(1+\alpha_1)^{x_0}\right) \cdot (1+\alpha_1+\beta_1) \\
&- \beta_1\alpha_1^2 x_0^2 + (-2\alpha_1\beta_1 x_0 - \alpha_1\beta_1 - 2\beta_1) \cdot (1+\alpha_1) \\
&+ \left(\alpha_1^2\beta_1 + 3\alpha_1\beta_1 + 2\beta_1\right) \cdot (1+\alpha_1)^{x_0}]. \qquad (3.10)
\end{aligned}
$$

For the conditional expectation of X_2 given $X_1 \leq x_0$, the expression is as follows

$$
\begin{aligned}
E(X_2 \mid X_1 \leq x_0) = &\frac{1}{P(X_1 \leq x_0)} \sum_{x_1=0}^{x_0} \sum_{x_2=0}^{\infty} \frac{x_2 \cdot \alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}} \\
&\cdot \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}} \cdot \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right]
\end{aligned}
$$

Upon solving for $E(X_2|X_1 \leq x_0)$, we obtain:

$$
\begin{aligned}
E(X_2|X_1 \leq x_0) = &\frac{1}{F(x_0|\alpha_1, \beta_1)}[F(x_0|\alpha_1, \beta_1) \cdot \mu_2 \\
&+ \theta \cdot (E - m_1 \cdot F(x_0|\alpha_1, \beta_1) \cdot (m_{22} - m_2 \cdot \mu_2)]
\end{aligned} \qquad (3.11)
$$

where

$$
E = \sum_{x_1=0}^{x_0} \frac{e^{-x_1} \cdot \alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}}
$$

The total effect at the left cut-off point is defined as

$$
T_l(x_0; \alpha, \beta) = E(X_2 - X_1|X_1 \leq x_0) = E(X_2|X_1 \leq x_0) - E(X_1|X_1 \leq x_0) \quad (3.12)
$$

The expression for $T_l(x_0; \alpha, \beta)$ at the left cut-off point, denoted by the subscript l, is obtained by subtracting equation 3.11 from equation 3.10

$$
\begin{aligned}
T_l(x_0; \alpha, \beta) = &\frac{1}{F(x_0|\alpha_1, \beta_1)}[F(x_0|\alpha_1, \beta_1) \cdot \mu_2 \\
&+ \theta \cdot (E - m_1 \cdot F(x_0|\alpha_1, \beta_1) \cdot (m_{22} - m_2 \cdot \mu_2)] - W
\end{aligned} \qquad (3.13)
$$

The variable W is defined as the conditional expectation of X_1 given $x_1 \leq x_0$, and this is represented by the equation 3.10.

3.3 Variance of total effect

To enable statistical inferences, it is essential to have an expression for the variance of total effect .The derivation of this variance involves combining the variances of pre-post measurements conditional on the cut-off point x_0 and considering the covariance, as outlined by the following formula

$$\mathrm{var}(X_1-X_2 \mid X_1 > x_0) = \mathrm{var}(X_1 \mid X_1 > x_0)+\mathrm{var}(X_2 \mid X_1 > x_0)-2\mathrm{cov}(X_1, X_2 \mid X_1 > x_0), \tag{3.14}$$

and

$$\mathrm{var}(X_2-X_1 \mid X_1 \leq x_0) = \mathrm{var}(X_1 \mid X_1 \leq x_0)+\mathrm{var}(X_2 \mid X_1 \leq x_0)-2\mathrm{cov}(X_1, X_2 \mid X_1 \leq x_0), \tag{3.15}$$

where

$$\mathrm{var}(X_1 \mid X_1 > x_0) = E(X_1(X_1-1) \mid X_1 > x_0)+E(X_1 \mid X_1 > x_0)-[E(X_1 \mid X_1 > x_0)]^2, \tag{3.16}$$

$$\mathrm{var}(X_2 \mid X_1 > x_0) = E(X_2(X_2-1) \mid X_1 > x_0)+E(X_2 \mid X_1 > x_0)-[E(X_2 \mid X_1 > x_0)]^2, \tag{3.17}$$

$$\mathrm{cov}(X_1, X_2 \mid X_1 > x_0) = E(X_1X_2 \mid X_1 > x_0) - E(X_1 \mid X_1 > x_0)E(X_2 \mid X_1 > x_0). \tag{3.18}$$

The expression with the left cut-off ($X_1 > x_0$) will be replaced by ($X_1 \leq x_0$) to obtain equations 3.14 and 3.15. The required results, along with the conditional expectations discussed and derived earlier. For the right cut-off,

$$E(X_1^2) = E(X_1(X_1-1)\,|\,X_1 > x_0) + E(X_1\,|\,X_1 > x_0)$$

$$E(X_1(X_1-1) \mid X_1 > x_0) = \sum_{x_1=x_0+1}^{\infty} \sum_{x_2=0}^{\infty} X_1(X_1-1)P(X_1 = x_1, X_2 = x_2 \mid X_1 > x_0)$$

Using the definition of CDF, the above expectation can be simplified to

$$E(X_1^2 \mid X_1 > x_0) = \frac{1}{1-F(x_0|\alpha_1, \beta_1)} \left[\frac{\alpha_1^3 + 3\alpha_1^2 + 2\alpha_1 + 2\alpha_1^2\beta_1 + 8\alpha_1\beta_1 + 6\beta_1}{\alpha_1^2(\alpha_1+\beta_1)(1+\alpha_1)} - P\right]$$

where

$$P = \sum_{x_1=0}^{x_0} \frac{x_1^2 \cdot \alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}}$$

$$E(X_2(X_2-1) \mid X_1 > x_0) = \sum_{x_1=x_0+1}^{\infty} \sum_{x_2=0}^{\infty} X_2(X_2-1)P(X_1 = x_1, X_2 = x_2 \mid X_1 > x_0)$$

$$
\begin{aligned}
E(X_2(X_2-1) \mid X_1 > x_0) &= \sum_{x_1=x_0+1}^{\infty}\sum_{x_2=0}^{\infty} X_2(X_2-1)P(X_1=x_1, X_2=x_2 \mid X_1 > x_0) \\
&= \frac{1}{1-F(x_0|\alpha_1,\beta_1)}\Bigg[(1-F(x_0|\alpha_1,\beta_1))\frac{2\alpha_2^2(\alpha_2+3\beta_2)(1+\alpha_2)}{\alpha_2^4(1+\alpha_2)(\alpha_2+\beta_2)}+ \\
&\quad \theta(P_1 - m1(1-F(x_0|\alpha_1,\beta_1)))(P_2 - m2\frac{2\alpha_2^2(\alpha_2+3\beta_2)(1+\alpha_2)}{\alpha_2^4(1+\alpha_2)(\alpha_2+\beta_2)})\Bigg]
\end{aligned}
$$

where

$$
P_1 = \sum_{x_1=x_0+1}^{\infty}\frac{e^{-x_1}\cdot\alpha_1^2(1+\alpha_1+\beta_1+\beta_1x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}},
$$

$$
P_2 = \sum_{x_2=0}^{\infty}\frac{e^{-x_1}\cdot x_2(x_2-1)\cdot\alpha_2^2(1+\alpha_2+\beta_2+\beta_2x_2)}{(\alpha_2+\beta_2)(1+\alpha_2)^{x_2+2}}
$$

And the cross product moment simplifies to

$$
E(X_1X_2|X_1 > x_0) = \frac{1}{1-F(x_0|\alpha_1,\beta_1)}\left[D_1\mu_2 + \theta(D_2 - m_1D_1)(m_{22} - m_2\mu_2)\right]
$$

where

$$
D_1 = \sum_{x_1=x_0+1}^{\infty}\frac{x_1\cdot\alpha_1^2(1+\alpha_1+\beta_1+\beta_1x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}},
$$

$$
D_2 = \sum_{x_1=x_0+1}^{\infty}\frac{e^{-x_1}x_1\cdot\alpha_1^2(1+\alpha_1+\beta_1+\beta_1x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}}
$$

Now for the left cut-off point,

$$
\begin{aligned}
E(X_1(X_1-1) \mid X_1 \le x_0) &= \sum_{x_1=0}^{x_0}\sum_{x_2=0}^{\infty} X_1(X_1-1)P(X_1=x_1, X_2=x_2 \mid X_1 \le x_0) \\
&= \frac{1}{F(x_0|\alpha_1,\beta_1)}\left[\sum_{x_1=0}^{x_0}\frac{x_1(x_1-1)\cdot\alpha_1^2(1+\alpha_1+\beta_1+\beta_1x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}}\right]
\end{aligned}
$$

$$
\begin{aligned}
E(X_2(X_2-1) \mid X_1 \le x_0) &= \sum_{x_1=0}^{x_0}\sum_{x_2=0}^{\infty} X_2(X_2-1)P(X_1=x_1, X_2=x_2 \mid X_1 \le x_0) \\
&= \frac{1}{F(x_0|\alpha_1,\beta_1)}\Bigg[F(x_0|\alpha_1,\beta_1)\frac{2\alpha_2^2(\alpha_2+3\beta_2)(1+\alpha_2)}{\alpha_2^4(1+\alpha_2)(\alpha_2+\beta_2)}+ \\
&\quad \theta(v_1 - m1\cdot F(x_0|\alpha_1,\beta_1))(P_2 - m2\frac{2\alpha_2^2(\alpha_2+3\beta_2)(1+\alpha_2)}{\alpha_2^4(1+\alpha_2)(\alpha_2+\beta_2)})\Bigg]
\end{aligned}
$$

where

$$
v_1 = \sum_{x_1=0}^{x_0}\frac{e^{-x_1}\cdot\alpha_1^2(1+\alpha_1+\beta_1+\beta_1x_1)}{(\alpha_1+\beta_1)(1+\alpha_1)^{x_1+2}},
$$

P_2, m_1 and m_2 have been defined earlier.

and the cross product moment reduces to

$$E(X_1X_2 \mid X_1 \leq x_0) = \frac{1}{F(x_0|\alpha_1, \beta_1)} \left[H_1\mu_2 + \theta(H_2 - m_1H_1)(m_{22} - m_2\mu_2)\right]$$

where

$$H_1 = \sum_{x_1=0}^{x_0} \frac{x_1 \cdot \alpha_1^2(1 + \alpha_1 + \beta_1 + \beta_1 x_1)}{(\alpha_1 + \beta_1)(1 + \alpha_1)^{x_1+2}},$$

$$H_2 = \sum_{x_1=0}^{x_0} \frac{e^{-x_1} x_1 \cdot \alpha_1^2(1 + \alpha_1 + \beta_1 + \beta_1 x_1)}{(\alpha_1 + \beta_1)(1 + \alpha_1)^{x_1+2}},$$

The expressions for the variance, pertaining to a right cut-off ($> x_0$), can be derived by substituting the values from equations 3.16, 3.17, and 3.18 into equation 3.14.

3.4 The effect of cut-off point, x_0, on RTM

It is well known that the selection criterion of subjects for inclusion in an intervention study plays an important role in the magnitude of RTM. Using $R_r(x_0; \alpha, \beta)$ and $R_l(x_0; \alpha, \beta)$, a graph for different cut-off points is given in figure 3.1. The variables X_1 and X_2 are considered for demonstrative purposes if they are negatively correlated and have specific values of $\alpha_1 = 1.5$, $\alpha_2 = 3$, $\beta_1 = 1$, $\beta_2 = 1$, and $\theta = -2$. and if they are positively correlated and have specific values of $\alpha_1 = 1.5$, $\alpha_2 = 3$, $\beta_1 = 1$, $\beta_2 = 1$, and $\theta = 2$. The graph visualization revealed that for both positive and negative correlation situations, the RTM effect increases with increasing cut-off point for the right cut-off, while the RTM effect is higher for farther left cut-off points, decreases in the tail of the distribution as the cut-off point increases. When there is a negative correlation between variables X_1 and X_2, the RTM effect is slightly larger in magnitude than when there is a positive correlation. The result is depicted in Figure 3.1. The farther is the cut-off point in the tail of a distribution the more severe is the RTM effect.

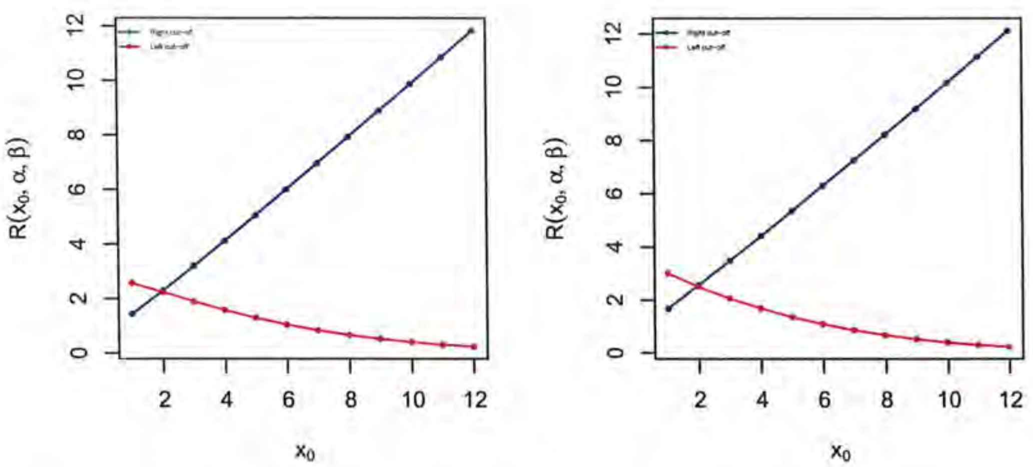

Figure 3.1: Graph of the derived formula of RTM for greater than or less than cut-off points when the underlying distribution is Aryuyuen and Bodhisuwan (2023)'s bivariate generalized Poisson Lindley distribution. Left panel: Positive correlation; Right panel: Negative correlation

3.5 RTM as a function of θ or correlation

The RTM as a function of θ is given in Figure 3.2. A fixed cut-off point $x_0 = 3$ and specific values of parameters ($\alpha_1 = 1.5$, $\alpha_2 = 3$, $\beta_1 = \beta_2 = 1$) are considered for demonstration purposes. As discussed by (Aryuyuen and Bodhisuwan, 2023), the θ parameter is responsible for the correlation between the variables, such that a positive value of θ results in positive correlation between the variables, and a negative value of θ gives rise to a negative correlation. For the specified parameter values, θ varies between -5 and 5. The RTM effect is computed using equations 3.8 and 3.13 and is visualized in Figure 3.2. For negative values of θ (correlations) -5, the RTM effect is maximum and then starts decreasing as the value of θ (correlation) approaches to zero. The RTM effect is small for both right and left cut-off point when the value of θ is 5. From the graph depicted in 3.2, it is evident that RTM is a linear function of θ

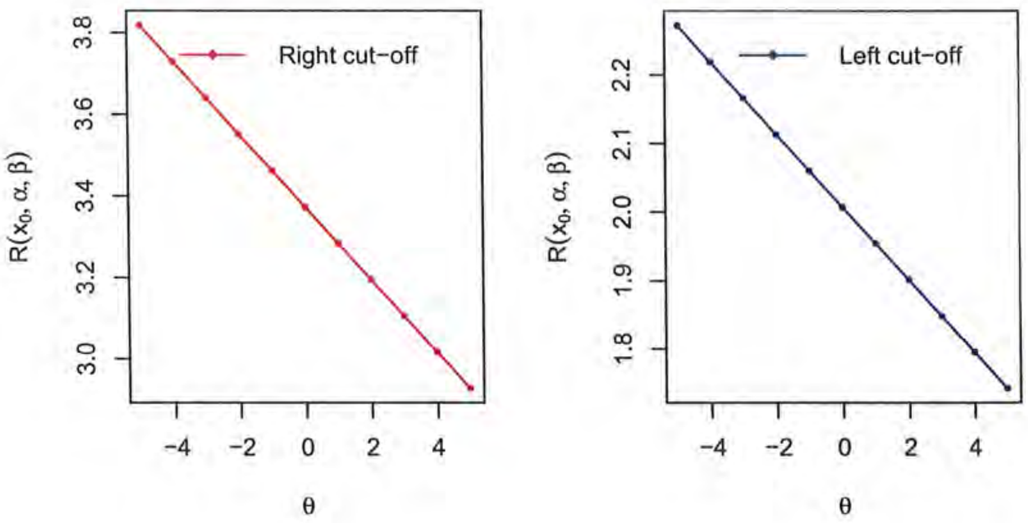

Figure 3.2: Graph illustrating the RTM effect, varying the parameter θ while keeping the cut-off point x_0 constant. Set $\alpha_1 = 1.5$, $\alpha_2 = 3$, and $\beta_1 = \beta_2 = 1$.

3.6 Maximum Likelihood Estimation (MLE)

Let $(x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})$ be pairs of pre and post observations of size n sampled from a BGPL(Θ), where $\Theta = (\alpha_1, \beta_1, \alpha_2, \beta_2, \theta)$, as proposed by Aryuyuen and Bodhisuwan (2023). Let $P_t(x_1, x_2)$ represent the truncated bivariate probability distribution. The likelihood function can then be expressed as follows

$$L(\Theta) = \prod_{i=1}^{n} f(x_{1i}, x_{2i}; \Theta)$$

and the log likelihood function is

$$
\begin{aligned}
\ell(\Theta) &= \log \prod_{i=1}^{n} f(x_{1i}, x_{2i}; \Theta) \\
&= 2n \log \alpha_2 + 2n \log \alpha_1 - (2 + x_{1i}) \sum_{i=1}^{n} \log(1 + \alpha_1) - n \log(\beta_2 + \alpha_2) \\
&\quad + \sum_{i=1}^{n} \log(\alpha_1 + \beta_1 + \beta_1 x_{1i} + 1) - n \log(\beta_1 + \alpha_1) \\
&\quad - (2 + x_{2i}) \sum_{i=1}^{n} \log(1 + \alpha_2) + \sum_{i=1}^{n} \log(\alpha_2 + \beta_2 + \beta_2 x_{2i} + 1) \\
&\quad + \sum_{i=1}^{n} \log \left[1 + \theta \left(e^{-x_{2i}} - \frac{\alpha_2^2(\alpha_2 + \beta_2 + 1 - e^{-1})}{(\alpha_2 + \beta_2)(1 + \alpha_2 - e^{-1})^2} \right) \right. \\
&\quad \left. \times \left(e^{-x_{1i}} - \frac{\alpha_1^2(\alpha_1 + \beta_1 + 1 - e^{-1})}{(\beta_1 + \alpha_1)(1 + \alpha_1 - e^{-1})^2} \right) \right].
\end{aligned}
$$

To estimate the unknown parameters Θ we take the partial derivatives with respect to the parameters α_i, β_i and θ for $i = 1, 2$ and equating the obtained results to zero yields five estimating equations. However, the above equations are not provided in closed forms and cannot be explicitly solved for the involved parameters. So a direct maximization of the $\ell(\Theta)$ would make it easier to obtain maximum likelihood estimates of the parameters. To achieve the objective, the log likelihood function has been maximized using the optim built in function R.

To find the Maximum Likelihood Estimate (MLE) of parameters for the truncated bivariate generalized Poisson Lindley distribution, a set of random numbers of size 10,000 is generated from the bivariate generalized Poisson Lindley distribution with $\alpha_1 = 1.5$, $\alpha_2 = 1.5$, $\beta_1 = 3, \beta_2 = 3$, and $\theta = 2$. The observation above cut-off point 3 are considered and the true parameters are estimated from the truncated density.

The log-likelihood plot depicted in figure 3.3 displays the log-likelihood values corresponding to parameters α_1, α_2, β_1, β_2, and θ. Notably, the plot suggests estimated values of 1.5 for α_1, 1.5 for α_2, and 2 for θ. However, a notable discrepancy is observed in the log-likelihood plot for the estimated parameter β_1, where the value is 10, deviating significantly from the true value of 3. This discrepancy indicates a potential inadequacy in accurately estimating the β_1 parameter. Similarly, the same behavior of the β_2 parameter. Therefore, we fixed both parameters β_1 and β_2, which are equal to 1.

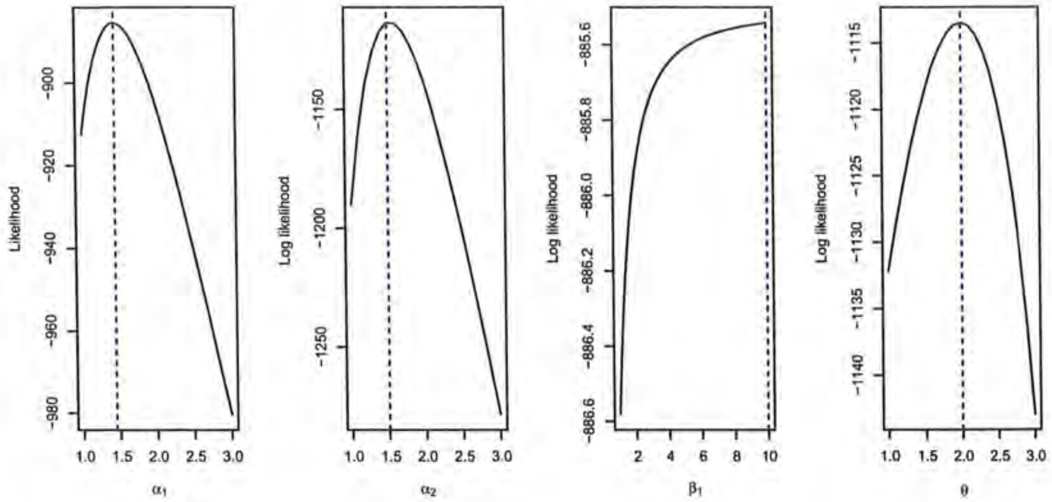

Figure 3.3: The log-likelihood plot depicts the maximum likelihood estimates of the parameters for a bivariate Generalized Poisson Lindley distribution, the estimates are represented by blue points.

The MLE has large sample property that when the sample size increase then the estimated values approach true values. In Fig. 3.4 the sample size between 125 and 250 also gives reasonable estimates but when the sample size increases up to 500 or above then the estimates are getting close to the true value. In Fig. 3.4 the estimated value is plotted against sample size n which shows that the values approach dashed lines that represent true values.

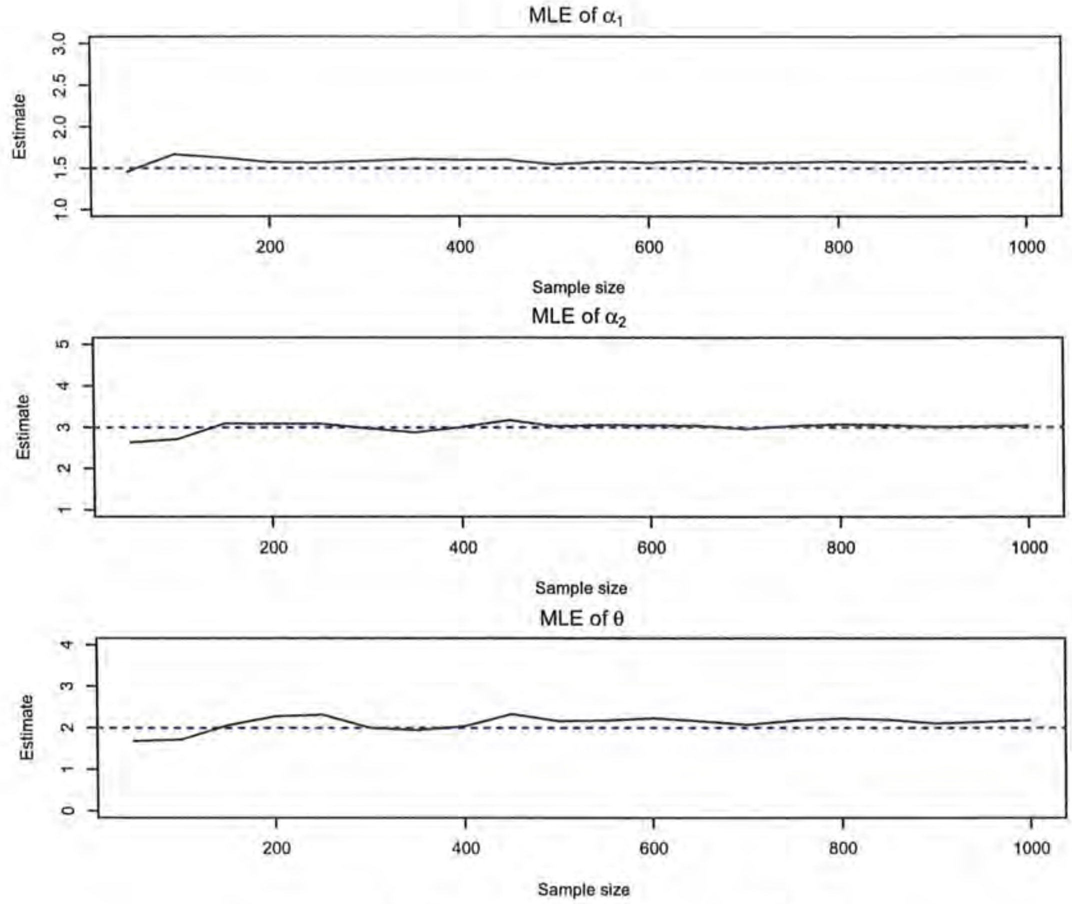

Figure 3.4: Large sample property of MLE

3.7 Data Generation and Simulation Study

One of the key goals before starting a simulation study is to generate data. Data generation from the bivariate generalized Poisson lindley distribution of Aryuyuen and Bodhisuwan (2023) is an important and tricky task. To generate the data, first a random sample on X_1 was generated from a univariate generalized Poisson lindley distribution. As in the current study the variables are not independent from each other, so the conditional pmf from the joint pmf of Aryuyuen and Bodhisuwan (2023) bivariate Poisson pmf was used to generate data on the second random variable X_2. To achieve the purpose of generating the bivariate data, conditional sampling method is adopted by first generating sample for the univariate generalized Poisson and then from the conditional pmf given as follow.

$$P(X_1 = x_1) = \frac{\alpha_1^2(1+\alpha_1+\beta_1+\beta_1 x_1)}{(\alpha_1+\beta_1)(\alpha_1+1)^{x_1+2}}$$

$$P(X_2|X_1 = x_1) = \frac{\alpha_2^2(1+\alpha_2+\beta_2+\beta_2 x_2)}{(\alpha_2+\beta_2)(\alpha_2+1)^{x_2+2}} \times \left[1+\theta(e^{-x_1}-m_1)(e^{-x_2}-m_2)\right]$$

The built in R function can facilitate to generate pseudo random numbers for simulation from the conditional distribution which is given below.

```
sample(x, size, replace = TRUE, prob)
```

For some specific values of parameters of the truncated bivariate generalized poisson lindley distribution a simulation study was conducted in order to compare the estimated RTM with true RTM effect to check its performance. Following are the steps taken to generate sets of observation, considering as a pre and post observation of intervention study.

1. A random sample was generated from univariate generalized Poisson Lindley with parameter α_1,β_1

2. In the conditional pmf, each datum x_i was substituted along with other parameters α_2, β_2, and θ. The built-in R function `sample(x, size, replace = TRUE, prob)` was used to generate pseudo-random numbers for simulation from the mentioned conditional distribution.

3. These sample realizations were denoted by x_{ij} for $i = 1, 2$, and $j = 1, 2, \ldots, n_1$.

4. x_{1j} and x_{2j} were considered as pre and post observations of an intervention study.

5. The first n number of observations of x_{1j} beyond/below a truncation point x_0 and the corresponding x_{2j} observations were then considered a random sample from Aryuyuen and Bodhisuwan (2023) truncated bivariate generalized Poisson lindley distribution.

6. This sampling procedure was repeated 1000 times and for each sample, the RTM effect was estimated using the maximum likelihood estimation.

3.7.1 Empirical distribution of $\hat{R}_k(x_0, x_1, x_2)$

The sampling distribution of $\hat{R}_k(x_0, x_1, x_2)$ using normal quantile-quantile is shown in Figure 3.5. The visualization is appealing; the sampling distributions for sample sizes 100 and 200 are both approximately normal. Sampling distributions of RTM at different cut-off points and parameters were also found to be normal but are not given for brevity.

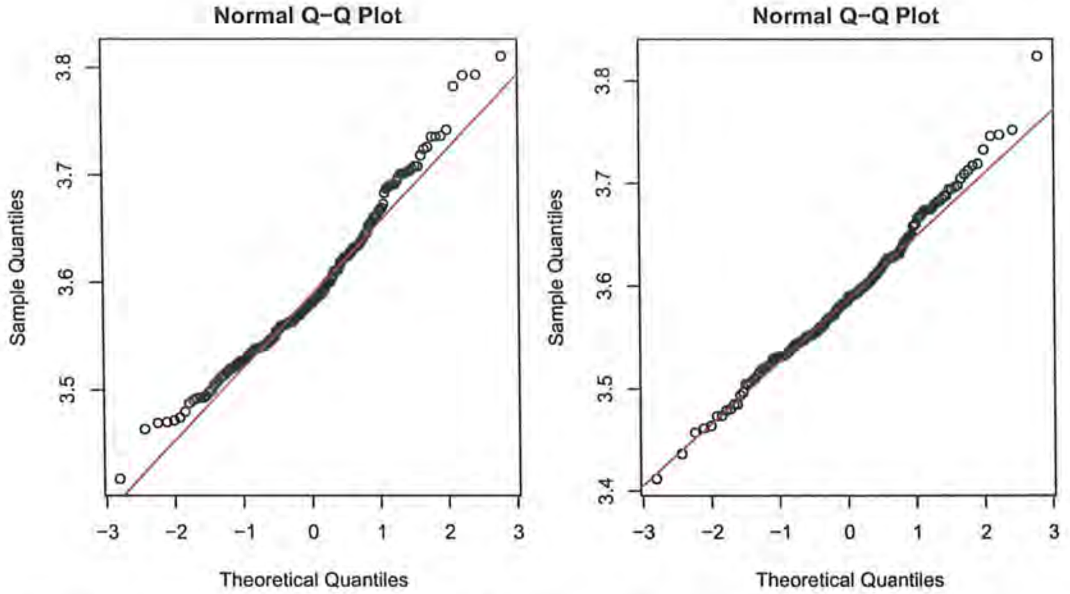

Figure 3.5: Normal Q-Q plot of the sampling distribution of RTM effect for $\alpha_1 = 1.5$, $\alpha_2 = 3$, $\beta_1 = \beta_2 = 1$, $\theta = 2$ and $x_0 = 3$. Left panel: $n = 100$; Right panel: $n = 200$

3.7.2 Empirical unbiasdness and consistency of $\hat{R}_k(x_0, x_1, x_2)$

Figure 3.7 shows the comparison of estimated and actual RTM (the red dotted line) for various sample sizes. The mean of estimated RTM (the blue line segments) are very close to the actual RTM for different sample sizes, as can be seen in figure 3.7, demonstrating the unbiasedness of RTM estimator. Increasing the sample size, from 50 to 250, minimizes the variation from the mean or center, implying that the RTM estimator is also consistent.

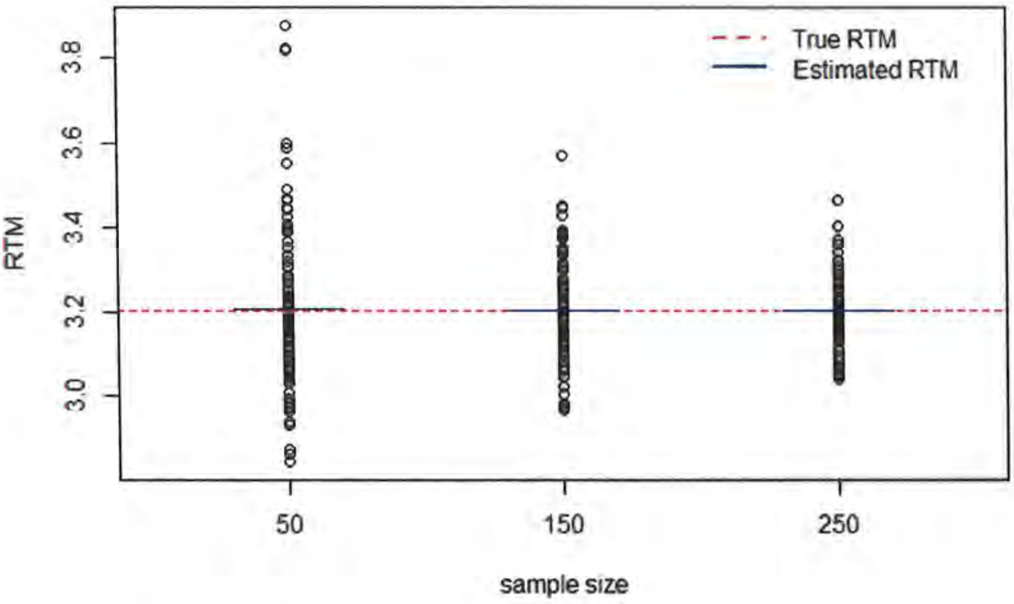

Figure 3.6: Estimates of RTM and its sampling distribution for different sample choices and parameters $x_0 = 3$, $\alpha_1 = 1.5$, $\alpha_2 = 3$, $\beta_1 = \beta_2 = 1$, and $\theta = 2$.

3.8 Comparison via simulations study

Simulation study was performed with various sample sizes while keeping the parameters fixed and utilizing a specific right cut-off point. The outcomes are presented in Table 3.1, indicating that as the sample size increases, the estimated RTM and treatment values approach the true RTM and treatment values. The simulation results strongly suggest that the proposed expressions provide estimations of the treatment effect that closely align with the true values. Similar conclusions were drawn for the RTM and treatment effect associated with the left cut-off point using various parameters, although these results are not presented here for brevity.

Table 3.1: Comparison of estimate and true RTM for Different Sample Sizes

Parameters: $\alpha_1 = 2$, $\alpha_2 = 2$, $\beta_1 = \beta_2 = 1$, $\theta = 2$, $x_0 = 3$

Sample Size	True		Estimate	
	Treatment	RTM	Treatment	RTM
100	0	3.3373	-0.1746	3.3063
200	0	3.3373	-0.1140	3.3285
300	0	3.3373	-0.0306	3.3290
400	0	3.3373	0.0716	3.3150
500	0	3.3373	0.0204	3.3304
700	0	3.3373	0.0042	3.3321

Parameters: $\alpha_1 = 2$, $\alpha_2 = 3$, $\beta_1 = \beta_2 = 1$, $\theta = 2$, $x_0 = 3$

Sample Size	True		Estimate	
	Treatment	RTM	Treatment	RTM
100	0.5166	3.5874	0.8199	3.7538
200	0.5166	3.5874	0.7230	3.6489
300	0.5166	3.5874	0.6235	3.6085
400	0.5166	3.5874	0.6056	3.6068
500	0.5166	3.5874	0.5762	3.5981
700	0.5166	3.5874	0.5380	3.5953

Parameters: $\alpha_1 = 3$, $\alpha_2 = 2$, $\beta_1 = \beta_2 = 1$, $\theta = 2$, $x_0 = 3$

Sample Size	True		Estimate	
	Treatment	RTM	Treatment	RTM
100	-0.5166	3.2857	-0.6563	3.2334
200	-0.5166	3.2857	-0.6243	3.2563
300	-0.5166	3.2857	-0.5952	3.2792
400	-0.5166	3.2857	-0.5696	3.2761
500	-0.5166	3.2857	-0.5564	3.2710
700	-0.5166	3.2857	-0.5413	3.2913

3.9 Data Example

The data set includes accident data from 122 experienced shunters, with random variables X and Y representing the number of accidents from 1937-1942 and 1943-1947, extracted from Aryuyuen and Bodhisuwan (2023) as shown in table 3.2 is used to quantify the RTM effect. The statistics calculated from the data are $\bar{x} = 1.2705$, $\bar{y} = 0.9754$, $\sigma_x^2 = 1.6535$, $\sigma_y^2 = 1.2969$, $\text{Cov}(x, y) = 0.37860$, and $\text{Cor}(x, y) = 0.2585$. Aryuyuen and Bodhisuwan (2023) used the method of moments to estimate the parameters. In this work, the parameters are estimated using the maximum likelihood method of estimation considering all data points, i.e., no truncation. The estimated parameters were found to be $\alpha_1 = 1.5540$, $\alpha_2 = 1.9843$, $\beta_1 = 44.091$, $\beta_2 = 30.548$, and $\theta = 1.5579$. On the basis of these estimated parameters the total, RTM, and treatment effect at different cut off points is shown in Figure 3.7. The RTM effect is increasing with the increase in the cut off point for right truncation as the total effect is increasing.

Table 3.2: Bivariate accident count data of 122 shunters

X	Y								Total
	0	1	2	3	4	5	6	7	
0	21	13	4	2	0	0	0	0	40
1	18	14	5	1	0	0	0	1	39
2	8	10	4	3	1	0	0	0	26
3	2	1	2	2	1	0	0	0	8
4	1	4	1	0	0	0	0	0	6
5	0	1	0	1	0	0	0	0	2
6	0	0	1	0	0	0	0	0	1
Total	50	43	17	9	2	0	0	1	122

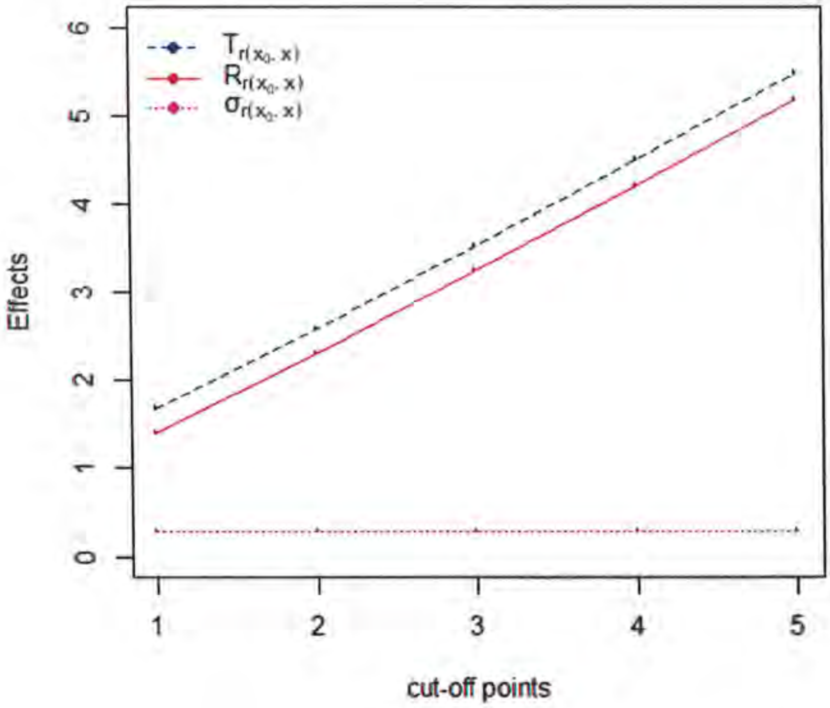

Figure 3.7: RTM effects for points greater than x_0.

As the cut-off point moves further into the tail, the RTM effect increases, leading to an increase in the average total effect while maintaining a constant average treatment effect. As a result, an observed average increase or decrease, which is the additive effect of the RTM and treatment, may be misinterpreted as a true change.

3.10 Discussion

RTM is a natural phenomenon which occur when repeated measurements are observed at different times. In pre post studies intervention are applied to the subjects of the studies based on some threshold point i.e., below or above and RTM can potentially effect the conclusions made about the interventions. Therefore, its quantification is an important statistical research problem. In literature, a number of methods are available to quantify RTM in bivariate Poisson distribution have negative correlated count variables. Poisson data may exhibit positive correlations among the study variables.

Aryuyuen and Bodhisuwan (2023) developed a bivariate generalized Poisson lindley distribution having flexible correlation structure and is used to formulate expression for RTM for the bivariate Poisson count data exhibiting flexible correlation. Our derivations assuming bivariate generalized Poisson lindley discussed by Aryuyuen and Bodhisuwan (2023) showed similar behavior with that of bivariate normal in terms of the covariance. The RTM is linearly related with the correlation and is maximum for the negative correlation.

Moreover, The maximum likelihood estimators were obtained by maximizing the log likelihood function using the optim function in R. The simulation study revealed that the maximum likelihood (ML) estimators of RTM are unbiased and consistent. The comparison of the proposed expression of RTM with actual values via simulations study revealed that as the sample size increases the proposed RTM approach to actual values.

Chapter 4

Conclusion

In practical scenarios, various situations arise where interventions or treatments are implemented on participants to assess improvements in study variables, aiming for betterment. When these interventions are administered to individuals with measurements at the extremes, either below or above a designated cut-off point, the conclusions regarding the effectiveness of the intervention or treatment effect may potentially be influenced by the regression to the mean effect (RTM).Thus, an ineffective intervention could be considered effective due the RTM effect if overlooked. Such erroneous conclusions have been reported in vast research areas such as business, economics, public health, sports, and managements as discussed in the introduction and are not limited to the clinical studies.

An accurate estimation of the RTM effect is needed in the intervention studies for accurate estimation of the intervention effect. So far, researchers have developed many methods to estimate the RTM effect, but existing methods are based on restricted assumptions, such as normality of the bivaraite data which may not be the case in real life. On the other hand, for non-normal populations the model or methods have limitation of in-applicability to empirical distribution, computational inconvenience and multi-modality problems.

In pre-post studies RTM effect could occur when a treatment is applied to individuals or subjects selected on the basis of a cut-off point for inclusion in a study. Intervention effect can be then estimated accurately by accounting for the RTM effect which could be a part of the total effect.

In this study, we delve into the estimation of both regression to the mean (RTM) and intervention effects. This is examined through the utilization of bivariate generalized Poisson lindley distributions, which permit consideration of positive correlation.

Moreover, as the right and left cut-off points increases, the RTM was observed to be monotonically increasing and decreasing, respectively. In cases of correlated and non-dispersed data the RTM intersects at some point for the right and left cut-off points, while for the bivariate generalized Poisson lindley distributions, which allow for dispersed data and positive correlation, it showed opposite behavior with no intersection.

The maximum likelihood estimators of the RTM and intervention effects were derived in the current work. The properties such as unbiasedness, consistency and asymptotic normality were verified through simulations for the bivariate generalized Poisson lindely distributions which allow for positive correlation.

The intervention and RTM effects were estimated using maximum likelihood estimation utilising data on the bivariate accident count data of 122 shunters . It was demonstrated that the observed change was driven by the RTM effect and should be accounted for to accurately estimate the intervention effect.

References

Aryuyuen, S. and Bodhisuwan, W. (2023). On new bivariate poisson-lindley distribution with application of correlated bivariate count data analysis. *Thailand Statistician*, 21(2):228–243.

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.

Barton, D. and Dennis, K. (1952). The conditions under which gram-charlier and edgeworth curves are positive definite and unimodal. *Biometrika*, 39(3/4):425–427.

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.

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

Campbell, J. (1934). The poisson correlation function. *Proceedings of the Edinburgh Mathematical Society*, 4(1):18–26.

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

Das, P. and Mulder, P. (1983). Regression to the mode. *Statistica Neerlandica*, 37(1):15–20.

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.

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

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

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.

Kario, K., Schwartz, J. E., and Pickering, T. G. (2000). Changes of nocturnal blood pressure dipping status in hypertensives by nighttime dosing of α-adrenergic blocker, doxazosin: results from the halt study. *Hypertension*, 35(3):787–794.

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.

Lakshminarayana, J., Pandit, S. N., and Srinivasa Rao, K. (1999). On a bivariate poisson distribution. *Communications in Statistics-Theory and Methods*, 28(2):267–276.

Lee, M. and Smith, G. (2002). Regression to the mean and football wagers. *Journal of Behavioral Decision Making*, 15(4):329–342.

McCambridge, J., Kypri, K., and McElduff, P. (2014). Regression to the mean and alcohol consumption: a cohort study exploring implications for the interpretation of change in control groups in brief intervention trials. *Drug and alcohol dependence*, 135:156–159.

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

Prior, J. O., van Melle, G., Crisinel, A., Burnand, B., Cornuz, J., and Darioli, R. (2005). Evaluation of a multicomponent worksite health promotion program for cardiovascular risk factors—correcting for the regression towards the mean effect. *Preventive medicine*, 40(3):259–267.

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

Retting, R. A., Ferguson, S. A., and Hakkert, A. S. (2003). Effects of red light cameras on violations and crashes: a review of the international literature. *Traffic injury prevention*, 4(1):17–23.

Roustit, M., Jullien, A., Jambon-Barbara, C., Goudon, H., Blaise, S., Cracowski, J.-L., and Khouri, C. (2022). Placebo response in raynaud's phenomenon clinical trials: The prominent role of regression towards the mean: Placebo response in raynaud's phenomenon. In *Seminars in Arthritis and Rheumatism*, volume 57, page 152087. Elsevier.

Schectman, G. and Hoffmann, R. G. (1988). A history of hypercholesterolemia influences cholesterol measurements. *Archives of internal medicine*, 148(5):1169–1171.

Schmidt, M. I., Bracco, P., Canhada, S., Guimarães, J. M., Barreto, S. M., Chor, D., Griep, R., Yudkin, J. S., and Duncan, B. B. (2021). Regression to the mean contributes to the apparent improvement in glycemia 3.8 years after screening: the elsa-brasil study. *Diabetes Care*, 44(1):81–88.

Shahane, A., George, V., and Johnson, W. D. (1995). Effect of bivariate regression toward the mean in uncontrolled clinical trials. *Communications in Statistics-Theory and Methods*, 24(8):2165–2181.

Tallis, G. M. (1961). The moment generating function of the truncated multinormal distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 23(1):223–229.

Teicher, H. (1954). On the multivariate poisson distribution. *Scandinavian Actuarial Journal*, 1954(1):1–9.

Wilcox, M. A., Chang, A. M., and Johnson, I. R. (1996). The effects of parity on birthweight using successive pregnancies. *Acta obstetricia et gynecologica Scandinavica*, 75(5):459–463.