

Quantification and Estimation of Regression to the mean for  
Bivariate Lognormal Distribution



By

Saddam Hussain

Department of Statistics  
Faculty of Natural Sciences  
Quaid-i-Azam University, Islamabad

2021

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

---

Quantification and Estimation of Regression to the mean for  
Bivariate Lognormal Distribution



By

Saddam Hussain

*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

2021

# CERTIFICATE

**Quantification and Estimation of Regression to the  
mean for Bivariate Lognormal Distribution**


By

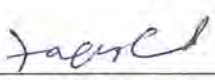
**SADDAM HUSSAIN**

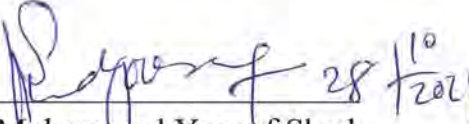
**(Reg.No. 02221913003)**

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.   
Dr. Manzoor Khan  
(Supervisor)

2.   
Prof. Dr. Faqir Muhammad  
(External Examiner)

3.   
Dr. Muhammad Yousaf Shad  
(Chairman)

**DEPARTMENT OF STATISTICS  
QUAID-I-AZAM UNIVERSITY  
ISLAMABAD, PAKISTAN  
2021**

---

## Declaration

I “Saddam Hussain” hereby solemnly declare that this thesis entitled “Quantification and Estimation of Regression to the mean for Bivariate Lognormal Distribution”, submitted by me for the partial fulfillment of Master of Philosophy in Statistics, is the original work and has not been submitted concomitantly or latterly to this or any other university for any other Degree.

Dated: \_\_\_\_\_

Signature: \_\_\_\_\_

---

# *DEDICATION*

I dedicate this thesis to my sweetened and affectionate Parents, whose love, adoration, help and prays create me capable to get such accomplishment and credit.

---

# *Acknowledgments*

All the praises to ALLAH ALMIGHTY, the most merciful and the entire source of knowledge and wisdom and a due respect for the Muhammad (Peace be upon him), who is forever a torch of guidance for mankind.

I would like to thank my supervisor and Chairman, Dr. Manzoor Khan, for his guidance and support throughout the completion of my research. I cannot thank him enough for his thoughtful guidance, generous support, and friendly discussions which helped me to improve my academic knowledge.

My deepest gratitude goes to my parents and my family for their invaluable support and constant encouragement throughout my academic endeavors. I also thank my friends and colleagues for their help and words of encouragement.

## Abstract

As a natural phenomenon, regression to the mean (RTM) happens when extreme observations are picked at the initial measurement and get closer to the mean throughout the subsequent measurements. Regression to the mean is a potential problem in data analysis that could lead to incorrect conclusions. Identifying and accounting for the RTM effect is an essential objective in any statistical study. RTM expressions for the Normal, Poisson and Binomial distributions are accessible in the literature. RTM expression is not available when the pre and post-variables are distributed according to the bivariate lognormal distribution. The RTM effect becomes more severe when the correlation between the two variables becomes weaker. Based on the correlation function, our derivations showed that a bivariate lognormal distribution behaved like a bivariate normal distribution. In pre-post experiments, the RTM impact decreases linearly as the correlation between variables increases. In a lognormal distribution, the RTM for the left and right cut-off points decrease differently with correlation. The proposed formulation for the RTM effect under bivariate lognormal distribution is substantially more satisfying than the Edgeworth series and Saddlepoint approximation. We conducted a simulation analysis to compare our suggested RTM expression to previously published approaches for non-normal populations. The RTM effect was assessed for 56 cyclosporin test pairs at various cut-off values. The study included blood samples from organ transplant patients. We get parameter estimates using maximum likelihood. It is unreasonable to assess cyclosporin's real efficacy without considering the RTM effect. The RTM effect becomes increasingly noticeable as the cut-off point approaches the tail of the distribution.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Background . . . . .	1
1.2	Historical Background . . . . .	4
1.3	Objectives . . . . .	7
<b>2</b>	<b>Literature</b>	<b>8</b>
2.1	RTM under the bivariate normal distribution . . . . .	8
2.1.1	James (1973) . . . . .	8
2.1.2	Davis (1976) . . . . .	10
2.1.3	Gardner and Heady (1973) . . . . .	10
2.1.4	Shahane et al. (1995) . . . . .	12
2.2	RTM for non-normal population . . . . .	13
2.2.1	Das and Mulder (1983) . . . . .	13
2.2.2	Beath and Dobson (1991) . . . . .	14
2.2.3	John and Jawad (2010) . . . . .	16
2.2.4	Khan and Olivier (2018) . . . . .	17
2.3	Designing studies to mitigate the RTM problem . . . . .	18
2.3.1	Controlled trials with randomization . . . . .	18
2.3.2	Two measurements approach of Ederer (1972) . . . . .	18
<b>3</b>	<b>Derivation of the Total/RTM effects</b>	<b>19</b>
3.1	Case 1: Right cut-off point . . . . .	21
3.2	Case 2: Left cut-off point . . . . .	22
3.2.1	Expression for RTM . . . . .	23
3.3	Expressions for Variances . . . . .	24
3.3.1	Variance of the total effect . . . . .	24

3.3.2	Variance of RTM . . . . .	27
3.4	RTM as a function of cut-off point $z_0$ . . . . .	27
3.5	RTM is a function of correlation $\rho_0$ . . . . .	28
3.6	Comparison of RTM under Lognormal and Normal distributions . . . . .	30
<b>4</b>	<b>Estimation of Total and RTM Effects</b>	<b>31</b>
4.1	Simulation Analysis of RTM Effect . . . . .	35
4.2	Comparison of methods . . . . .	35
4.2.1	Comparison the RTM effect . . . . .	35
4.2.2	Comparison of the estimating methods for treatment effect . . . . .	39
4.3	Data Example: the Cyclosporin Study . . . . .	42
<b>5</b>	<b>Discussion</b>	<b>44</b>
	<b>Reference</b>	<b>45</b>

# List of Figures

1.1	Height of Participants with an extreme height in the right at 190 cm of a parent . . . . .	2
1.2	Height of Participants . . . . .	2
1.3	Height of Participants with height of an individual whose parent's height was initially extreme . . . . .	3
3.1	Graph showing the RTM effect generated on the basis of the derived formula for points more than or less than a cut-off point, when the underlying distribution is bivariate lognormal with parameter $\mu_x = 2, \mu_y = 2, \sigma_x = 1.5, \sigma_y = 1.5$ and $\rho_{xy} = 0.6$ . . . . .	28
3.2	The graph shows RTM effect for distinct points of $\rho$ and predetermined cut-off point $x_0 = 12$ , when the underlying distribution is bivariate lognormal with parameter $\mu_x = 2, \mu_y = 2, \sigma_x = 1.5$ and $\sigma_y = 1.5$ . . . . .	29
3.3	The graph shows the comparison of RTM under the lognormal and normal distributions for different cut-off points. The bivariate lognormal/normal parameters are $\mu_x = 6, \mu_y = 6, \sigma_x = 0.1, \sigma_y = 0.1$ and $\rho = 0.6$ . . . . .	30
4.1	RTM at correlation coefficient $\rho=0.4$ . . . . .	36
4.2	RTM at correlation coefficient $\rho=0.6$ . . . . .	36
4.3	RTM at correlation coefficient $\rho=0.4$ and $\mu_1=2.1$ . . . . .	36
4.4	RTM at correlation coefficient $\rho=0.6$ and $\mu_1=2.1$ . . . . .	36
4.5	RTM at correlation coefficient $\rho=0.4$ and $\sigma_1 = 1.6$ . . . . .	37
4.6	RTM at correlation coefficient $\rho=0.6$ and $\sigma_1 = 1.6$ . . . . .	37
4.7	RTM at correlation coefficient $\rho=0.4$ and $\sigma_2 = 1.6$ . . . . .	38
4.8	RTM at correlation coefficient $\rho=0.6$ and $\sigma_2 = 1.6$ . . . . .	38
4.9	RTM at correlation coefficient $\rho=0.4$ and $\mu_2=2.1$ . . . . .	38
4.10	RTM at correlation coefficient $\rho=0.6$ and $\mu_2=2.1$ . . . . .	38

## List of Figures

---

4.11 Treatment at correlation coefficient $\rho=0.6$ . . . . .	40
4.12 Treatment at correlation coefficient $\rho=0.4$ . . . . .	40
4.13 Treatment at correlation coefficient $\rho=0.6$ and $\mu_1=2.1$ . . . . .	40
4.14 Treatment at correlation coefficient $\rho=0.4$ and $\mu_1=2.1$ . . . . .	40
4.15 Treatment at correlation coefficient $\rho=0.6$ and $\sigma_1=1.6$ . . . . .	41
4.16 Treatment at correlation coefficient $\rho=0.4$ and $\sigma_1=1.6$ . . . . .	41
4.17 Treatment at correlation coefficient $\rho=0.6$ and $\sigma_2=1.6$ . . . . .	42
4.18 Treatment at correlation coefficient $\rho=0.4$ and $\sigma_2=1.6$ . . . . .	42
4.19 Graph shows that the total, RTM, and treatment effects generated on the basis of the derived formula for points more than a cut-off value for $\hat{\mu}_x = 4.88, \hat{\mu}_y = 4.96, \hat{\sigma}_x = 0.92, \hat{\sigma}_y = 0.81$ and $\hat{\rho}_{xy} = 0.96$ . . . . .	43

# Chapter 1

## Introduction

### 1.1 Background

Despite being remembered by Sir Francis Galton as early as 1886, regression toward the mean has gained much attention lately. Galton first observed regression to the mean (RTM) when he noticed that parents taller than the average population height had children shorter than their parents' height, but closer to the average population height. On the other hand, parents shorter than the average population height had taller children than them and more closer to the average population height. The phenomenon is that a subsequent measurements on a random variable that was extreme on its first observation appear to regress toward the distribution's center. In biological studies, regression toward the mean is commonly used to account for the statistical evaluation of a treatment/intervention effect ([Ibrahim, 2015](#)). Researchers are also interested in analyzing the effect of a treatment on a group of respondents who have an exceptionally high or low quantitative characteristic ([Shahane et al., 1995](#)).

RTM could take place due to the individual observation when observed with random error (variation). In [Figure 1.1](#), the hypothetical heights of individuals are depicted which is normally distributed with a mean of 170 cm and a standard deviation of 10 cm. The [Figure 1.2](#) shows an observed value of 190 cm which is extremely high, let us call it parent's height. Upon measuring the height of his child, it would be less than 190 cm and closer to the true population mean. In the [Figure 1.3](#), the measured value of child is given which is closed to the true population mean, 170 cm ([Barnett et al., 2005](#)).

There is as well the possibility of the occurrence of RTM at the group level. Assume the height of participants follows a normal distribution with a mean of 170 cm and a standard deviation of 10 cm. Based on the initial readings, we choose a group of individuals with an extreme height greater than 180 cm from the population. Due to the random variation, there would be more participants in the group whose height greater or below 180 cm. The group's mean height value will decline on follow-up measurements, the participants with highly extreme initial height are due to the random variation are closer to the population mean 170 cm (Barnett et al., 2005).

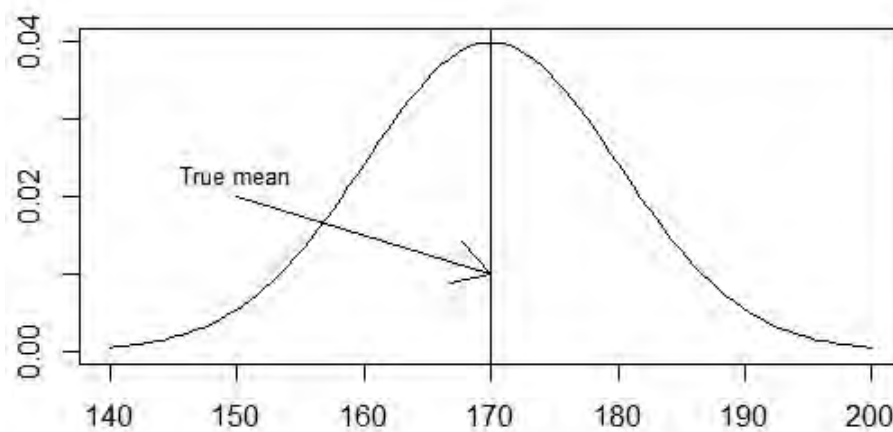


Figure 1.1: Height of Participants with an extreme height in the right at 190 cm of a parent

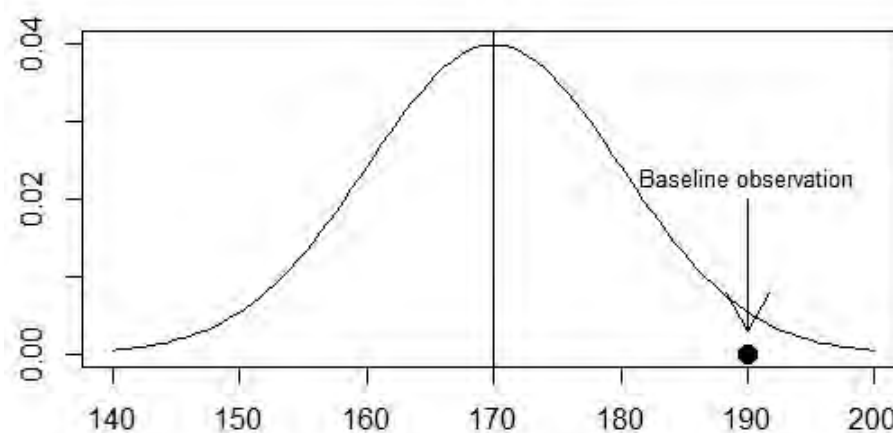


Figure 1.2: Height of Participants

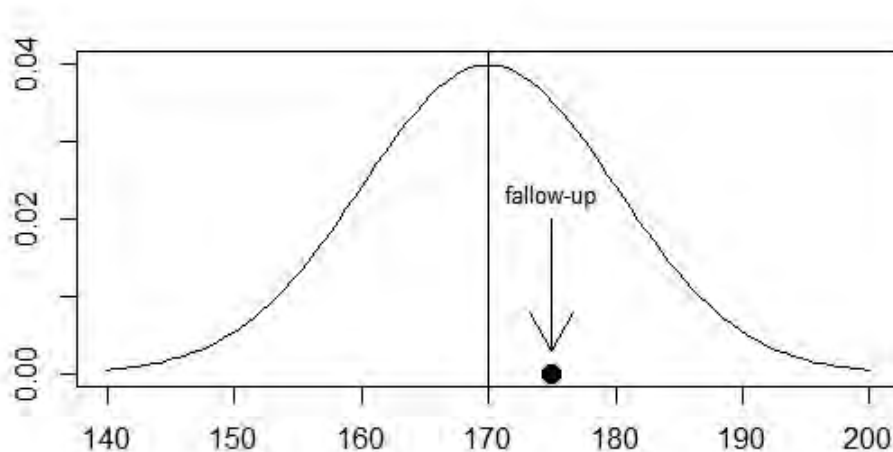


Figure 1.3: Height of Participants with height of an individual whose parent's height was initially extreme

Researchers are also interested in identifying the treatment effect on a group of individuals with quantitative characteristics, for example the estimation of the drug's impact on subjects with high blood pressure. A typical analysis that ignores regression towards the mean will lead to fallacious estimates of treatment effect. Also, subjects from a "high risk" category, defined as those with a value of the characteristic of interest greater than a cut-off value (such as diastolic blood pressure greater than 90 mmHg) or a value greater than a pre-selected population percentile, are included in some medical and biologic studies (such as greater than the 95th percentile). For example, such type situations occur in selecting hypertensive or hypercholesterolemic patients. Occasionally, the characteristic of interest can only be observed with significant error and the design of a reliable selection procedure becomes a serious issue. The design leads to obtaining subjects with the desired characteristics as well as regression towards the mean (McMAHAN, 1982).

In psychological studies the researchers mostly use repeated measurements due to this there is a high possibility that an RTM effect could bias the conclusion of an intervention study. For example, Yu and Chen (2015) showed evidence in favor of the efficiency of social conformity and unrealistic optimism effects, however, there is not a significant effect after controlling the effect of RTM. Burke et al. (2014) studied the HealthMPower program's for the prevention of childhood obesity through a holistic three years program in elementary school, HealthMPower program has a significant effect to reduce the risk of childhood obesity. Skinner et al. (2015) clarified that HealthMPower causes a favorable change in

body composition is erroneous because the findings obtained are most likely due to the RTM effect. Another study in literature by [Moore et al. \(2018\)](#) concludes that the Parenting, Eating, and Activity for child health (PEACH), an investigative program was significant in the reduction of standardized BMI and waist scores. ([Hannon et al., 2018](#)) claimed that the finding was incorrect since the reported reduction was most likely caused by RTM effect.

Random fluctuations or measurement errors in a subject increase the effect of regression to the mean. RTM is a common problem in data analysis since data without random errors are rare in real life. Further, the effect of RTM is proportional to the measure of the dispersion of the random error component ([Shahane et al., 1995](#)). RTM can also occur when subjects with a particular attribute are chosen for study based on their baseline measurements at the extreme of a distribution ([Shahane et al., 1995](#); [Johnson and George, 1991](#)).

## 1.2 Historical Background

[Galton \(1886\)](#) first studied phenomena of the RTM about a century ago. It was discovered in pea growth experiments that offspring from tall plants were shorter than either of the parent plants. Likewise, the offspring of two shorter plants were to be taller than any of their parents on average. According to Galton this phenomenon is referred to as 'regression towards mediocrity'.

[James \(1973\)](#) derived a formula for RTM to accurately estimate the effect of true treatment and RTM effects for the case of a bivariate normal distribution which has been truncated on the basis of baseline measurements. The results demonstrate that if the correlation between before and after treatment is small then the regression effect will be large. Also, the method of moments is used for parameter estimation in the by assuming the fraction in the truncated portion to be known.

[Davis \(1976\)](#) extended the derivation of RTM expression when numerous measurements were collected before giving treatment to patients and highlighted how these techniques were beneficial in minimizing the RTM effect.



[Johnson and George \(1991\)](#) proposed a model to estimate the RTM effect in the presence of correlated within-subject variability and independent random error.

[Barnett et al. \(2005\)](#) introduced the RTM problem and illustrate practical approaches to deal with the problem during the design and analysis stages. The results suggested that the influence of RTM in a sample become more evident when measurement error increases and when follow-up measurements are only preferred on a sub-sample selected using a baseline value

[Khan and Olivier \(2018\)](#) studied the effect of regression to the mean in the case of both homogenous and inhomogeneous Poisson processes. The expression to quantify the effect of regression to the mean for bivariate Poisson distribution for both cases has been derived. Through the method of the maximum likelihood, the estimator of the RTM effect is derived and these estimators were shown to be consistent, unbiased, and approximately normal. The result also suggested that for both cases the effect of RTM is different due to the equality of mean and variance of the distribution.

Similarly, for the case of binary data, [Khan and Olivier \(2019\)](#) derived the effect of regression to the mean for the bivariate binomial distribution. The results suggested that the RTM is severe, whenever there is a negative correlation coefficient, whereas the correlation coefficient is always positive in normal and Poisson distribution. Results show that the change in the number of nonconforming cardboards is due to RTM which couldn't be due to the intervention effect. The treatment effect obtained by subtracting the RTM effect from the total observed effect was biased due to the dependency of the true and random error component.

[Müller et al. \(2003\)](#) derived the generalization of regression to the mean paradigm in the nonparametric situation, where both population distribution of the given observation and also the contaminating errors are unknown. The method of the emergence of nonparametric in regression to the mean with data shrinkage idea also used to find the mode of the distribution. Results suggested that plug-in approaches were used for the estimation of smoothing parameters.

Beath and Dobson, 1991 studied the estimation of the effect of RTM in non-normal distribution based on Edge-worth series and saddle point approximation. Results concluded that Edge-worth series approximation to estimates the effect of RTM is considered accurate as compare to Gram-Charlie approximation. From the results, it is clear that saddle-point approximation is more accurate but is computationally difficult.

The lognormal distribution is an important continuous distribution in statistics. The distribution has a wide range of uses in biological and medical sciences. A lognormal distribution is widely used to characterize financial asset distributions, such as share and stock prices. Since asset value can't be negative lognormal distribution is well suited for this purpose. Also, the lognormal distribution finds the widest variety of applications in ecology. Modern ecological science focuses on the consequences of most species on the planet's great capacity for growth (Crow and Shimizu, 1987). There are many real-life situations which fallows bivariate lognormal distribution. Some of them are briefly discussed below.

Yue (2000) used the multivariate lognormal probabilistic model for the prediction of flood-frequency analysis. It has been examined that lognormal distribution is often a possible choice for flood frequency analysis. The pair of mutually correlated variables used for the flood-frency analysis are the peak volume and duration.

Yue (2002) has also used the bivariate lognormal as a probabilistic model for the prediction of storm events, such events are characterized by their peak and the total amount which are mutually correlated.

Yerel and Konuk (2009) have model impurities in magnesite ore deposits through bivariate lognormal distribution. The impurities in the magnesite ore deposits with a higher grade than the cut-off grade are considered for the plant process.

Dehghani and Fadaee (2020) used the bivariate lognormal distribution as a probabilistic model for earthquake prediction purposes. The pair of correlated variables were earthquake magnitude and recurrence time, respectively.

## 1.3 Objectives

The objectives of this thesis are:

1. Derivation of formula for quantifying the Total, RTM, and treatment effect for the BLD.
2. Estimation of the Total, RTM, and treatment effects using the method of Maximum likelihood estimation.
3. Comparison with some existing methods using simulation.

# Chapter 2

## Literature

In bivariate distribution regression to the mean effect is accounted for to accurately estimate the treatment effect. Researchers have developed several measures to extract the RTM effect in various distributions. In this chapter, some of the work related to quantifying the RTM effect in literature has been discussed.

### 2.1 RTM under the bivariate normal distribution

The literature below focused on pre and post variable which follows bivariate normal distribution and the variable must be positively correlated and also fulfilled the assumption of stationary.

#### 2.1.1 James (1973)

Suppose the random variable  $X$  and  $Y$  represent measurements of the before and after treatment of a patient. Both  $X$  and  $Y$  are normally distributed  $N(\mu, \sigma^2)$ . Only patients with above truncation point  $x_0$  are considered for treatment and the treatment effect is measured by comparing the pre-post mean. The density function of truncated bivariate standardized normal distribution is,

$$f(x, y) = \frac{[1 - \Phi(x_0)]^{-1}}{2\pi\sqrt{1 - \rho^2}} \exp \left[ \left( \frac{-1}{2(1 - \rho^2)} \right) \left( \left( \frac{x - \mu_x}{\sigma_x} \right)^2 + \left( \frac{y - \mu_y}{\sigma_y} \right)^2 - 2\rho \left( \frac{x - \mu_x}{\sigma_x} \right) \left( \frac{y - \mu_y}{\sigma_y} \right) \right) \right] \quad (2.1)$$

The mean and variance of the variable representing the pre-treatment are, respectively, given by

$$E(X | X \geq x_0) = \mu + \frac{\phi(z_0)}{1 - \Phi(z_0)}\sigma,$$

$$Var(X | X \geq x_0) = \sigma^2 \left[ \frac{\phi(z_0)}{1 - \Phi(z_0)} \left( z_0 - \frac{\phi(z_0)}{1 - \Phi(z_0)} \right) + 1 \right].$$

Similarly, the mean and variance of post-treatment are

$$E(Y | X \geq x_0) = \mu + \frac{\phi(z_0)}{1 - \Phi(z_0)}\rho\sigma,$$

$$Var(Y | X \geq x_0) = \sigma^2 \left[ \rho^2 \frac{\phi(z_0)}{1 - \Phi(z_0)} \left( z_0 - \frac{\phi(z_0)}{1 - \Phi(z_0)} \right) + 1 \right].$$

The effect of regression to mean (RTM) was derived by [James \(1973\)](#) as

$$E(Y - X | X \geq x_0) = \frac{\phi(z_0)}{1 - \Phi(z_0)}\sigma(\rho - 1)$$

where,  $\mu$  and  $\sigma$  are the unconditional mean and standard deviation of random variable  $X$  and  $Y$  and  $\phi(\cdot)$  and  $\Phi(\cdot)$  are the respective density and distribution functions of the the random variable  $X$ .

Usign the sample conditional means  $(\bar{x}, \bar{y})$  and variances  $(S_x^2, S_y^2)$ , and the sample regression coefficient  $b_{yx}$ , [James \(1973\)](#) estimated the parameters  $(\mu, \sigma^2, \rho, \gamma)$  of the model through the method of moments, and are given below.

$$\begin{aligned} \hat{\mu} &= \bar{x} - \frac{\phi(z_0)}{1 - \Phi(z_0)}\hat{\sigma} \\ \hat{\sigma}^2 &= \frac{S_x^2}{\left[ \frac{\phi(z_0)}{1 - \Phi(z_0)} \left( x_0 - \frac{\phi(z_0)}{1 - \Phi(z_0)} \right) + 1 \right]} \\ \hat{\rho} &= \left[ b^2 \left( \frac{\phi(z_0)}{1 - \Phi(z_0)} \left( x_0 - \frac{\phi(z_0)}{1 - \Phi(z_0)} \right) + 1 \right) - \frac{S_y^2}{\hat{\sigma}^2} + 1 \right]^{\frac{1}{2}} \\ \hat{\gamma} &= \frac{b}{\hat{\rho}}, \end{aligned}$$

where  $(\mu, \sigma^2, \rho, \gamma)$  are the mean, variance, correlation coefficient, and treatment parameters.

### 2.1.2 Davis (1976)

Davis (1976) introduced various methods to mitigate the regression to the mean effect. Let  $y_i$  be the measure of variable of interest the cutoff point is  $k$ . Assume that variable of interest  $y_i$  is normally distributed  $N(\mu, \sigma)$ , and  $\rho_{ij}$  is the correlation coefficient between  $i^{th}$  and  $j^{th}$  measure. Then

$$E(Y_1 | y_1 > k_1) = \mu + c_1\sigma$$

where,  $c_1 = \frac{\phi(Z_1)}{1-\Phi(Z_1)}$  and  $Z = \frac{y-\mu}{\sigma}$ .

Similarly, the mean of  $y_2$  given that  $y_1$  exceeds  $k_1$  of the distribution is given below.

$$E(Y_2 | y_1 > k_1) = \mu + \rho c_1\sigma$$

The effect of Regression to the mean is,

$$\begin{aligned} RTM &= E(Y_1 | y_1 > k_1) - E(Y_2 | y_1 > k_1) \\ &= c_1\sigma(1 - \rho) \end{aligned}$$

Davis (1976) propose a model-based multiple measurement method, through the use of two or even more observations on the same subject. Multiple measurements on the same individual prior to performing an intervention can substantially decrease the quantity of RTM.

By the RTM equation we determine that if the  $\rho$  value is unity than there may be no regression to the mean impact. And as  $\rho$  end up smaller the impact of regression to the mean become significant. The result also shows that regression had a sizable effect which leads to fallacious conclusions concerning treatment effects.

### 2.1.3 Gardner and Heady (1973)

Gardner and Heady (1973) adress the impact of within-person variability, which is the variability of repeated observations of the same variable on the same individual at various times. Consider a basic model in which the component to be evaluated is expected to be subject to an additive error, or deviation from the true' value. let,

$$Y_i = Y_0 + \epsilon$$

where  $Y_0$  is the actual level of the variable for a specific participant and  $Y$  is the empirical observation. Assume that the  $\epsilon$  is  $N(0, \delta_0^2)$ , and also independent of the true value of  $y$ . And,  $Y_0$  is distributed  $N(\mu, \sigma^2)$  with variance  $\lambda^2 = \sigma^2 + \delta_0^2$  and correlation  $\text{corr}(Y, Y_0) = \sigma/\sqrt{\sigma^2 + \delta_0^2}$ . The truncated distribution of observed values within a certain range is,

$$f(Y | Y > L) = \frac{f(y)\phi(\frac{Y-\mu}{\lambda})}{1 - \Phi(\frac{L-\mu}{\lambda})}$$

where  $\phi(u)$  is the standardized normal PDF and  $\Phi(u)$  is the CDF at an specific point  $u$ .

That is

$$\phi(u) = \frac{\exp(-\frac{u^2}{2})}{\sqrt{2\pi}}$$

$$\Phi(u) = \int_{-\infty}^u \phi(t)dt$$

A truncated Normal curve with no observations below  $Y_L = l$  is,

$$v(u) = \frac{\phi(u)}{1 - \Phi(u)},$$

The expected value of  $Y$  given  $L$  in the truncated distribution is,

$$E(Y | Y > L) = \mu + \epsilon v(\frac{L - \mu}{\lambda}).$$

Similarly, the function of  $y$  is given

$$f(y | Y > L) = \frac{\int_L^\infty f(y, Y)dX}{\int_L^\infty f(y)dy}$$

$$= \frac{\phi(\frac{y-\mu}{\sigma})[1 - \Phi(\frac{L-y}{\delta_0})]}{1 - \Phi(\frac{L-\mu}{\lambda})}$$

It is useful for comparative purposes to write (1) and (2) respectively as

$$E(y | Y > L) = \mu + \frac{\sigma^2}{\lambda} v(\frac{L - \mu}{\lambda})$$

$$E(Y | Y > L) = \mu + \frac{\lambda}{\sigma} v\sigma(\frac{L - \mu}{\lambda})$$

It has been observed that the mean of the observed values is beyond the mean of true value  $y$ , since  $\lambda > \sigma$  except where  $\delta_0$  is zero. This is the RTM effect.

### 2.1.4 Shahane et al. (1995)

Shahane et al. (1995) discussed two more scenarios: sampling from a truncated bivariate normal distribution based on (i) at least one variable surpasses the threshold or cut-off point (ii) a linear combination of the two variables that exceeds a certain threshold. Under the measurement error of Gardner and Heady (1993) and the subject effect model of Johnson and George (1991), the author derived the expected regression effects for both scenarios.

Gardner and Heady (1973) describe that X and Y can be defined as the sum of the two components.

$$X_i = U_1 + e_{1i} \quad \text{and} \quad Y_i = U_2 + e_{2i}$$

The effect of regression towards the mean is,

$$R(X_2 | T) = \frac{\sigma_{e1}^2 \phi(\alpha) \Phi(A)}{\sigma_x P(T)}$$

$$R(Y_2 | T) = \frac{\sigma_{e2}^2 \phi(\beta) \Phi(B)}{\sigma_y P(T)}$$

$$T = (X_1 > k_1 \text{ and } Y_1 > k_2)$$

where,  $\alpha = \frac{k_1 - \mu_1}{\sigma_x}$ ,  $\beta = \frac{k_2 - \mu_2}{\sigma_y}$ ,  $A = \frac{(\beta - \rho\alpha)}{\sqrt{1 - \rho^2}}$  and  $B = \frac{(\alpha - \rho\beta)}{\sqrt{1 - \rho^2}}$

The Johnson and George (1991) model regression effect can be minimized by taking replicated measurements. Assume we identify the subjects for eligibility using the mean of X and Y of n replication measurements.

$$R(X_{(n+1)} | T_n) = \frac{\sigma_{e1}^2 \phi(\alpha) \Phi(A)}{nP(A)\sigma_{\bar{x}}}$$

$$R(Y_{(n+1)} | T_n) = \frac{\sigma_{e2}^2 \phi(\beta) \Phi(A)}{nP(B)\sigma_{\bar{y}}}$$

Johnson and George (1991) introduced a model, where an additional component the effect of "within subject" variability is defined, namely  $S_{ij}$  the subject effect also assumed to lead to regression to the mean. For this the model we have,

$$X_{ij} = U_1 + S_{1i} + e_{1ij}$$

$$Y_{ij} = U_2 + S_{2i} + e_{2ij}$$



The regression effect we have

$$R_x = R(X_{(mn+1)}) | T_{mn} = \frac{n\sigma_{s1}^2(1 - \rho_{s1}) + \sigma_{e1}^2}{mn\sigma_{\bar{x}}} \left[ \frac{\phi(\alpha)\Phi(A)}{P(T_{mn})} \right]$$

$$R_x = R(Y_{(mn+1)}) | T_{mn} = \frac{n\sigma_{s2}^2(1 - \rho_{s2}) + \sigma_{e2}^2}{mn\sigma_{\bar{y}}} \left[ \frac{\phi(\beta)\Phi(B)}{P(T_{mn})} \right]$$

The RTM effect due to measurement error can be reduced by increasing  $m$  and  $n$  number of replicates but the RTM effect in case of within subject variability is only reduced by increasing  $m$ .

Under the model of linear function of the bivariate response  $(X_i, Y_i)$  with the known constants  $a$  and  $b$  such as

$$Z_i = aX_i + bY_i \quad i=1,2$$

Under Gardner and Heady model (1973), the expected regression effect on  $Z_2$ , conditioned on the truncation point  $T = (Z_1 > k_1)$

$$R(Z_2 | T) = \frac{a^2\sigma_{e1}^2 + b^2\sigma_{e2}^2}{\sqrt{a^2(\sigma_u^2 + \sigma_{e1}^2) + b^2(\sigma_v^2 + \sigma_{e2}^2) + 2ab\rho_{uv}\sigma_u\sigma_v}} \left[ \frac{\phi(\alpha)}{Q(c)} \right]$$

From the results, we conclude that the correlation between the subject effect increases and the effect of regression towards mean becomes smaller and vice versa.

## 2.2 RTM for non-normal population

### 2.2.1 Das and Mulder (1983)

Das and Mulder (1983) introduced a simple generic formula to estimate the effect of RTM for an arbitrary random variable of the stationary population of subjects. The most necessary assumption here is that the within-subject variance (the disturbance) is normally distributed. Suppose  $X$  is a continuous random variable which is measured twice on the individual subjects of a population. The subjects of stationary population both  $X_1$  and  $X_2$  are considered to have a common distribution with identical mean  $\mu_x$ , variance  $\sigma_x^2$  and the correlation coefficient  $\rho$ . The quantify effect of regression to the mean of a normally distributed variable is,

$$E(X_1 - X_2 | X_1 = x_1) = (1 - \rho)(x - \mu_x)$$

In case of without time effect stationary measurement model. In this model, the length of time between the two measurements  $X_1$  and  $X_2$  is considered to have no influence.

$$X_j = W + E_j \quad j = 1, 2$$

Where  $W$  is the true mean and an error term  $E$ . So, the effect of regression to the mean is,

$$R(x) = -(1 - \rho)\sigma_x^2 \frac{1}{g(x)} \frac{dg(x)}{dx}$$

In the case of a unimodal density  $g$ , it is more acceptable to refer to regression to the mode instead of regressions to the mean. In the normality of density function  $g(x)$ , the regression effect reduced in linear form, so the mode equal to mean.

In case of with time effect stationary measurement model. In this model, the length of time between the two measurements  $X_1$  and  $X_2$  is considered to have a significant effect.

$$\rho_T = \rho + (1 - \rho)\gamma_T$$

The effect of regression to the mean is,

$$\begin{aligned} R_T(x) &= -(1 - \rho_T)\sigma_x^2 \frac{d}{dx} \ln [g(x)] \\ &= -(1 - \rho)(1 - \gamma_T)\sigma_x^2 \frac{d}{dx} \ln [g(x)] \end{aligned}$$

Where  $\gamma_T$  represent the correlation coefficient between the disturbance term  $E_1$  and  $E_2$  with  $\lim_{T \rightarrow \infty} \gamma_T = 0$  and  $\rho = \frac{\sigma_w^2}{\sigma_x^2} = \lim_{T \rightarrow \infty} \rho_T$ .

### 2.2.2 Beath and Dobson (1991)

Beath and Dobson (1991) have quantified the regression to the mean of non-normal distributions using Edgeworth series and saddlepoint approximation. Consider random variables  $X_1$  and  $X_2$  that represent consecutive measurements on the identical respondents. It is determined that  $X_1 = M + e_1$  and  $X_2 = M + e_2$ , where  $M$  is the random variable reflecting the individual's 'actual' value and  $e_1$  and  $e_2$  are random variables reflecting measurement mistakes or within-subject variability, where  $M$ ,  $e_1$  and  $e_2$  are assumed to be mutually independent. The random variable  $M$  had arbitrary probability density

function  $f(m)$  with mean  $\mu$  and variance  $\theta^2 = \rho\sigma^2$  and the variable  $e_1$  and  $e_2$  are normally distributed with mean 0 and  $\Delta^2$ , where  $\delta^2 = (1 - \rho)\sigma^2$ , then  $X_1$  and  $X_2$  are also normally distributed  $N(\mu, \sigma^2)$ .

Assume the individuals are picked on the basis that the initial measurement  $X_1$ , was more than truncated point  $X_L$ . Then the use of the technique of Das and Mulder (1983) the regression to the mean are,

$$R(x_L) = E(X_1 - X_2 | X_1 > x_L) = \frac{(1 - \rho)\sigma^2 g(x_L)}{1 - G(x_L)}$$

where  $g(x)$  is the PDF of random variable  $X_1$  and  $X_2$  given by,

$$g(x) = \frac{1}{\Delta} \int_{-\infty}^{\infty} f(m) \phi\left(\frac{x - m}{\Delta}\right) dm$$

From Edgeworth approximation method the feasible estimates of  $g(x_L)$  the PDF and  $G(x_L)$  the distribution function is,

$$\begin{aligned} g(x_L) &= \frac{1}{\sigma} \phi\left(\frac{x_L - \mu}{\sigma}\right) \sum_j \left(\frac{\theta}{\sigma}\right)^j H_j\left(\frac{x_L - \mu}{\sigma}\right) \\ G(x_L) &= \sum_{j \geq 0} c_j \left(\frac{\theta}{\sigma}\right)^j \int_{-\infty}^{\frac{x_L - \mu}{\sigma}} \phi(v) H_j(v) dv \\ &= \Phi\left(\frac{x_L - \mu}{\sigma}\right) - \phi\left(\frac{x_L - \mu}{\sigma}\right) \sum_{j \geq 0} c_j \left(\frac{\theta}{\sigma}\right)^j H_{j-1}\left(\frac{x_L - \mu}{\sigma}\right) \end{aligned}$$

The Edgeworth series give sometime negative values of multimodel approximation for some specific values of skewness and kurtosis. In this case the feasible estimation of  $R(x_L)$  through saddlepoint method, which provides an better fit to the probability density function. The approximation of measurement distribution  $g(x)$  through saddlepoint method is,

$$V(x) = \frac{\exp\{k(t_0) - t_0 x\}}{\{2\pi k''(t_0)\}^{\frac{1}{2}}}$$

Where  $K(t_0)$  is the cumulative generating function and  $K'(t_0) = x$ .

The methods presented in the literature are not applicable if the underlying distribution is nonnormal. In this case, the approximation methods described in this paper provides

feasible results.

### 2.2.3 John and Jawad (2010)

As previously stated, the Das and Mulder (1983) approach can't be applied to evaluate the RTM effect in case of an empirical distribution. The John and Jawad (2010) decided to utilize the method proposed by Das and Mulder (1983) data adaptive through kernel density estimation and kernel estimation algorithms for the hazard rate function. The term for assessing the mean regression for high initial levels subjects as suggested by Das and Mulder is,

$$R(X_L) = E((X_1 - X_2) | X_1 > x_L) = \frac{(1 - \rho)\sigma^2 g(x_L)}{1 - G(x_L)} \quad (1)$$

Kernel density estimations have been thoroughly researched and applied in a variety of applications across the literature. For a given set of starting values of  $X_{1i}$ , the probability density estimator kernel function  $g$  is provided.

$$\hat{g}_s(x) = n^{-1} \sum K_s(x - X_{1i})$$

$$K_s(\cdot) = \frac{1}{h} K(\cdot/h)$$

where "s" represents the smoothing parameter and  $k_s(\cdot)$  is the kernel function. The mean integrated square error (MISE) is a popular method of determining the estimated error of  $\hat{g}_h(x)$  is,

$$MISE(s) = E \int (\hat{g}_h - g)^2$$

And the Asymptotic mean integrated square error (AMISE) is given

$$AMISE(s) = R(K)/ns + s^4 R(g'') \left( \int_{-\infty}^{\infty} x^2 K(x) dx / 2 \right)^2$$

It is the optimal value of hAMISE that minimises the AMISE(s) and also effectively approximates the optimum value of hMISE, which results in the optimal value of MISE

(s) and can be calculated as

$$h_{AMISE} = \left( \frac{R(K)}{nR(g'')(\int_{-\infty}^{\infty} x^2 K(x) dx / 2)^2} \right)^{1/5}.$$

### 2.2.4 Khan and Olivier (2018)

Khan and Olivier (2018) focused to quantify the expression for the effect of RTM in of bivariate Poisson distribution for both cases of Poisson process homogeneous and inhomogeneous.

*The RTM effect for the bivariate Poisson assuming a right cut-off point*

$$R_r(y_0; \theta) = \theta_1 \frac{1 - F(y_{0-1} | \theta_0 + \theta_1)}{1 - F(y_0 | \theta_0 + \theta_1)} - \theta_2$$

*The RTM effect for the bivariate Poisson assuming a left cut-off point*

$$R_l(y_0; \theta) = \theta_2 - \theta_1 \frac{1 - F(y_{0-1} | \theta_0 + \theta_1)}{1 - F(y_0 | \theta_0 + \theta_1)}$$

Moreover, the expression for intervention/treatment effect has been derived,

$$\delta(\theta) = \theta_1 - \theta_2$$

In case of the null intervention effect meaning the pre and post observations are identically distributed, the authors obtained the RTM effect by putting  $\theta_2 = \theta_1$

$$R_r(y_0; \theta) = \theta_1 \frac{1 - F(y_{0-1} | \theta_0 + \theta_1)}{1 - F(y_0 | \theta_0 + \theta_1)} - \theta_1$$

Finally, the total effect has been derived as following,

$$\begin{aligned} T(y_0; \theta) &= R_r(y_0 + \delta(\theta)) \\ T(y_0; \theta) &= \left[ \theta_1 \frac{1 - F(y_{0-1} | \theta_0 + \theta_1)}{1 - F(y_0 | \theta_0 + \theta_1)} - \theta_1 \right] + [\theta_1 - \theta_2] \end{aligned}$$

The simulation results suggested that the maximum likelihood estimator of RTM is consistent, unbiased and also approximately normally distributed. The results also show that the RTM effect for the homogenous Poisson process is different from the

inhomogeneous Poisson process due to the equality of mean and variance of the distribution.

### 2.3 Designing studies to mitigate the RTM problem

In intervention studies, the study design can assist to reduce the RTM effect ([Yudkin, 1996](#)). The subsection that follow explain some well-known study designs and their possible impact on RTM.

#### 2.3.1 Controlled trials with randomization

If subjects are randomly allocated to treatment and control groups, this could help in reducing the RTM effect in the sense that any change in the control group is considered as the RTM effect. This effect upon subtraction from the total change in the treatment group can help in separating the treatment effect. However, due to ethical constraints the randomization process is not always achievable [Khan \(2019\)](#).

#### 2.3.2 Two measurements approach of [Ederer \(1972\)](#)

[Ederer \(1972\)](#) has proposed two measurement approaches for mitigating the effect of RTM. The initial measurement is used to select participants, and the second measurement is used as a baseline against which the treatment impact is measured. The RTM is assumed to occur between the first and second measurements, so the intervention effect is evaluated as the mean change from baseline.

# Chapter 3

## Derivation of the Total/RTM effects

The normal distribution is considered as the backbone of statistical inference due to its applications in diverse research areas. Likewise, the log normal distribution is also an important continuous distribution and has a wide range of uses in biological and medical sciences, financial asset financial asset distributions and stock prices etc (Kenton, 2020). In the field of radiation protection, the lognormal distribution has is frequently used (Gale, 1967). Lognormal distribution has also applications in quality control, in cases when the conventional quality control process failed, the modified quality control method led to a relatively correct interpretation of the data (Morrison, 1958). Using lognormal control instead of normal control is recommended for skewed data (Morrison, 1958). Similarly, in radiological and environmental studies, it's known that radiological data are positive and significantly skewed and can be modeled by the lognormal distribution (Blackwood, 1992). The expression of RTM under the bivariate lognormal distribution is missing in literature. To derive its expression, consider the density function of bivariate lognormal distribution as

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \frac{1}{xy} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right]. \quad (3.2)$$

Usually, the intervention is applied to subjects on a baseline criterion which could be in the right or left tail of a distribution, thereby leading to a truncated distribution. Let the pre post observations be  $X$  and  $Y$ , then the truncated bivariate lognormal distribution of  $X$  and  $Y$  is given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \frac{x^{-1}y^{-1}}{1-\Phi(x_0)} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] \quad (3.3)$$

where,  $F_X(x) = \Phi(x_0)$  is the cumulative density function (CDF) of standard normal distribution.

In general the Total effect is quantified as the difference between the conditional expectation of  $X$  and  $Y$ , denoted by  $T_r(x_0, \alpha)$  and given by

$$T_r(x_0, \theta) = E(X - Y | X > x_0) = E(X | X > x_0) - E(Y | X > x_0). \quad (3.4)$$

Similarly, the expression for RTM for a right cut-off point is given by  $R_r(x_0, \alpha)$  and given by

$$R_r(x_0, \theta) = E(X - Y | X > x_0, E(X) = E(Y)), \quad (3.5)$$

where  $E(X)$  and  $E(Y)$  are the unconditional means of the bivariate lognormal distribution. This can be interpreted as the total effect under identically distributed pre-post variables, i.e.,  $\mu_x = \mu_y$  and  $\sigma_x = \sigma_y$ .

As the lognormal distribution is skewed one, so the behaviour of RTM would be different for the left and right cut-off points. Both cases are separately considered.



### 3.1 Case 1: Right cut-off point

Assume that an intervention or treatment group is determined on the basis of the variable  $X$  greater than the threshold point  $x_0$ , then the conditional expectation of  $X$  is,

$$\begin{aligned}
 E(X | X > x_0) &= \frac{[1 - \Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \int_{x_0}^{\infty} \int_0^{\infty} \frac{1}{y} \exp \left[ \frac{-1}{2(1 - \rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\
 &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\
 &= \frac{1}{2\pi\sqrt{1 - \rho^2}} \frac{1}{1 - \Phi\left(\frac{\log(x_0) - \mu_x}{\sigma_x}\right)} \int_{\frac{\log(x_0) - \mu_x}{\sigma_x}}^{\infty} \int_{-\infty}^{\infty} \exp(\mu_x + z\sigma_x) \\
 &\quad \exp\left(\frac{-1}{2(1 - \rho^2)}\right) (z^2 + p^2 - 2\rho zp) dp dz
 \end{aligned}$$

This can be simplified to the following expression

$$E(X | X > x_0) = \frac{\exp(\mu_x)}{\sqrt{2\pi} \left[ 1 - \Phi\left(\frac{\log(x_0) - \mu_x}{\sigma_x}\right) \right]} \int_{\frac{\log(x_0) - \mu_x}{\sigma_x}}^{\infty} \exp\left(\frac{-1}{2} (z^2 + 2z\sigma_x \pm \sigma_x^2)\right) dz$$

After further simplification the conditional expectation of  $X$  we have,

$$E(X | X > x_0) = E(X) \frac{1 - \Phi(z_0 - \sigma_x)}{1 - \Phi(z_0)} \tag{3.6}$$

where,  $E(X) = \exp\left(\mu_x + \frac{\sigma_x^2}{2}\right)$

Similarly, solving the conditional expectation of  $E(Y|X > x_0)$  by using the same procedure

$$\begin{aligned}
 E(Y | X > x_0) &= \frac{[1 - \Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \int_{x_0}^{\infty} \int_0^{\infty} \frac{1}{x} \exp \left[ \frac{-1}{2(1 - \rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\
 &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\
 &= \frac{\left[ 1 - \Phi\left(\frac{\log(x_0) - \mu_x}{\sigma_x}\right) \right]^{-1}}{2\pi\sqrt{1 - \rho^2}} \int_{\frac{\log(x_0) - \mu_x}{\sigma_x}}^{\infty} \int_{-\infty}^{\infty} \exp(\mu_y + p\sigma_y) \\
 &\quad \exp\left(\frac{-1}{2(1 - \rho^2)}\right) (z^2 + p^2 - 2\rho zp \pm \rho^2 z^2) dp dz
 \end{aligned}$$

After simplification the preceding expression can be shown to be given by

$$E(Y | X > x_0) = E(Y) \frac{1 - \Phi(z_0 - \rho\sigma_x)}{1 - \Phi(z_0)} \quad (3.7)$$

where,  $E(Y) = \exp\left(\mu_y + \frac{\sigma_y^2}{2}\right)$ .

We get the total effect of bivariate lognormal for a right cut-off point by substituting Equation (3.5) and (3.6) in (3.4) is given

$$T_r(x_0, \theta) = E(X) \left[ \frac{1 - \Phi(z_0 - \sigma_x)}{1 - \Phi(z_0)} \right] - E(Y) \left[ \frac{1 - \Phi(z_0 - \rho\sigma_x)}{1 - \Phi(z_0)} \right] \quad (3.8)$$

where the expectations of  $X$  and  $Y$  are the unconditional means of the respective univariate lognormal distributions of  $X$  and  $Y$  as  $E(X) = \exp(\mu_x + \sigma_x^2/2)$  and  $E(Y) = \exp(\mu_y + \sigma_y^2/2)$ , and

$$\theta = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho_{xy}).$$

## 3.2 Case 2: Left cut-off point

Assume that an intervention or treatment group is formed on the basis of the variable  $X$  less than the threshold point  $x_0$ , then the conditional expectation of  $X$  is

$$\begin{aligned} E(X | X < x_0) &= \frac{[\Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_0^{x_0} \int_0^\infty \frac{1}{y} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\ &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\ &= \frac{1}{2\pi\sqrt{1-\rho^2}} \frac{1}{\Phi\left(\frac{\mu_x - \log(x_0)}{\sigma_x}\right)} \int_{-\infty}^{\frac{\mu_x - \log(x_0)}{\sigma_x}} \int_{-\infty}^\infty \exp(\mu_x + z\sigma_x) \\ &\quad \exp\left(\frac{-(z^2 + \rho^2 - 2\rho zp)}{2(1-\rho^2)}\right) dp dz \end{aligned}$$

After simplification the expression of conditional mean of  $X$ , we get

$$E(X | X < x_0) = E(X) \frac{[\Phi(-z_0 - \sigma_x)]}{[\Phi(-z_0)]} \quad (3.9)$$

Similarly, solving the conditional expectation of  $E(Y|X < x_0)$  and following the same steps we get

$$\begin{aligned}
 E(Y | X < x_0) &= \frac{[\Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_0^{x_0} \int_0^\infty \frac{1}{x} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\
 &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\
 &= \frac{\exp(\mu_y)}{2\pi\sqrt{1-\rho^2}\Phi\left(\frac{\mu_x - \log(x_0)}{\sigma_x}\right)} \int_{-\infty}^{\frac{\mu_x - \log(x_0)}{\sigma_x}} \exp\left(\frac{-z^2}{2}\right) \\
 &\quad \left( \int_{-\infty}^\infty \exp(p\sigma_y) \exp\left\{\frac{-(p - z\rho)^2}{2(1-\rho^2)}\right\} dp \right) dz
 \end{aligned}$$

After simplification the preceding expression of conditional mean of Y reduces to

$$E(Y | X < x_0) = E(Y) \frac{\Phi(-z_0 - \rho\sigma_x)}{\Phi(-z_0)} \quad (3.10)$$

Following the similar steps for left cut-off point, and using the definition of the total effect, we get the resulting expression of the total effect for the left cut-off point as

$$T_l(x_0, \theta) = E(X) \left[ \frac{\Phi(-z_0 - \sigma_x)}{\Phi(-z_0)} \right] - E(Y) \left[ \frac{\Phi(-z_0 - \rho\sigma_x)}{\Phi(-z_0)} \right] \quad (3.11)$$

### 3.2.1 Expression for RTM

The expressions of RTM for both left and right cut-off point can be obtained by substituting  $\mu_x = \mu_y$  and  $\sigma_x = \sigma_y$  in the respective equations (3.8) and (3.11).

The expression of RTM effect for right cut-off point is

$$R_r(x_0, \theta) = E(X) \left[ \frac{1 - \Phi(z_0 - \sigma_x)}{1 - \Phi(z_0)} \right] - E(X) \left[ \frac{1 - \Phi(z_0 - \rho\sigma_x)}{1 - \Phi(z_0)} \right]. \quad (3.12)$$

Likewise, the RTM effect of left cut-off point is

$$R_l(x_0, \theta) = E(X) \left[ \frac{\Phi(-z_0 - \sigma_x)}{\Phi(-z_0)} \right] - E(X) \left[ \frac{\Phi(-z_0 - \rho\sigma_x)}{\Phi(-z_0)} \right]. \quad (3.13)$$

### 3.3 Expressions for Variances

In this section we derive expressions for the variances of the total and RTM effects. Details are given in the following subsections.

#### 3.3.1 Variance of the total effect

The variance of the total effect in case of right cut-off point  $T_r(x_0, \theta)$  and  $R_r(x_0, \theta)$  can be calculate by merging the  $Var(X|X > x_0)$ ,  $Var(Y|X > x_0)$  and  $Cov(XY|X > x_0)$  as

$$Var(X - Y | X > x_0) = Var(X | X > x_0) + Var(Y | X > x_0) - 2Cov(XY|X > x_0). \quad (3.14)$$

Some important results required to evaluate the variance of the total effect, are given below. The  $E(X^2|X > x_0)$  for right cut-off point is

$$\begin{aligned} E(X^2|X > x_0) &= \frac{[1 - \Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1 - \rho^2}} \int_{x_0}^{\infty} \int_0^{\infty} \frac{x}{y} \exp \left[ \frac{-1}{2(1 - \rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\ &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\ &= \frac{\left[ 1 - \Phi \left( \frac{\log(x_0) - \mu_x}{\sigma_x} \right) \right]^{-1}}{2\pi\sqrt{1 - \rho^2}} \int_{\frac{\log x_0 - \mu_x}{\sigma_x}}^{\infty} \int_{-\infty}^{\infty} \exp(2\mu_x + 4\sigma_x) \exp \left( \frac{-(z^2 + p^2 - 2\rho zp)^2}{2(1 - \rho^2)} \right) dp dz \end{aligned}$$

After simplifying the formula for  $E(X^2|X > x_0)$ , we obtained

$$E(X^2 | X > x_0) = \exp(2(\mu_x + \sigma_x^2)) \frac{1 - \Phi(z_0 - 2\sigma_x)}{1 - \Phi(z_0)} \quad (3.15)$$

Now, we obtain  $Var(X|X > x_0)$  by using equations (3.5) and (3.12)

$$\begin{aligned} Var(X | X > x_0) &= E(X^2 | X > x_0) - (E(X | X > x_0))^2 \\ &= \frac{1}{1 - \Phi(z_0)} \left[ \exp(2(\mu_x + \sigma_x^2))(1 - \Phi(z_0 - 2\sigma_x)) \right. \\ &\quad \left. - \exp(2\mu_x + \sigma_x^2) \frac{(1 - \Phi(z_0 - \sigma_x))^2}{1 - \Phi(z_0)} \right] \end{aligned} \quad (3.16)$$

Likewise, to find the  $Var(Y|X > x_0)$  we also need some results. The  $E(Y^2|X > x_0)$  for right cut-off point is

$$\begin{aligned} E(Y^2 | X > x_0) &= \frac{[1 - \Phi(x_0)]^{-1}}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \int_{x_0}^{\infty} \int_0^{\infty} \frac{y}{x} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 \right. \right. \\ &\quad \left. \left. - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] dx dy \\ &= \frac{[1 - \Phi \left( \frac{\log(x_0) - \mu_x}{\sigma_x} \right)]^{-1}}{2\pi\sqrt{1-\rho^2}} \int_{\frac{\log(x_0) - \mu_x}{\sigma_x}}^{\infty} \int_{-\infty}^{\infty} \exp(2\mu_y + 2\rho\sigma_y) \exp \left( \frac{-(z^2 + p^2 - 2\rho zp)}{2(1-\rho^2)} \right) dp dz \end{aligned}$$

After simplifying the formula for  $E(Y^2|X > x_0)$ , we obtain

$$E(Y^2 | X > x_0) = \exp(2(\mu_y + \sigma_y^2)) \frac{\Phi(z_0 - 2\rho\sigma_y)}{\Phi(z_0)} \quad (3.17)$$

$$\begin{aligned} Var(Y | X > x_0) &= E(Y^2 | X > x_0) - [E(Y | X > x_0)]^2 \\ &= \frac{1}{1 - \Phi(z_0)} \left[ \exp(2(\mu_y + \sigma_y^2))(1 - \Phi(z_0 - 2\rho\sigma_x)) \right. \\ &\quad \left. - \exp(2\mu_y + \sigma_y^2) \frac{(1 - \Phi(z_0 - \rho\sigma_x))^2}{1 - \Phi(z_0)} \right], \end{aligned} \quad (3.18)$$

and the  $Cov(X, Y | X > x_0)$  for right cut-off point is

$$\begin{aligned} Cov(X, Y | X > x_0) &= E(XY | X > x_0) - E(X | X > x_0)E(Y | X > x_0) \\ &= \exp \left( \mu_y + \sigma_y^2 + \rho\sigma_x\sigma_y - \frac{\rho^2\sigma_y^2}{2} \right) \frac{[1 - \Phi(z_0 - \sigma_x - \rho\sigma_y)]}{[1 - \Phi(z_0)]} \\ &\quad - E(Y) \frac{[1 - \Phi(z_0 - \sigma_x)][1 - \Phi(z_0 - \rho\sigma_x)]}{[1 - \Phi(z_0)]^2} \end{aligned} \quad (3.19)$$

Putting equation (3.11), (3.12) and (3.13) in (3.10), we get the variance of Total effect for right cut-off point

$$\begin{aligned} Var(X - Y | X > x_0) &= \frac{1}{1 - \Phi(z_0)} \left[ \exp(2(\mu_x + \sigma_x^2)) (1 - \Phi(z_0 - 2\sigma_x)) - \exp(2\mu_x - \sigma_x^2) \right. \\ &\quad \left. \frac{(1 - \Phi(z_0 - \sigma_x))^2}{(1 - \Phi(z_0))} + \exp(2\mu_y + 2\sigma_y^2) (1 - \Phi(z_0 - 2\rho\sigma_x)) \right. \\ &\quad \left. - \exp(2\mu_y + \sigma_y^2) \frac{(1 - \Phi(z_0 - \rho\sigma_x))^2}{(1 - \Phi(z_0))} - 2 \exp \left( \mu_y + \sigma_y^2 + \rho\sigma_x\sigma_y - \frac{\rho^2\sigma_y^2}{2} \right) \right] \end{aligned}$$

$$-E(Y) (1 - \Phi(z_0 - \sigma_x - \rho\sigma_y)) \frac{(1 - \Phi(z_0 - \sigma_x)) (1 - \Phi(z_0 - \rho\sigma_x))}{(1 - \Phi(z_0))} \Big] \quad (3.20)$$

Similarly, the variance of Total effect for left cut-off point can be calculated through the same procedure

$$Var(X - Y | X < x_0) = Var(X | X < x_0) + Var(Y | X < x_0) - 2Cov(XY | X < x_0) \quad (3.21)$$

To evaluate the variance of the Total effect for the left cut-off point, we need some additional results given below. So the conditional variance of X is

$$\begin{aligned} Var(X | X < x_0) &= E(X^2 | X < x_0) - (E(X | X < x_0))^2 \\ &= \frac{1}{\Phi(-z_0)} \left[ \exp(2(\mu_x + \sigma_x^2))(\Phi(-z_0 - 2\sigma_x)) - \exp(2\mu_x + \sigma_x^2) \frac{(\Phi(-z_0 - \sigma_x))^2}{\Phi(-z_0)} \right] \end{aligned} \quad (3.22)$$

Moreover, the conditional variance of Y for the left cut-off point is

$$\begin{aligned} Var(Y | X < x_0) &= E(Y^2 | X < x_0) - [E(Y | X < x_0)]^2 \\ &= \frac{1}{\Phi(-z_0)} \left[ \exp(2(\mu_y + \sigma_y^2))(\Phi(-z_0 - 2\rho\sigma_x)) - \exp(2\mu_y + \sigma_y^2) \frac{(\Phi(-z_0 - \rho\sigma_x))^2}{\Phi(-z_0)} \right] \end{aligned} \quad (3.23)$$

and the  $Cov(X, Y | X < x_0)$  for left cut-off point is

$$\begin{aligned} Cov(X, Y | X < x_0) &= E(XY | X < x_0) - E(X | X < x_0)E(Y | X < x_0) \\ &= \exp\left(\mu_y + \sigma_y^2 + \rho\sigma_x\sigma_y - \frac{\rho^2\sigma_y^2}{2}\right) \frac{\Phi(-z_0 - \sigma_x - \rho\sigma_y)}{\Phi(-z_0)} \\ &\quad - E(y) \frac{\Phi(-z_0 - \sigma_x)\Phi(-z_0 - \rho\sigma_x)}{[\Phi(-z_0)]^2} \end{aligned} \quad (3.24)$$

By putting the equation (3.16), (3.17) and (3.18) in (3.15) we get the variance of Total for left cut-off point

$$\begin{aligned} Var(X - Y | X < x_0) &= \frac{1}{\Phi(-z_0)} \left[ \exp(2(\mu_x + \sigma_x^2)) \Phi(-z_0 - 2\sigma_x) - \exp(2\mu_x - \sigma_x^2) \right. \\ &\quad \left. \frac{(\Phi(-z_0 - \sigma_x))^2}{\Phi(-z_0)} + \exp(2\mu_y + 2\sigma_y^2) \Phi(-z_0 - 2\rho\sigma_x) \right] \end{aligned}$$

$$\begin{aligned}
 & - \exp(2\mu_y + \sigma_y^2) \frac{(\Phi(-z_0 - \rho\sigma_x))^2}{\Phi(-z_0)} - 2 \exp\left(\mu_y + \sigma_y^2 + \rho\sigma_x\sigma_y - \frac{\rho^2\sigma_y^2}{2}\right) \\
 & - E(Y)\Phi(-z_0 - \sigma_x - \rho\sigma_y) \frac{\Phi(-z_0 - \sigma_x) (\Phi(-z_0 - \rho\sigma_x))}{\Phi(-z_0)} \Big] \quad (3.25)
 \end{aligned}$$

### 3.3.2 Variance of RTM

The variance of RTM can be obtained by assuming the pre-post variables as stationary, i.e.,  $\mu_1 = \mu_2$ ,  $\sigma_1 = \sigma_2$  and substituting these values in the equations (3.17) and (3.22). The following expressions are deduced.

The variance of RTM for right cut-off point is

$$\begin{aligned}
 Var(X - Y | X > x_0) &= \frac{1}{1 - \Phi(z_0)} \left[ \exp(2(\mu_x + \sigma_x^2)) (1 - \Phi(z_0 - 2\sigma_x)) - \exp(2\mu_x - \sigma_x^2) \right. \\
 & \frac{(1 - \Phi(z_0 - \sigma_x))^2}{(1 - \Phi(z_0))} + \exp(2\mu_y + 2\sigma_x^2) (1 - \Phi(z_0 - 2\rho\sigma_x)) \\
 & - \exp(2\mu_x + \sigma_x^2) \frac{(1 - \Phi(z_0 - \rho\sigma_x))^2}{(1 - \Phi(z_0))} - 2 \exp\left(\mu_x + \sigma_x^2 + \rho\sigma_x^2 - \frac{\rho^2\sigma_x^2}{2}\right) \\
 & \left. - E(X) (1 - \Phi(z_0 - \sigma_x - \rho\sigma_x)) \frac{(1 - \Phi(z_0 - \sigma_x)) (1 - \Phi(z_0 - \rho\sigma_x))}{(1 - \Phi(z_0))} \right] \quad (3.26)
 \end{aligned}$$

Similarly, the variance of RTM effect for left cut-off point is

$$\begin{aligned}
 Var(X - Y | X < x_0) &= \frac{1}{\Phi(-z_0)} \left[ \exp(2(\mu_x + \sigma_x^2)) \Phi(-z_0 - 2\sigma_x) - \exp(2\mu_x - \sigma_x^2) \right. \\
 & \frac{(\Phi(-z_0 - \sigma_x))^2}{\Phi(-z_0)} + \exp(2\mu_x + 2\sigma_x^2) \Phi(-z_0 - 2\rho\sigma_x) \\
 & - \exp(2\mu_x + \sigma_x^2) \frac{(\Phi(-z_0 - \rho\sigma_x))^2}{\Phi(-z_0)} - 2 \exp\left(\mu_x + \sigma_x^2 + \rho\sigma_x^2 - \frac{\rho^2\sigma_x^2}{2}\right) \\
 & \left. - E(X) \Phi(-z_0 - \sigma_x - \rho\sigma_x) \frac{\Phi(-z_0 - \sigma_x) (\Phi(-z_0 - \rho\sigma_x))}{\Phi(-z_0)} \right] \quad (3.27)
 \end{aligned}$$

## 3.4 RTM as a function of cut-off point $z_0$

To see the effect of cut-off point on the RTM effect we plot RTM as a function of cut-off point  $x_0$ . By using the RTM expressions in equation (3.12) and (3.13), the graph for various cut-off values is shown in Figure 3.1. The parameters  $\mu_x = 2$ ,  $\mu_y = 2$ ,  $\sigma_x = 1.5$ ,  $\sigma_y = 1.5$  and  $\rho_{xy} = 0.6$  are used for explanation purposes. The graph shows that the RTM effect is at the maximum for the extreme cut-off point at both ends. When the cut-off

value  $x_0$  increases, the probability  $P(X > x_0)$  falls, thereby the related RTM increases. In the case of left cut-off points, as the cut-off value  $X_0$  increases, the probability  $P(X < x_0)$  also increases, and corresponding RTM decreases.

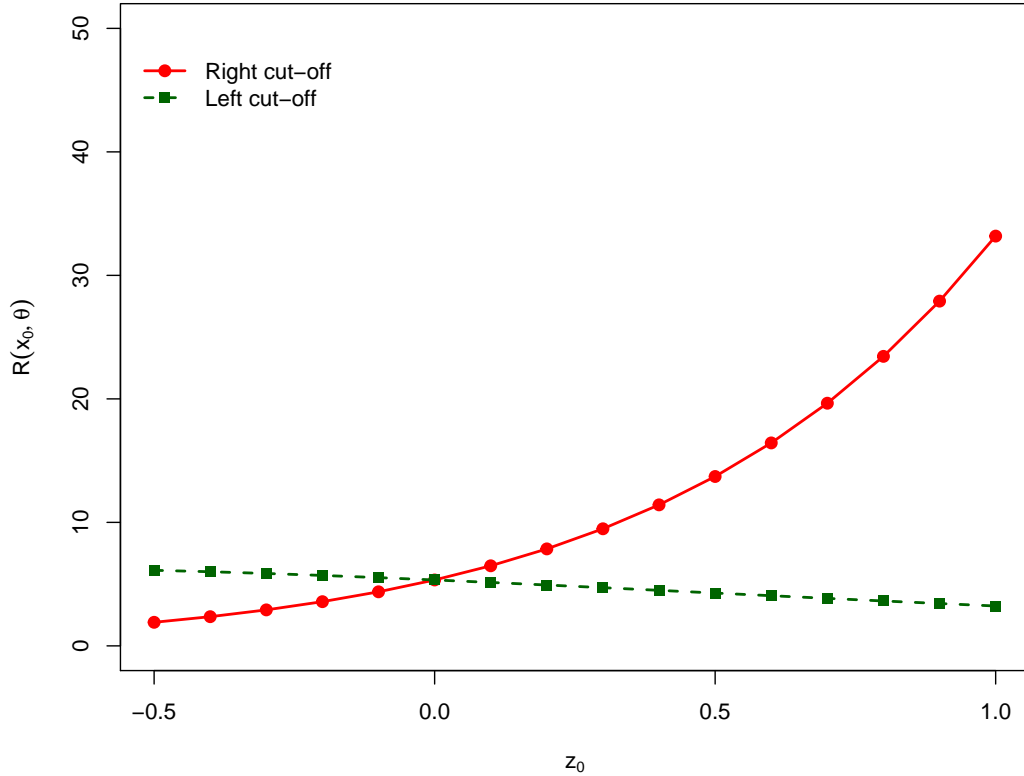


Figure 3.1: Graph showing the RTM effect generated on the basis of the derived formula for points more than or less than a cut-off point, when the underlying distribution is bivariate lognormal with parameter  $\mu_x = 2$ ,  $\mu_y = 2$ ,  $\sigma_x = 1.5$ ,  $\sigma_y = 1.5$  and  $\rho_{xy} = 0.6$

### 3.5 RTM is a function of correlation $\rho_0$

To see the effect of correlation coefficient on the RTM effect we plot RTM as a function of correlation coefficient  $\rho_0$ . By using the RTM expressions in equation (3.12) and (3.13), the graph for various values of correlation coefficient is shown in Figure 3.2. We have fixed the cut-off point  $x_0 = 12$  and the parameters  $\mu_x = 2$ ,  $\mu_y = 2$ ,  $\sigma_x = 1.5$  and  $\sigma_y = 1.5$  are used for explanation purposes. The graph shows that as the correlation  $\rho$  value between variable  $X$  and  $Y$  decreases, the RTM effect rises. When the correlation  $\rho$  between the variables increases, the RTM effect approaches zero. Also, in the case of perfect correlation, there is no RTM effect. Hence, in data analysis, the pre-post variables are highly correlated then the expected RTM effect will be minimum, whereas it would be maximum when



they are independent.

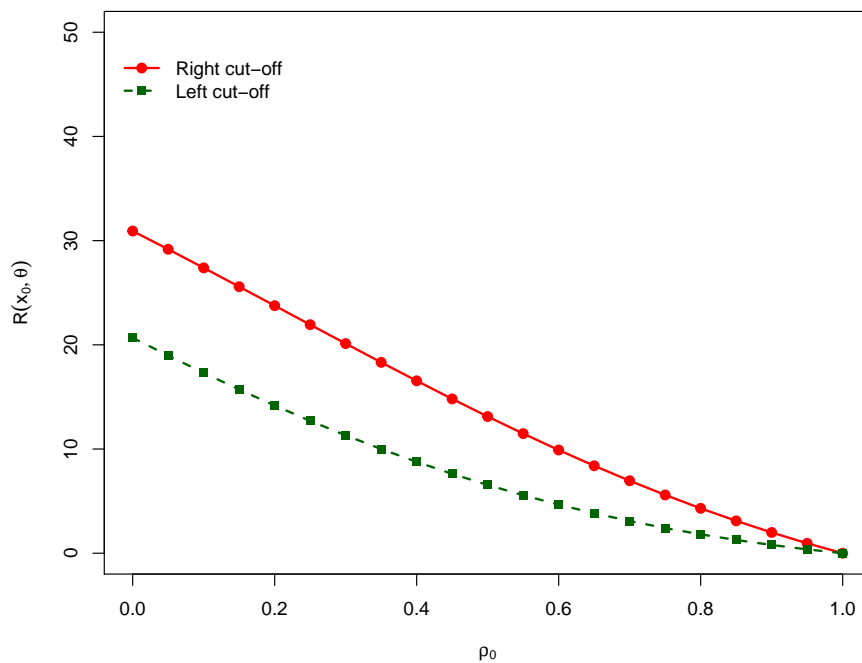


Figure 3.2: The graph shows RTM effect for distinct points of  $\rho$  and predetermined cut-off point  $x_0 = 12$ , when the underlying distribution is bivariate lognormal with parameter  $\mu_x = 2$ ,  $\mu_y = 2$ ,  $\sigma_x = 1.5$  and  $\sigma_y = 1.5$

### 3.6 Comparison of RTM under Lognormal and Normal distributions

According to the general rule of thumb, when  $\mu > 6\sigma$  the lognormal distribution could be approximated by the normal distribution. However, in the case of RTM, this approximation does not work for approximating the RTM effect. Figure 3.3 shows that when the rule of thumb holds, the RTM under the assumption of the bivariate lognormal distribution increases exponentially, while that based on the normality assumption stays flat. From the graph, we can infer that the approximation should not be used while dealing with RTM when the data follows the bivariate lognormal distribution.

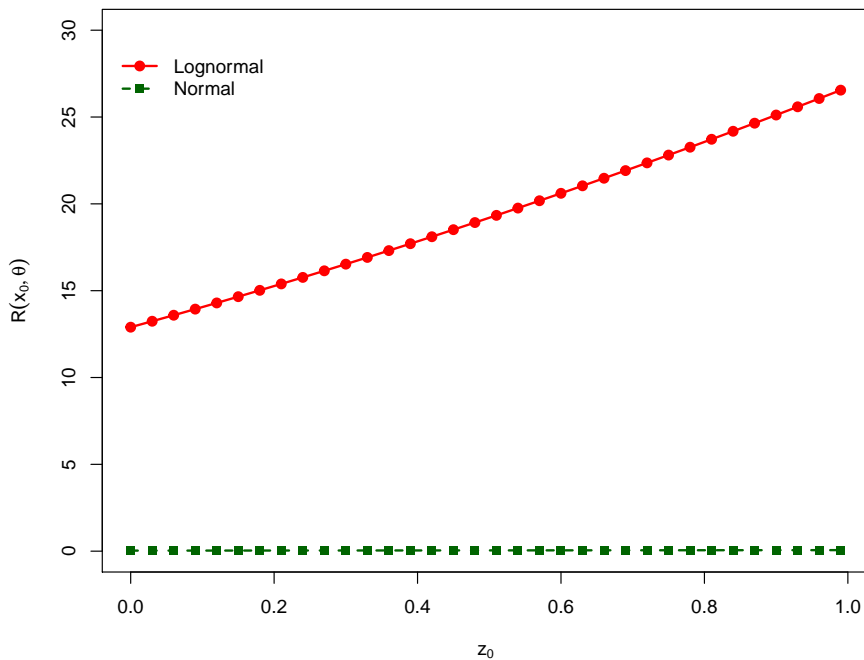


Figure 3.3: The graph shows the comparison of RTM under the lognormal and normal distributions for different cut-off points. The bivariate lognormal/normal parameters are  $\mu_x = 6$ ,  $\mu_y = 6$ ,  $\sigma_x = 0.1$ ,  $\sigma_y = 0.1$  and  $\rho = 0.6$

# Chapter 4

## Estimation of Total and RTM Effects

This chapter discusses the estimation of the total, RTM, and intervention effects. This objective can be achieved by estimating the parameters of the truncated bivariate lognormal distribution with the probability density function given by

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \frac{x^{-1}y^{-1}}{1-\Phi(x_0)} \exp \left[ \frac{-1}{2(1-\rho^2)} \left( \left( \frac{\log(x) - \mu_x}{\sigma_x} \right)^2 + \left( \frac{\log(y) - \mu_y}{\sigma_y} \right)^2 - 2\rho \left( \frac{\log(x) - \mu_x}{\sigma_x} \right) \left( \frac{\log(y) - \mu_y}{\sigma_y} \right) \right) \right] \quad (4.1)$$

To find the maximum likelihood estimates of the bivariate lognormal distribution, we use the method developed by (Cohen Jr, 1955). For this purpose, the probability density function can be written as the product of the marginal frequency function of  $X$  and the conditional frequency function of  $Y$ , without loss of generality. Following Cohen (1955), equation (4.1) becomes

$$f(x, y) = \frac{\left(1 - \Phi\left(\frac{\log(x_0) - \mu_x}{\sigma_x}\right)\right)^{-1}}{2\pi\sigma_x\sqrt{\sigma}} \exp\left(\frac{-1}{2}\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)^2\right) \exp\left(\frac{-1}{2\sigma}\left(\log(y) - \alpha - \beta\left(\log(x) - \overline{\log(x)}\right)\right)^2\right). \quad (4.2)$$

where  $\beta = \frac{\rho\sigma_y}{\sigma_x}$ ,  $\alpha = \mu_y - \beta\left(\mu_x - \overline{\log(x)}\right)$  and  $\sigma^2 = \sigma_y(1 - \rho^2)$

Let  $(x_{11}, y_{21}), (x_{12}, y_{22}), \dots, (x_{1n}, y_{2n})$  be the pairs of independently distributed observation of size  $n$  from the bivariate lognormal distribution. The expressions of likelihood and log

likelihood functions are given below

$$L(x, y; \theta) = \frac{\left(1 - \Phi\left(\frac{\log_0 - \mu_x}{\sigma_x}\right)\right)^{-n}}{2\pi\sigma_x\sqrt{\sigma^2}} \prod_{i=1}^n x_i^{-1} y_i^{-1} \exp\left[-\frac{1}{2} \sum_{i=1}^n \left(\frac{\log(x) - \mu_x}{\sigma_x}\right)^2\right] \exp\left[-\frac{1}{2\sigma^2} \left(\log(y) - \alpha - \beta(\log(x) - \overline{\log(x)})\right)\right]$$

and

$$\begin{aligned} \ell(x, y; \theta) &= -n \log(2\pi) - n \log(\sigma^2) - \frac{n}{2} \log(\sigma^2) - n \log\left[1 - \Phi\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)\right] - \sum_{i=1}^n x_i \\ &\quad - \sum_{i=1}^n y_i - \frac{1}{2} \sum_{i=1}^n \left(\frac{\log(x) - \mu_x}{\sigma_x}\right)^2 - \frac{1}{2\sigma^2} \sum_{i=1}^n \left(\log(y) - \alpha - \beta(\log(x) - \overline{\log(x)})\right) \end{aligned} \quad (4.3)$$

Differentiating the equation (4.3) with respect to  $\mu_x$  and then equating to zero, we get

$$\begin{aligned} \frac{-n\phi\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)}{\sigma_x \left(1 - \Phi\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)\right)} + \frac{\sum_{i=1}^n (\log(x) - \mu_x)}{\sigma_x^2} &= 0 \\ \frac{-n\phi(z_0')}{\sigma_x (1 - \Phi(z_0'))} + \frac{\sum_{i=1}^n (\log(x) - \mu_x)}{\sigma_x^2} &= 0 \end{aligned} \quad (4.4)$$

where  $z_0' = \frac{\log(x) - \mu_x}{\sigma_x}$ . After simplifying equation (4.4), we get

$$\hat{\mu}_x = \overline{\log(x)} - \sigma_x \frac{\phi(z_0')}{1 - \Phi(z_0')}. \quad (4.5)$$

Differentiating the log likelihood function with respect to  $\sigma_x$  and then equating to zero, we get

$$\begin{aligned} \frac{-n}{\sigma_x} - \frac{n\phi\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)}{1 - \Phi\left(\frac{\log(x) - \mu_x}{\sigma_x}\right)} \left(\frac{\log(x) - \mu_x}{\sigma_x^2}\right) + \frac{2}{\sigma_x^3} \sum_{i=1}^n \left(\overline{\log(x)} - \mu_x\right)^2 &= 0 \\ -n \left[ \frac{1}{\sigma_x} + \frac{nz_0'\phi(z_0')}{\sigma_x (1 - \Phi(z_0'))} - \frac{1}{n\sigma_x^3} \sum_{i=1}^n (\log(x) - \mu_x)^2 \right] &= 0 \end{aligned} \quad (4.6)$$

Rearranging terms, and using  $c(z_0') = \frac{\phi(z_0')}{1 - \Phi(z_0')}$ , it can be shown that

$$\hat{\sigma}_x^2 = \frac{S_x^2}{1 + z_0'c(z_0')}, \quad (4.7)$$

where  $S_x^2 = \frac{1}{n} \sum_{i=1}^n (\log(x_i) - \hat{\mu}_x)^2$ .

Now, differentiating the  $\ell(x, y; \theta)$  with respect to  $\alpha$  and then setting the equation to zero, we get

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n \left( \log(y) - \alpha - \beta(\log(x) - \overline{\log(x)}) \right) &= 0 \\ \sum_{i=1}^n \log(y) - n\alpha - \beta \sum_{i=1}^n \left( \log(x) - \overline{\log(x)} \right) &= 0 \end{aligned} \quad (4.8)$$

By simplification of equation (4.8), we have

$$\hat{\alpha} = \overline{\log(y)}. \quad (4.9)$$

Similarly, differentiating  $\ell(x, y; \theta)$  with respect to  $\beta$  and then equating to zero, we get

$$\begin{aligned} \frac{1}{\sigma^2} \sum_{i=1}^n \left( \log(y) - \alpha - \beta(\log(x) - \overline{\log(x)}) \right) (\log(x) - \overline{\log(x)}) &= 0 \\ \sum_{i=1}^n \log(y)(\log(x) - \overline{\log(x)}) - \alpha \sum_{i=1}^n (\log(x) - \overline{\log(x)}) &= \beta \sum_{i=1}^n (\log(x) - \overline{\log(x)})^2 \end{aligned}$$

It can be simplified to

$$\hat{\beta} = \frac{\sum_{i=1}^n (\log(x) - \overline{\log(x)})(\log(y) - \overline{\log(y)})}{\sum_{i=1}^n (\log(x) - \overline{\log(x)})^2}$$

and

$$\hat{\beta} = \frac{\bar{r} S_y}{S_x} \quad (4.10)$$

where  $\bar{r} = \frac{\sum_{i=1}^n (\log x_i - \overline{\log x_1})(\log x_2 - \overline{\log x_2})}{n S_x S_y}$  and  $S_y^2 = \frac{1}{n} \sum_{i=1}^n (\log x_2 - \hat{\mu}_y)^2$

Differentiating the log likelihood function with respect to  $\sigma$  and then equating to zero, we get

$$\frac{-n}{\sigma} + \frac{\sum_{i=1}^n \left( \log(y) - \alpha - \beta(\log(x) - \overline{\log(x)}) \right)^2}{\sigma^3} = 0 \quad (4.11)$$

The equation (4.11) can be further simplified to

$$\begin{aligned}\sigma &= \frac{1}{n} \sum_{i=1}^n \left( \log(y) - \overline{\log(y)} - \beta(\log(x) - \overline{\log(x)}) \right)^2 \\ &= \frac{1}{n} \left( \sum_{i=1}^n (\log(y) - \overline{\log(y)})^2 + \beta^2 \sum_{i=1}^n (\log(x) - \overline{\log(x)})^2 \right. \\ &\quad \left. - 2\beta \sum_{i=1}^n (\log(x) - \overline{\log(x)})(\log(y) - \overline{\log(y)}) \right)\end{aligned}$$

and

$$\hat{\sigma} = S_y \sqrt{1 - \hat{r}^2} \quad (4.12)$$

Now the MLE's of truncated bivariate lognormal distribution is respectively given

$$\hat{\mu}_x = \overline{\log(x)} - \hat{\sigma}_x \frac{\phi(z_0)}{1 - \Phi(z_0)} \quad (4.13)$$

$$\hat{\sigma}_x^2 = \frac{S_x^2}{[1 + z_0 c(z_0)]} \quad (4.14)$$

$$\hat{\mu}_y = \hat{\alpha} + \hat{\beta} \left( \overline{\log(x)} - \hat{\mu}_x \right) \quad (4.15)$$

$$\hat{\sigma}_y = \sqrt{\hat{\sigma}^2 + \hat{\sigma}_x^2 \hat{\beta}^2} \quad (4.16)$$

$$\hat{\rho} = \frac{\hat{\sigma}_x^2 \hat{\beta}^2}{\sqrt{\hat{\sigma}^2 + \hat{\sigma}_x^2 \hat{\beta}^2}} \quad (4.17)$$

The MLE of  $T_r(x_0, \theta)$  is

$$\hat{T}_r(x_0, x) = E(X) \left[ \frac{1 - \Phi(z_0 - \hat{\sigma}_x)}{1 - \Phi(z_0)} \right] - E(Y) \left[ \frac{1 - \Phi(z_0 - \hat{\rho} \hat{\sigma}_x)}{1 - \Phi(z_0)} \right] \quad (4.18)$$

By putting  $E(Y) = E(X)$ , we get the  $\hat{R}_r(x_0, x)$  from equation (4.18)

$$\hat{R}_r(x_0, x) = E(X) \left[ \frac{1 - \Phi(z_0 - \hat{\sigma}_x)}{1 - \Phi(z_0)} - \frac{1 - \Phi(z_0 - \hat{\rho} \hat{\sigma}_x)}{1 - \Phi(z_0)} \right] \quad (4.19)$$

Where the expectation of X and Y are the unconditional means of the sample, the expressions are  $E(X) = \exp\left(\hat{\mu}_x + \frac{\hat{\sigma}_x^2}{2}\right)$  and  $E(Y) = \exp\left(\hat{\mu}_y + \frac{\hat{\sigma}_y^2}{2}\right)$ .

Similarly, the MLE of  $T_l(x_0, \theta)$

$$\widehat{T}_l(x_0, x) = E(X) \left[ \frac{\Phi(-z_0 - \hat{\sigma}_x)}{\Phi(-z_0)} \right] - E(Y) \left[ \frac{\Phi(-z_0 - \hat{\rho}\hat{\sigma}_x)}{\Phi(-z_0)} \right] \quad (4.20)$$

To obtain the MLE of  $\widehat{R}_l(x_0, x)$  by putting  $E(Y) = E(X)$  in equation (4.20)

$$\widehat{R}_l(x_0, x) = E(X) \left[ \frac{\Phi(-z_0 - \hat{\sigma}_x)}{\Phi(-z_0)} - \frac{\Phi(-z_0 - \hat{\rho}\hat{\sigma}_x)}{\Phi(-z_0)} \right] \quad (4.21)$$

## 4.1 Simulation Analysis of RTM Effect

A simulation study was carried out to estimate the RTM effect and empirically evaluate its characteristics. The method of [Aitchison and Egozcue \(2005\)](#) was used to generate sample observations from bivariate lognormal distribution for different parameters using R package. The corresponding probability  $P(X > x_0)$  is relatively small if a cut-off point is chosen far in the tail on either side. Therefore, the number of observations above/below a cut-off would be very small in the generated sample. To have enough observation above/below a certain threshold say,  $x_0 = 14$ , samples sizes of  $n1 = 50000$  were generated from the bivariate lognormal distribution for different permutations of the parameters. The first  $n$  observations greater than  $x_0$  along with the associated  $Y$  were considered as the bivariate random samples from a truncated bivariate lognormal distribution. The sampling method was performed  $i = 1000$  times, and the RTM and intervention effects were estimated by using the maximum likelihood for each sample.

## 4.2 Comparison of methods

In literature [Beath and Dobson \(1991\)](#) estimates the RTM effect for empirical nonnormal distribution by adopting the Edgeworth series and Saddlepoint approximations. In this section, we graphically compare the estimated RTM and intervention effects by the proposed, the Edgeworth series, and Saddlepoint approximation methods for a nonnormal population.

### 4.2.1 Comparison the RTM effect

This subsection compares the estimated RTM by the proposed method with the estimated RTM by the Edgeworth series and Saddlepoint approximation for different parameters.

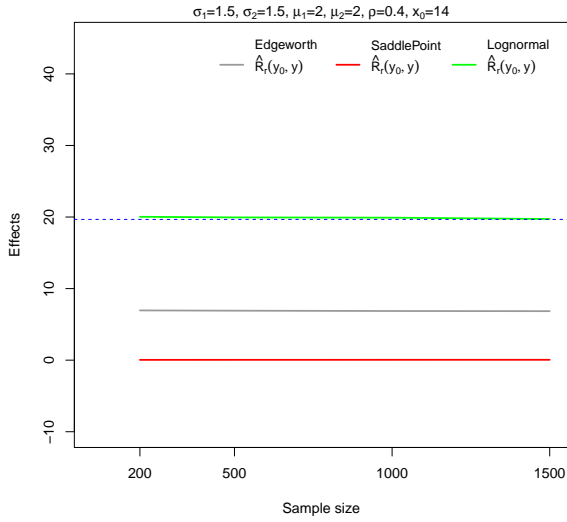


Figure 4.1: RTM at correlation coefficient  $\rho=0.4$

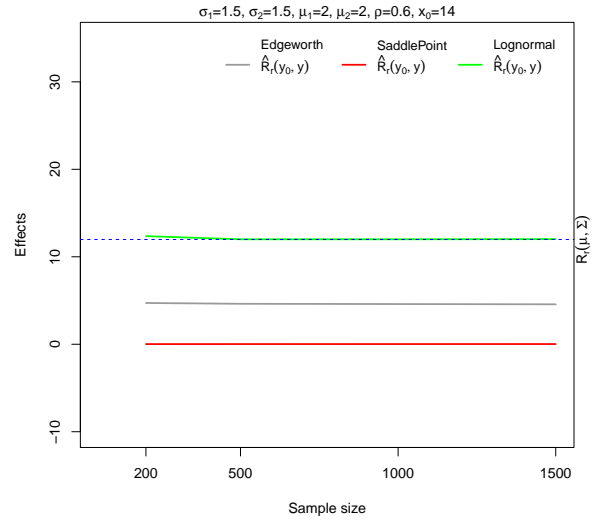


Figure 4.2: RTM at correlation coefficient  $\rho=0.6$

In Figure 4.1, the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ . The resulting true RTM was 20. The estimated RTM based on lognormal distribution is close to its true value, suggesting unbiasedness of the estimator, whereas the estimated RTM based on Edgeworth series and Saddle-point approximation underestimated the RTM effect by more than half in error of its true value. And in Figure 4.2, the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = 1.5$ ,  $\sigma_2^2 = 1.5$  and  $\rho = 0.6$ , and the value of the true RTM effect is near 10. As expected, the RTM decreased with increasing value of the correlation coefficient. Here, also the Edgeworth and Saddlepoint methods underestimated the true RTM, in contrast to the method based on lognormal distribution which unbiasedly estimated the true RTM.

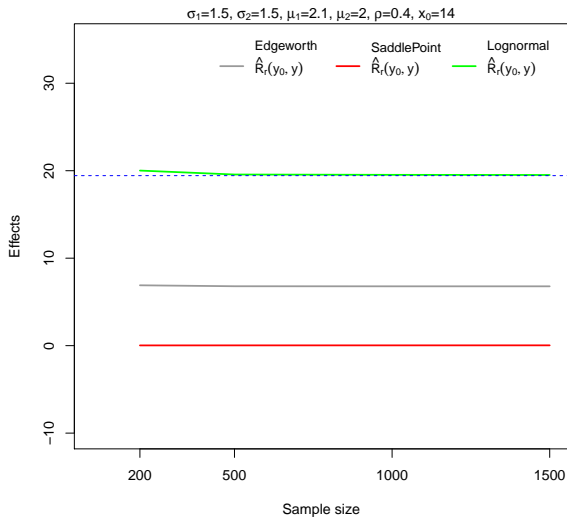


Figure 4.3: RTM at correlation coefficient  $\rho=0.4$  and  $\mu_1=2.1$

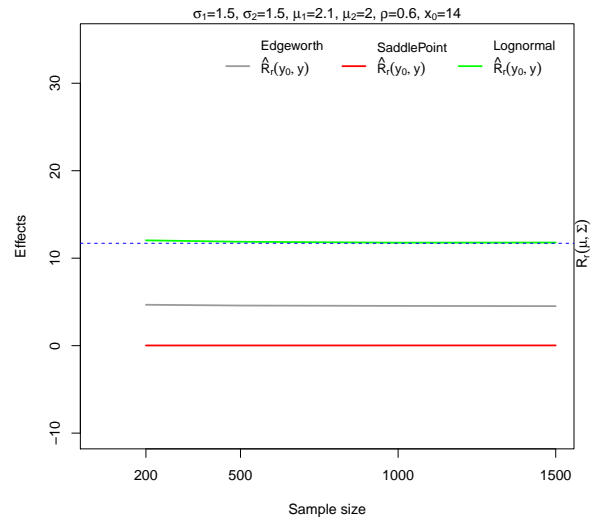


Figure 4.4: RTM at correlation coefficient  $\rho=0.6$  and  $\mu_1=2.1$



In Figure 4.3, the parameters were fixed at  $\mu_1 = 2.1$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ . As we changed the parameter  $\mu_1 = 2$  to  $\mu_1 = 2.1$  the true RTM slightly decreased. The estimated RTM based on lognormal distribution is close to its true value, indicating that the estimator is unbiased. However, the estimated RTM based on the Edgeworth series and Saddle-point approximation underestimated the RTM effect by more than half in the error of its true value, indicating that the estimator is not reliable. And in Figure 4.4, the correlation was increased from  $\rho = 0.4$  to  $\rho = 0.6$ . The RTM effect declined with the value of the new correlation coefficient. In this case, the Edgeworth and Saddlepoint techniques again underestimated the true RTM, but the method based on lognormal accurately estimated the true RTM.

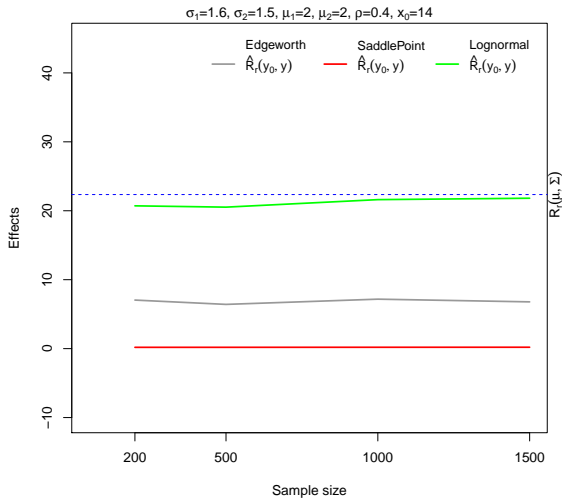


Figure 4.5: RTM at correlation coefficient  $\rho=0.4$  and  $\sigma_1 = 1.6$

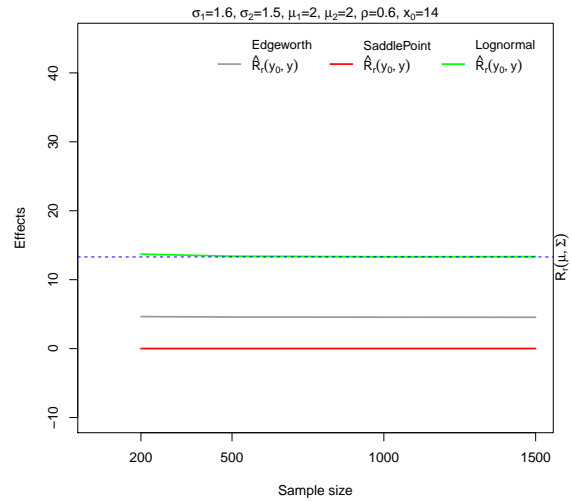


Figure 4.6: RTM at correlation coefficient  $\rho=0.6$  and  $\sigma_1 = 1.6$

In Figure 4.5, we allowed the variables to have different variances and the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.6$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ . As the parameter  $\sigma_1$  increased the true RTM rose marginally. The RTM based on the lognormal is close to its true value. According to the estimated RTM based on the Edgeworth series and the Saddle-point approximation, the RTM impact was underestimated by more than half in the inaccuracy of its true value, demonstrating the unreliability of the estimator. In Figure 4.6, the only parameter changed was  $\rho = 0.6$ . Similar pattern appeared again.

In Figure 4.7, the case  $\sigma_1 < \sigma_2$  was considered to see its effect on the true RTM and estimation of the parameters. The parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.6$  and  $\rho = 0.4$ . As we increased the parameter  $\sigma_2$ , the resulting true RTM was 20. The estimated RTM based on lognormal distribution is close to its true value, suggesting

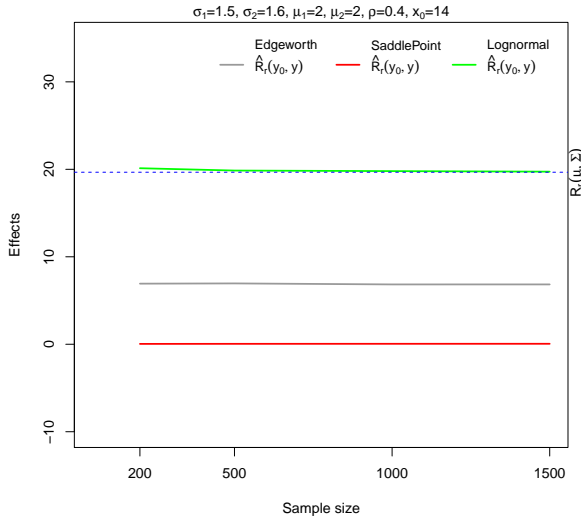


Figure 4.7: RTM at correlation coefficient  $\rho=0.4$  and  $\sigma_2 = 1.6$

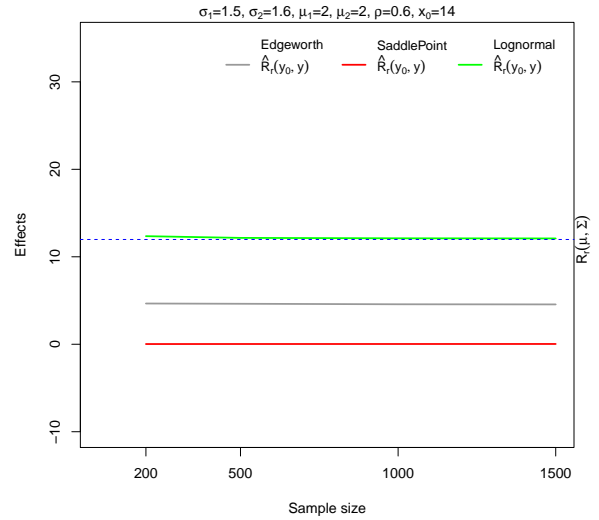


Figure 4.8: RTM at correlation coefficient  $\rho=0.6$  and  $\sigma_2 = 1.6$

unbiasedness of the estimator, while the estimated RTM based on Edgeworth series and Saddle-point approximation underestimated the RTM effect with an error of more than 10 in absolute value with the later even worse. In Figure 4.8, the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.6$  and  $\rho = 0.6$ , and the value of the true RTM effect is near 10. As expected, the RTM decreased with increasing value of the correlation coefficient. The pattern of estimation of the two methods remained the same.

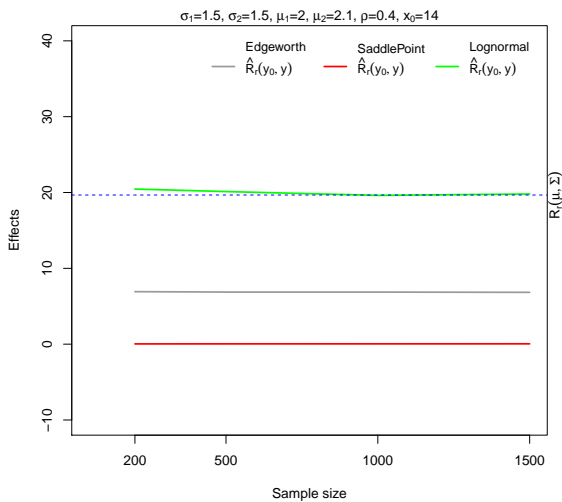


Figure 4.9: RTM at correlation coefficient  $\rho=0.4$  and  $\mu_2=2.1$

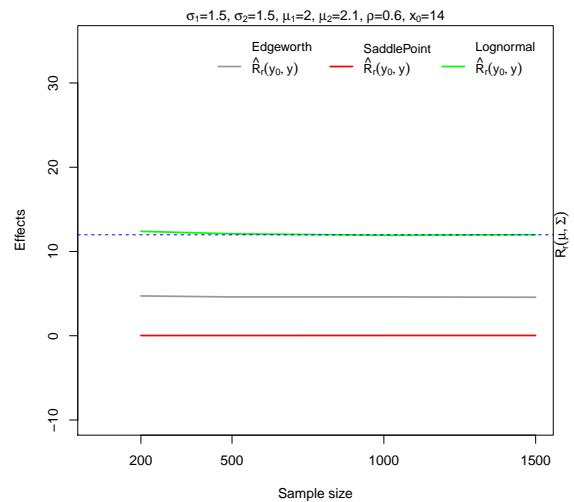


Figure 4.10: RTM at correlation coefficient  $\rho=0.6$  and  $\mu_2=2.1$

The last case  $\mu_1 < \mu_2$  was also considered. In Figure 4.9, the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2.1$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ , as we increased the parameter  $\mu_2$  the true RTM is 20. The RTM based on the lognormal was close to its true value.

The estimated RTM based on the Edgeworth series and the Saddle-point approximation underestimated the true RTM. Similarly, in Figure 4.10, as usual the parameters  $\rho = 0.6$  was increased, and the value of the true RTM effect was near 10. As expected, the RTM decreased with increasing value of the correlation coefficient. Here, also the Edgeworth and Saddlepoint methods underestimated the true RTM, in contrast to the method based on lognormal distribution.

### 4.2.2 Comparison of the estimating methods for treatment effect

This section discusses the graphical comparison of the estimated treatment effect by the proposed method based on the lognormal distribution and the Edgeworth series and Saddlepoint approximation for a different choices of the population parameters.

Firstly, we consider the case of the stationary variables where the distribution of the pre-post variables are identical. In this case the intervention/treatment effect is zero, i.e.,  $\delta\mu = 0$ . In Figure 4.11, the choices of parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ . The estimated treatment effect based on the lognormal distribution is zero which coincides with the true treatment effect indicating the unbiasedness of estimator. But, the lines of the estimated treatment effect based on the Edgeworth series and Saddlepoint approximation is close to 10 which overestimates the true treatment effect. This could lead to inaccurate conclusion if not taken care off. To see the effect of correlation on the treatment effect, the correlation was increased to  $\rho = 0.6$  In Figure 4.12 with other parameters unchanged. Here, also the Edgeworth and Saddlepoint methods overestimated the true treatment, whereas the method based on lognormal unbiasedly estimated the true treatment.

Secondly, the case of non-stationary pre-post variables with  $\mu_1 > \mu_2$  was considered. This allowed the treatment effect to be greater than zero, i.e.,  $\delta\mu > 0$ . In Figure 4.13, the parameters were fixed at  $\mu_1 = 2.1$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$ . With these choices of the parameters, the true treatment is around 8, and the green line of the estimated treatment effect for different sample sizes by the proposed method coincides with the true line. On the other hand, the estimated treatment by Edgeworth series and Saddlepoint approximation is well above the true treatment. Figure 4.14 depicted similar

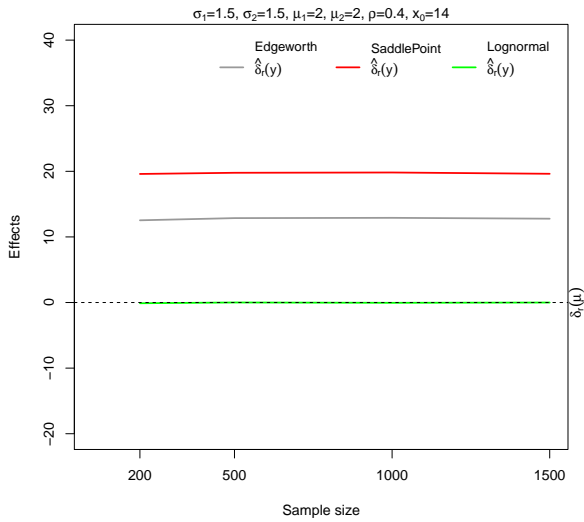


Figure 4.11: Treatment at correlation coefficient  $\rho=0.6$

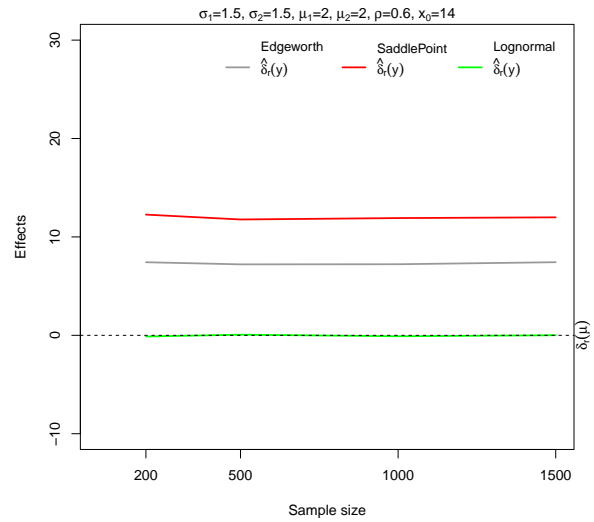


Figure 4.12: Treatment at correlation coefficient  $\rho=0.4$

behaviour for  $\rho = 0.6$ . The Edgeworth and Saddlepoint methods overestimated the true treatment by more than double of its true value.

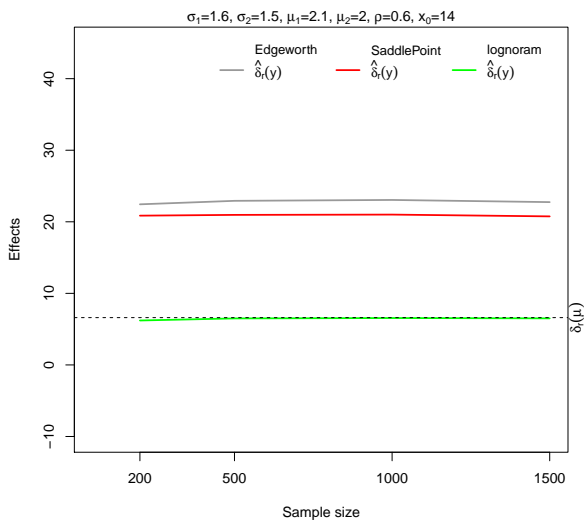


Figure 4.13: Treatment at correlation coefficient  $\rho=0.6$  and  $\mu_1=2.1$

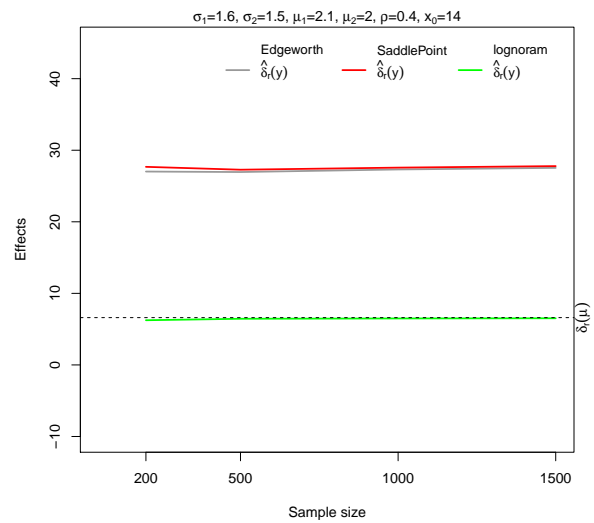


Figure 4.14: Treatment at correlation coefficient  $\rho=0.4$  and  $\mu_1=2.1$

Thirdly, the case of  $\sigma_1 > \sigma_2$  was considered. As the mean of the lognormal distribution is also a function of the dispersion parameter  $\sigma$ , so the treatment effect is non-zero here as well, particularly in this case,  $\delta(\mu) > 0$ . For this purpose, the choices of the parameters were  $\mu_1 = 2.1$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.6$ ,  $\sigma_2 = 1.5$  and  $\rho = 0.4$  and the result is depicted in Figure 4.15. By increasing the parameter from  $\sigma_1 = 1.5$  to  $\sigma_1 = 1.6$ , the lines of the estimated treatment effect by the proposed method and the true treatment are flat and close to 4, which indicates that the estimator is unbiased. But the estimated treatment effect by the

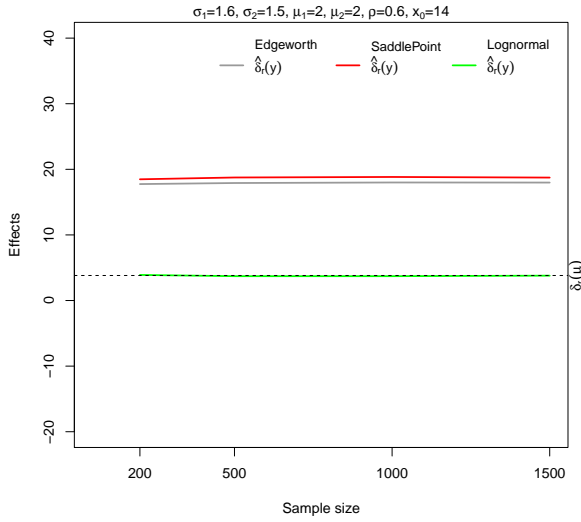


Figure 4.15: Treatment at correlation coefficient  $\rho=0.6$  and  $\sigma_1=1.6$

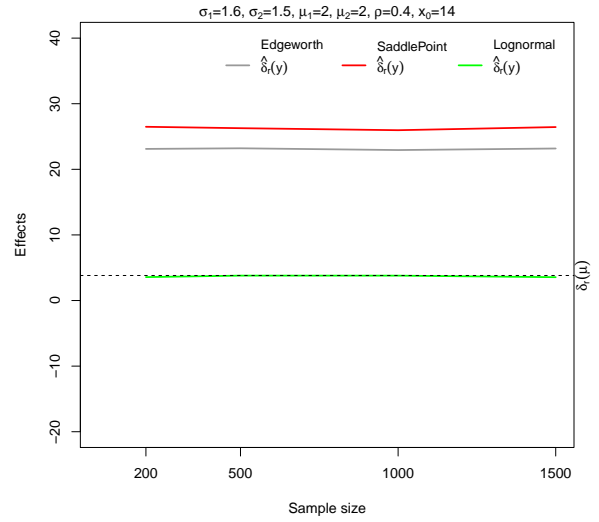


Figure 4.16: Treatment at correlation coefficient  $\rho=0.4$  and  $\sigma_1=1.6$

Edgeworth series and Saddlepoint approximation overestimated the true treatment. In Figure 4.16 the parameters were fixed at  $\mu_1 = 2.1$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.6$ ,  $\sigma_2 = 1.5$ ,  $\rho = 0.6$ , and the true treatment effect remained around the same value. The performance of the Edgeworth and Saddlepoint methods further deteriorated, whereas the proposed method unbiasedly estimated the true treatment effect.

Lastly, we considered the case  $\sigma_1 < \sigma_2$ . Here, the treatment effect is negative, i.e.,  $\delta(\mu) < 0$ . In Figure 4.17, the parameters were fixed at  $\mu_1 = 2$ ,  $\mu_2 = 2$ ,  $\sigma_1 = 1.5$ ,  $\sigma_2 = 1.6$  and  $\rho = 0.4$ . By increasing the parameter from  $\sigma_2 = 1.5$  to  $\sigma_2 = 1.6$ , the true treatment is around  $\delta(\mu) = -3$ . The estimated treatment effect by the proposed method is almost equal the true value for different sample sizes. However, the estimated treatment by the Edgeworth series and Saddlepoint approximation overestimated the true treatment and the later estimated it by more the three fold of its true value. In another permutation of the parameters, the results of the simulation are given in Figure 4.18. Here only correlation coefficient was changed  $\rho = 0.6$ . The performance of the Edgeworth series and Saddlepoint further deteriorated from estimation point of view.

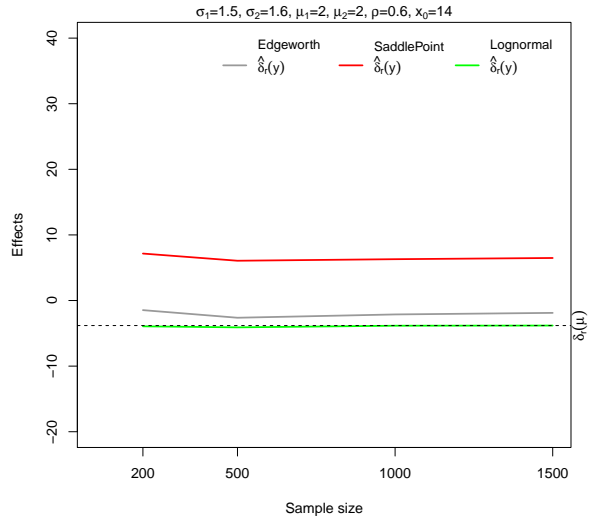


Figure 4.17: Treatment at correlation coefficient  $\rho=0.6$  and  $\sigma_2=1.6$

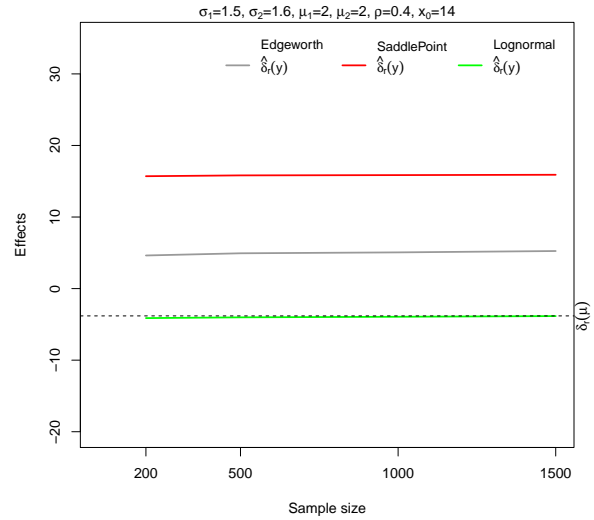


Figure 4.18: Treatment at correlation coefficient  $\rho=0.4$  and  $\sigma_2=1.6$

### 4.3 Data Example: the Cyclosporin Study

The data set encompassing  $n = 56$  cyclosporin assay pairs in this manuscript was obtained from a previously published paper (Gupta et al., 2013). Cyclosporine is a calcineurin inhibitor used as an immunosuppressant medication. Cyclosporine is given with other medications to prevent transplant rejection in people who have received kidney, liver, and heart transplants. It is usually used along with other medications to allow your new organ to function normally. It works by weakening the immune system by stopping white blood cells from attacking a transplanted organ. Cyclosporine is a potent immunomodulatory agent with an increasing number of clinical applications.

The method of maximum likelihood method was used for estimating the parameters of the bivariate lognormal distribution. As there was no information about the baseline criterion, so the data were assumed to be coming from un-truncated bivariate lognormal distribution. The estimated parameter were  $\hat{\mu}_x = 4.88$ ,  $\hat{\mu}_y = 4.96$ ,  $\hat{\sigma}_x = 0.92$ ,  $\hat{\sigma}_y = 0.81$  and  $\hat{\rho}_{xy} = 0.96$ . The estimated treatment effect here is  $\hat{\delta}(\mu) = 3.02$  which is independent of the cut-off point. To see how RTM exaggerate the treatment effect, we assume different cut-off points for the data example under study. From Figure 4.19, it is evident that as the cut-off point increases the RTM effect (the red line) increases which ultimately increases the total effect (the green dotted line). If RTM is not accounted for then the total effect would be mistakenly associated with the estimated treatment effect which is

3.02 and this would bias the conclusion of this study as the total effect is at least 15.

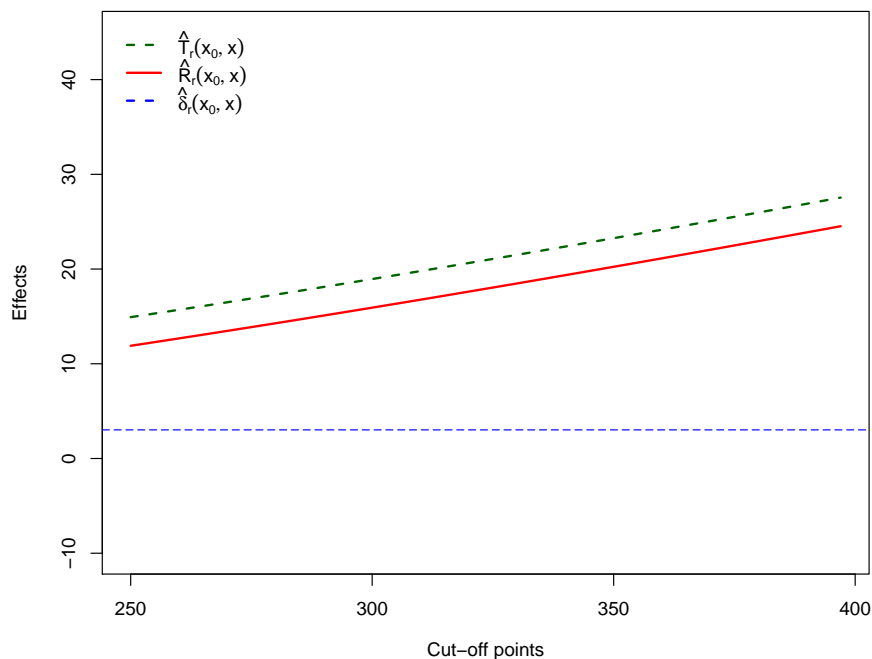


Figure 4.19: Graph shows that the total, RTM, and treatment effects generated on the basis of the derived formula for points more than a cut-off value for  $\hat{\mu}_x = 4.88$ ,  $\hat{\mu}_y = 4.96$ ,  $\hat{\sigma}_x = 0.92$ ,  $\hat{\sigma}_y = 0.81$  and  $\hat{\rho}_{xy} = 0.96$

# Chapter 5

## Discussion

Regression to the mean is a serious problem in data analysis that can lead to incorrect conclusions if overlooked, for this reason, RTM need to be addressed. Due to moral limitations, random allocation procedures or numerous baseline measurements intended to mitigate the RTM effect are not always viable. Thus, quantifying and accounting for the RTM effect is an important objective in an intervention study. RTM expressions for the Normal, Poisson and Binomial distributions are accessible in the literature. RTM expressions are not available when the pre and post-variables are distributed according to the bivariate lognormal distribution.

According to the literature, the RTM effect decreases when the correlations between pre and post variables increases and the pattern depends on the underlying bivariate distribution. Our derivations suggested that a bivariate lognormal distribution behaves similarly to the bivariate normal distribution based on the correlation function. The RTM impact reduces linearly for the normal distribution when the correlation between variables increase in pre/post studies. Likewise, in lognormal distribution, RTM for the left and the right cut-off point also increases as the correlation increase. The RTM effect becomes more severe when the correlation between the two variables becomes weaker. Similarly, as the subjects for an intervention are selected far in the tail of a distribution, the effect of RTM would be more adverse.

RTM is more likely to occur in a pre/post research design when intervention or therapy is administered to subjects who have been selected based on particular thresholds. As the cut-off point moves further into the tail of the baseline distribution, the severity of



RTM grows proportionally. By splitting the total effect, the treatment and RTM can be determined.

A simulation study was conducted to compare estimation of the RTM under the bivariate lognormal distribution to previously published approaches for nonnormal populations, such as the Edgeworth series and Saddlepoint approximation. According to our simulation results, the proposed formulation for the RTM effect under bivariate lognormal distribution is substantially more satisfying than the Edgeworth series and Saddlepoint approximation.

The RTM effect was evaluated for different cut-off points for the data set containing 56 cyclosporin assay pairs. The study's findings are based on 56 blood samples collected from organ transplant patients. Utilizing the maximum likelihood method, we obtain the estimates of parameters. It is inaccurate to examine cyclosporin's actual effectiveness without taking into account the RTM effect.

We conclude from the results that the RTM effect needs to be accounted for using the derived method in this thesis when the pre-post variable follows the bivariate lognormal distributions. As other existing methods would underestimate the RTM effect, thereby overestimating the intervention effect that would lead to erroneous conclusions.

# References

- Aitchison, J. and Egozcue, J. J. (2005). Compositional data analysis: where are we and where should we be heading? *Mathematical Geology*, 37(7):829–850.
- 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.
- Blackwood, L. G. (1992). The lognormal distribution, environmental data, and radiological monitoring. *Environmental Monitoring and Assessment*, 21(3):193–210.
- Burke, R. M., Meyer, A., Kay, C., Allensworth, D., and Gazmararian, J. A. (2014). A holistic school-based intervention for improving health-related knowledge, body composition, and fitness in elementary school students: an evaluation of the healthmpowers program. *International Journal of Behavioral Nutrition and Physical Activity*, 11(1):1–12.
- Cohen Jr, A. C. (1955). Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association*, 50(271):884–893.
- Crow, E. L. and Shimizu, K. (1987). *Lognormal distributions*. Marcel Dekker New York.
- 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.
- Dehghani, H. and Fadaee, M. J. (2020). Probabilistic prediction of earthquake by bivariate distribution. *Asian Journal of Civil Engineering*, 21:977–983.
- Ederer, F. (1972). Serum cholesterol changes: effects of diet and regression toward the mean. *Journal of Chronic Diseases*, 25(5):277–289.

## References

---

- Gale, H. (1967). Some examples of the application of the lognormal distribution in radiation protection. *Annals of Occupational Hygiene*, 10(1):39–45.
- 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.
- Gupta, R. C., Ghitany, M., and Al-Mutairi, D. (2013). Estimation of reliability from a bivariate log-normal data. *Journal of Statistical Computation and Simulation*, 83(6):1068–1081.
- Hannon, B. A., Thomas, D. M., Siu, C., and Allison, D. B. (2018). The claim that effectiveness has been demonstrated in the parenting, eating and activity for child health (peach) childhood obesity intervention is unsubstantiated by the data. *British Journal of Nutrition*, 120(8):958–959.
- Ibrahim, Q. I. (2015). Adjustment for the regression to the mean effects in studies with repeated measures.
- James, K. E. (1973). Regression toward the mean in uncontrolled clinical studies. *Biometrics*, 29(1):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.
- Kenton (2020). Lognormal distribution. <https://www.investopedia.com/terms/l/log-normal-distribution.asp>.
- Khan, M. (2019). *Quantification and Estimation of Regression to The Mean for Bivariate Distributions*. PhD thesis, The University of New South Wales.
- Khan, M. and Olivier, J. (2018). Quantifying the regression to the mean effect in poisson processes. *Statistics in Medicine*, 37(26):3832–3848.

## References

---

- Khan, M. and Olivier, J. (2019). Regression to the mean for the bivariate binomial distribution. *Statistics in Medicine*, 38(13):2391–2412.
- McMAHAN, C. A. (1982). Regression toward the mean in a two-stage selection program. *American Journal of Epidemiology*, 116(2):394–401.
- Moore, C. J., Miller, J., Daniels, L. A., Vidgen, H. A., and Magarey, A. M. (2018). Pre–post evaluation of a weight management service for families with overweight and obese children, translated from the efficacious lifestyle intervention parenting, eating and activity for child health (peach). *British Journal of Nutrition*, 119(12):1434–1445.
- Morrison, J. (1958). The lognormal distribution in quality control. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 7(3):160–172.
- 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.
- 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.
- Skinner, A. C., Heymsfield, S. B., Pietrobelli, A., Faith, M. S., and Allison, D. B. (2015). Ignoring regression to the mean leads to unsupported conclusion about obesity. *International Journal of Behavioral Nutrition and Physical Activity*, 12(1):1–3.
- Yerel, S. and Konuk, A. (2009). Bivariate lognormal distribution model of cutoff grade impurities: A case study of magnesite ore deposit. *Scientific Research and Essays*, 4(12):1500–1504.
- Yu, R. and Chen, L. (2015). The need to control for regression to the mean in social psychology studies. *Frontiers in Psychology*, 5:1574.
- Yudkin, P. (1996). How to deal with regression to the mean in intervention studies. *Lancet*, 347:241–243.
- Yue, S. (2000). The bivariate lognormal distribution to model a multivariate flood episode. *Hydrological Processes*, 14(14):2575–2588.

## References

---

- Yue, S. (2002). The bivariate lognormal distribution for describing joint statistical properties of a multivariate storm event. *Environmetrics: The official journal of the International Environmetrics Society*, 13(8):811–819.