KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY, KUMASI

COMPARISON OF QUANTILE REGRESSION TO LOGNORMAL AND GAMMA REGRESSION USING BIRTH WEIGHT DATA

BY

OWUSU MENSAH ISAAC

A THESIS SUBMITTED TO THE DEPARTMENT OF MATHEMATICS, KWAME NKRUMAH UNIVERSITY OF SCIENCE AND TECHNOLOGY IN PARTIAL FUFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF M.PHIL MATHEMATICAL STATISTICS

NOVEMBER 2015

DECLARATION

I hereby declare that this submission is my own work towards the award of the Mphil degree and that, to the best of my knowledge, it contains no material previously published by another person nor material which had been accepted for the award of any other degree of the university, except where due acknowledgment had been made in the text.

Owusu Mensah Isaac	.....	.....
Student	Signature	Date

Certified by:

Dr.A.Y. Omari-Sasu	.....	.....
Supervisor	Signature	Date

Certified by:

Prof. S. K. Amponsah	.....	.....
Head of Department	Signature	Date

ABSTRACT

Over the years, positively skewed data such as data from insurance, economics, laboratory, health and so on, have been analysed using conditional mean models suchassimplelinearregressionandlogisticregression. Estimationofthesemodels can be seriously deficient if constructed on some non-gaussian settings and cannot be readily extended to non-central location which is precisely where the interest of a social science research often reside. This study therefore seeks to employ a methodology to deal with these problems.

This study seeks to estimate the quantiles that describe the entire distribution and also to obtain an appropriate statistical distribution for the birth weight data. Our study used birth weight data from Komfo Anokye Teaching Hospital. Quantile, Lognormal and Gamma regression were used in the analysis and Quantile-Quantle plot and Akaike's Information Criterion(AIC) were the goodness of fit test for the selection of the distribution that fitted the data well. Finally we estimated 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} quantile regression to describe the entire distribution of the data. The lognormal also was selected as a better distribution than the gamma distribution based on their AIC values and the graph of the Q-Q plot.

DEDICATION

I dedicate this work to my father Mr. Ofori Atta who has been of great help to my academic career.

ABSTRACT

Over the years, positively skewed data such as data from insurance, economics, laboratory, health and so on, have been analysed using conditional mean models suchassimplelinearregressionandlogisticregression. Estimationofthesemodels can be seriously deficient if constructed on some non-gaussian settings and cannot be readily extended to non-central location which is precisely where the interest of a social science research often reside. This study therefore seeks to employ a methodology to deal with these problems.

This study seeks to estimate the quantiles that describe the entire distribution and also to obtain an appropriate statistical distribution for the birth weight data. Our study used birth weight data from Komfo Anokye Teaching Hospital. Quantile, Lognormal and Gamma regression were used in the analysis and Quantile-Quantle plot and Akaike's Information Criterion(AIC) were the goodness of fit test for the selection of the distribution that fitted the data well. Finally we estimated 5^{th}, 25^{th}, 50^{th}, 75^{th} and 95^{th} quantile regression to describe the entire distribution of the data. The lognormal also was selected as a better distribution than the gamma distribution based on their AIC values and the graph of the Q-Q plot.

ACKNOWLEDGMENTS

I express my profound gratitude to the Almighty God for his grace and mercies without which this work would not have been completed. I also wish to thank the following persons:

The first goes to Dr.A.Y. Omari-Sasu, the supervisor of this thesis, with whose guidance, constructive criticisms, and suggestions this work has become successful. The next thanks the head of the Komfo Anokye Teaching Hospital(KATH) records room(Mr. Sarpong) for giving me the data for the analysis. Another appreciation goes to all my colleagues at the department, especially Isaac Adjei Mensah, for their support. Last but not the least, the diverse support my family especially my parents, gave in doing this study is fully acknowledged and gratified.

CONTENTS

ABBREVIATIONS/ACRONYMS

BMI .....Body Mass Index

AIC .....Akaike's Information Criterion

Q-Q .....Quantile-Quantile

GLM ..... Generalized Linear Model

MSE ..... Mean Square Error

LBW .....Low Birth Weight

WHO ..... World Health Organization

DOH ..... Department of Health

SGA .....Small for Gestational Age

PTB ..... Protein Birth

GAMLSS .....Generalized Addictive Model Location,Scale and Shape

NHANESIII .....National Health And Nutrition Examination Survey

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

According to Boan Health(Chinese Manufacturer of Dietary Nutrition Supplement), at full term, the average baby will be about 20 inches (51cm) long and will weigh approximately 6 to 9 pounds(2700 to 4000 grams). Birth weight is the body weight of a baby at its birth. There have been numerous studies that have attempted with varying degrees of success, the links between birth weight and later life conditions, including diabetes, obesity, tobacco smoking and intelligence. Boan Health(www.boanhealth.com).

MedlinePlus (USA National Library for Medicine) also defined birth weight as the first weight of a baby, taken just after he or she is born. A low birth weight is less than 5.5 pounds and a high birth weight is more than 8.8 pounds. A low birth weight baby can be born too small, too early (premature), or both. This can happen for many different reasons. They include health problems in the mother, genetic factors, problems with the placenta and substance abuse by the mother. Some low birth weight babies may be more at risk for certain health problems. Some may become sick in the first days of life or develop infections. Others may suffer from long term problems such as delayed motor and social development or learning disabilities. High birth weight babies are often big because the parents are big, or mother had diabetes during pregnancy. These babies may be at risk of birth injuries and problems with blood sugar.

The prevalence of childhood obesity increased dramatically during the last decade in industrialized countries (Toschke et al, 2005). The increase in prevalence seems rather to be due to a shift of the upper part of the body mass index

(BMI) distribution than a shift of the entire BMI distribution as an example observed in the NHANES III survey from 1988 to 1994 (Flegal et al, 2005). This increased positive skewness could be due to exposure to obesogenic environmental determinants among a subpopulation with high degree susceptibility. TV watching, formula feeding, smoking in pregnancy, maternal obesity or parental social class are well known environmental, constitutional or sociodemographic risk factors (Toschke et al, 2005). However, it remains unknown if these factors affect the entire BMI distribution or only part of it.

A recent descriptive study reported an effect of several risk factors for childhood obesity on upper BMI percentiles, whiles the middle part of the BMI distribution was virtually unaffected. However, this study did not adjust for potential confounders (Toschke et al, 2005). BMI data usually have skewed distribution for which common statistical modeling approaches such as simple linear or logistic regression have limitations.

Many types of laboratory data and epidemiologic data studies involve the analysis of highly skewed data. When the distributions of outcome variables are highly skewed, the mean is sensitive to outliers and is not a good measure of central tendency. A transformation of the outcome variable is a popular approach to improve symmetry and normality for a linear regression. Quantile regression is another approach to analyse such data. Quantile regression analysis of hospital charges provide unbiased estimates even when lognormal and equal variance assumptions are violated.

Estimating loss severity distribution from a historical data is an important actuarial activity in insurance. A standard reference on the subject is Hogg and Klugman(1984). The limited data in the available statistical tables(such as Deininger and Squire, 1996; Tabatabai, 1996) addresses the question of how to obtain a reasonable picture of the income distribution within a country

Quantile regression as proposed by Koenker and Bassett (1978), has immerged as an important statistical methodology for addressing the limitations of simple

linear regression. The quantile regression model is a natural extension of the liner regression model by estimating various conditional quantile functions. This offers a strategy for examining how the covariates influence the entire response distribution.

Koenker and Hallock (2001) in their journal said that most of the analysis of birth weights has employed conventional least square regression method. However, the resulting estimates of the various effects on the conditional mean of birth weights were not necessarily indicative of the size and nature of the effects on the lower tail of the birth weight distribution. A more complete picture of the covariates effects can be provided by estimating a family of conditional quantile functions.

Kenneth(2011), defined lognormal distribution as a distribution whose logarithm is normally distributed but whose untransformed scale is skewed. Generally positive data are analyzed by lognormal and gamma models (McCullagh et al, 1989; Das and Lee, 2009; Das and Park, 2012; Firth, 1988) as variance of some positive data set may have relation with the mean. Recently lognormal and models (Myers RH et al, 2002) are of interest in fitting positive data arising from quality improvement experiment. Das and Lee (2009) studied positive data for quality improvement experiment under both lognormal and gamma joint generalized linear models. Das et al (2012) found that the lognormal models (with non-constant variance) are much more effective than either traditional simple , multiple and logistic regression with constant variance. They also suggested that to reduce infant mortality due to low birth weight, mothers should be non-smokers.

Positively skewed distribution of a random variable occurs in many statistical applications. Since it is difficult to compare population mean for several populations, the sample mean which is the natural estimate is compared. The sample mean is also very non-robust. For example, the mean cost of medically homogenous groups of patients are used for hospital budgeting and it is common to compare cost means among different hospitals or over different period of time.

It is easy to give examples where a few atypical stays drastically change the mean estimate and whose common test of means (e.g. t-test and its variants) lead to different decisions when these outliers are removed from the data set.

In recent years procedures for automatic outlier detections, robust mean estimationandcomparisonofrobustmeansofasymmetricdatahavebeenstudied. For instance procedures based on robust fitting of parametric models have been showntobeusefulinapplications(VictoriaFeserandRonchetti, 1994,1997; Feser, 2000; Marazzi et al, 1998). In this framework robust parametric mean is defined as the mean of the estimated model.

1.2 Problem Statement

A lot of work has been done in this area, especially comparing quantile regression to simple linear regression and multiple linear regression. Koenker and Bassett (1978) said that the conventional least square estimators may be seriously deficient in linear models constructed on some non-Gaussian settings, where quantile regression would provide more robust and consequently more efficient estimators. When a distribution is skewed, the mean is not a good measure of central tendency and for that matter the conventional least squares models are not good for such distributions. Quantile regression is therefore good for such distributions since we can determine the stochastic relationship between the covariates and the response variable at every quantile. Distributions such as gamma and lognormal distributions are also good for skewed data. Das and Park (2012), said the lognormal and gamma regression models (with non-constant variance) are much more effective than either traditional simple, multiple, logistic regression and log-Gaussian models (with constant variance), because they better fit the data. There has not been enough study to compare these models which are good for skewed data (positively skewed) to come out with the best. These problems have motivated my interest to dive into the problem and seek for solutions.

1.3 Objectives of the Study

The objectives of the research includes the following;

- Determination of the quantiles that describe the entire distribution of the data
- Selection of the appropriate statistical model for the birth weight data.

1.4 Significance of the Study

This study was geared towards demonstrating the importance of understanding statistical distributions for positively skewed data. This information will enable researchers to make important decisions regarding data that is positively skewed. The paper analyzes the theoretical back-ground of the modeling process which takes place with birth weight data. By using statistical distributions to model birth weight, one has a much added insight into the complexity handling birth weight data.

1.5 Methodology

In this study we were interested in a data that is positively skewed, so we used a birthweight data or Body Mass Index(BMI) data which is one of the positively skewed data. Any other positively skewed data(like data on finance, wealth, economic etc.) could have been used for the study. Quantile,lognormal and gamma regression models were all fitted using the data. AlkaikeInformationCriterion(AIC),Log-LikelihoodestimateandQuantile-Quantile(Q-Q) Plot were used to determine the distribution that fit the data. The response variable(dependent variable) was the birthweight of a child or the BMI of the child. The covariates(independent variables) were the age of the mother,

the occupation of the mother, marital status of the mother, the educational background of the mother and so on.

1.6 Organization of the Study

In summary the first chapter introduced the topic of the study and gave an overview of the background of the study. This chapter also discussed the statement of the research problem and the objectives of the study. A brief overview of the methods and the originators of the models used in this study has been discussed in this chapter. Chapter two elaborated on the works that have been done with the BMI data using some of the models used in this study. the chapter conceptualized the effectiveness the models with the BMI data and gave an overview of why the models are good in analysing BMI data. Comparative studies of the models and the traditional least squares models by different researchers have been discussed in this chapter. The chapter again entailed an accounts of using the models to analyse other positively skewed data (like data on finance, wealth, economic, etc.). Chapter three discussed thoroughly the methods and procedures for the study. The mathematical background of all the models used in the study(quantile, lognormal and gamma models), and how the models can be generated using Generalized Linear Models(GLM). The tools used for the assessment of the performance of these models(AIC,Log-Likelihood estimate etc) were also discussed in this chapter. The next chapter, which is chapter four, entails the analysis of the data and the performance of the models. All the summary statistics and the inferential analysis were given and elaborated in this chapter. Tables and Figures necessary for the analysis of our research questions and objectives and its interpretations have been given in the chapter. Chapter five further build on chapter four. Conclusions and recommendations of areas for further studies have been discussed in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews all the relevant work that relates to the our study. This include works on lognormal, gamma and quantile regression with positively skewed data like insurance data, economic data, body mass index(BMI) data and so on.

2.2 Related Works on Quantile regression

In literature, most authors used linear or logistic regression to model effects on BMI measures. However, BMI data are usually positively skewed and therefore a transformation of the response variable and/or other regression methods might be more appropriate. Possible approaches include lognormal Box Cox power transformation of the BMI prior to linear regression modelling, gamma regression, quantile regression or GAMLSS models.

Gardosi et al. (1995) used stepwise regression to model the conditional percentiles. They wanted to understand the relationship between the predictors andthetailsofthesevariables, sotheymodelledtheconditionalquantilefunctions of the birth outcomes.

Infants who are born preterm (gestational period less than 37 weeks) or small for gestational age (below the 10th percentile of birth weight after controlling for gestational age) have elevated rates of morbidity and mortality (Garite et al., 2004). Reasons for these associations include poorly functioning organs, reduced metabolism, insulin resistance, and increased susceptibility to adverse

environmental events later in life (Barker, 2006).

In a literature review Sr'am et al.(2005) argued that the relationships between air pollution and gestational age and intrauterine growth warrant further analysis. Their second scientific objective was to investigate the effect of maternal exposure to tropospheric ozone, one of the criteria pollutants regulated under the Environmental Protection Agency's Clean Air Act, on SGA and preterm birth (PTB).

Classical frequentist (Koenker, 2005) models a conditional quantile rather than the conditional mean as a function of predictors. This enables inference of noncentral parts of the distribution, makes fewer assumptions, and is more robust to outliers than mean regression. One limitation with these approaches is that multiple levels can produce crossing quantiles, where for some values of the predictors the quantile function is decreasing in quantile level.

Children classified as having a low or very low birth weight (below 2500g and 1500g, respectively) or a premature birth (less than 37 full weeks of gestation at birth) face increased risks of a range of problems including those in the physical , behavioral and mental (Lorenz et al., 1998) domains. An estimated 8 percent of births in the U.S. are low weight and 12.8 percent are preterm, so that the average birth in the U.S. is well above the cutoffs (Behrman and Butler, 2007). Thus, if some exposure decreases the birth weights of babies who would otherwise have average to high birth weights by some modest amount, but it does not decrease the birth weights of babies who would otherwise have low birth weights, it might not be considered deleterious. On the other hand, if an exposure lowers the birth weights of babies by the same modest amount who already would have low birth weights, but it does not decrease the birth weights of babies who otherwise would have average to high birth weights, the exposure would be troubling. Thus, for modeling birth weights it is appropriate to consider quantile regression over standard linear regression, which implies that the exposure has the same effect across the entire response distribution.

Although the 1500g and 2500g cutoffs are useful benchmarks, we prefer not to discretize birth weights into very low, low, and normal when modeling. Discretizations have scientifically unjustifiable consequences. For example, it is much worse to be born at 1501g than 2499g, yet a discretization would treat both of these as equivalent. Similarly, being born at 2501g is not appreciably better than being born at 2499g. Quantile regression allows us to model non-central aspects of the birth weight distribution while considering the information that discretization masks (Abrevaya, 2001).

In practice, it is not clear which single exposure metric of tobacco smoke exposure oneshoulduse. Labassaysofcotinine(ametaboliteofnicotine)levelsinmaternal blood or urine are a common measure of tobacco smoke exposure, but cotinine has a half-life of around nine hours in pregnant women . Hence, a single cotinine measurement may inaccurately reflect exposure over the course of the pregnancy. Alternatively, self-reported smoking measures can be biased by poor recall and misreporting. Wang et al. (2009) struggle with this exact issue. They find that cotinine levels are an important predictor of lower average birth weights, but that theevidenceislessclearwhenmaternalself-reportsofsmokingwereused(p. 984). They go on to write that, the stronger exposure-response relationship for cotinine concentrations suggests that this objective measure more accurately represents the individual differences in smoking behaviour (p. 984). While this may be true, it seems risky to judge the reliability of competing measurements based on the strength of a relationship that one simultaneously attempts to estimate. Using a confirmatory factor structure partially resolves this issue, in that it enables analysts to pool the information from these multiple, imperfect measurements in hopes of more accurately representing the exposure in the quantile regression.

As a related measurement issue, some individual exposures arguably affect birth outcomes through a common biological pathway, so that in actuality they are indicators of an underlying factor. For example, suppose that psychological stress presents differently for many mothers. Some may feel stress because

they are socially isolated, some because their pregnancy was unwanted, and others because they feel they are incapable of influencing events that affect their lives (Bandura, 2010). Further, suppose that high levels of psychological stress, however presented, activate biological processes that have a negative effect on birth weights. If the indicators are modestly correlated and have low incidence rates marginally, individually they may not be strongly associated with birth weight in analyses, even though their underlying factor is. The factor structure offers analysts a way to represent and estimate such underlying constructs in regression models.

Quantile regression has been applied in various BMI related studies. Several risk factors for increased adult body size had different effect on specific quantiles. Andreas et al, (2008) in their paper, Alternative regression models to assess increase in childhood BMI said that GAMLSS and quantile regression seem to be more appropriate than common GLMs for risk factor modelling of BMI data. Narchi et al. (2010) found that adjusting the conditional distribution of birth weight for biological variables better identified at risk infants.

The first scientific objective was to better define the conditional distributions of gestational age and birth weight by incorporating personal characteristics and environmental factors. They use information from Texas birth certificate records, including maternal parity, sex of the infant, parental education level, parental age and race. Modeling multiplequantile levelsthrough constraintson the coefficients ensures monotonicity of the quantile function, as in (Bondell et al., 2010) and the references therein.

The aforementioned approaches model a finite number of quantile levels and do not share information across quantile level. In applications where we expect inference at proximate quantile levels to be similar, it is useful to encourage communication across the distribution. Specifying the full quantile function, which entails separate parameter at an uncountable number of quantile levels, fosters this all-encompassing approach. Recent examples of quantile function

modeling include Reich et al. (2011), who investigated the effects of temperature on tropospheric ozone using Bernstein polynomials, and Tokdar and Kadane (2011), who analyzed birth weights using stochastic integrals.

Multiple conditional extremal methods exist in the literature. Wang and Tsai (2009)modeledthetailindex, whichdeterminesthethicknessofthetails, through a linear log link function of the parameters. Wang et al. (2012) quantile regressed in the shallow tails and extrapolated the results into the deep tails for thickly-tailed data. Our application requires inference across the distribution, so we follow the approaches of (Reich et al., 2011) and (Zhou et al., 2012), who modeled the middle of the distribution semiparametrically and a parametric form above a threshold. In these applications either zero (Zhou et al., 2012) or one (Reich et al., 2011) covariate affected the distribution above the threshold. Their methodological challenge was modeling a discrete response. The gestational age measurements have been rounded into values of 25,26,...42 weeks. Canonical discrete regression models make restrictive assumptions about the relationship between the response and the predictors. Dichotomizing the response by PTB restricts inference to the cut point between 36 and 37 weeks. Previous approaches in the literature (Machado and Silva, 2005) modeled one quantile by adding random noise to compel the response to behave continuously.

Lane and Jerome (2011) used Bayesian quantile regression model in the study of the predictors of birth weight. The results suggested that smoking during pregnancy is associated with decreased birth weight, even at the lower end of the response distribution. It was in accordance with the meta-analysis of Shah and Bracken (2000). However, the results did not suggest a significant effect of psychological factors on birth weight. Of course they missed the important confounders that mask effects in the study, as in the case with any observational study.

Luke et al, (2013) introduced a semi-parametric Bayesian quantile approach that model the full quantile function rather than just a few quantile levels.

Their multilevel quantile function model established relationship between birth weight and predictors separately for each week of gestational age and between gestational age and the predictors separately across Texas Public Health Region. They showed that pooling information across gestational age quantile level substantially reduce MSE of predictor effects relative to standard frequentist quantile regression. They found ozone to be negatively associated with lower tail gestational age in South Texas and across the distribution of birth weight for high gestational age.

Infants who are both preterm and small for gestational age (SGA) are at higher mortality risk than infants with either condition singly (Katz et al., 2013). Reich and Smith (2013) extended quantile function methodology to censored data. They faced three methodological hurdles in our application. PTB and low birth weight are closely related, but distinct, concerns. Researchers prefer to use SGA infants to isolate effects on birth weight from those on gestational age, so it is important to allow the relationship between birth weight and the predictors to vary by gestational age. While multilevel regression models are well-suited for jointly modeling a collection of distributions, standard hierarchical models assume the predictors affect only the conditional mean of the response. Second, considerable interest lies in the tails (particularly in very premature, SGA or large-for-gestational age births), so it is important to enable the tails of these distributions to be affected differentially by the predictors relative to the center of the distribution. Estimation of parameter effects at very low or very high quantiles is generally the purview of extreme value analysis.

2.3 Related Works on Gamma and Lognormal Regression

Lognormal distributions (with two parameters) have a central role in human and ecological risk assessment for at least three reasons. First, many physical,

chemical, biological, toxicological, and statistical processes tend to create random variables that follow Lognormal distributions (Hattis and Burmaster, 1994). For example, the physical dilution of one material (say, a miscible or soluble contaminant) into another material (say, surface water in a bay) tends to create non equilibrium concentrations which are Lognormal in character (Ott, 1995; Ott, 1990). Second, when the conditions of the Central Limit Theorem obtain (Mood, Graybill, and Boes, 1974), the mathematical process of multiplying a seriesofrandomvariableswillproduceanewrandomvariable(theproduct)which tends (in the limit) to be Lognormal in character, regardless of the distributions from which the input variables arise (Benjamin and Cornell, 1970). Finally, Lognormaldistributionsareself-replicatingundermultiplicationanddivision, i.e., products and quotients of Lognormal random variables are themselves Lognormal distributions(CrowandShimizu, 1988; AitchisonandBrown, 1957), aresultoften exploited in back-of-the-envelope calculations.

Most literatures (Lewit et al., 1995; Lavado et al. 2010; Reolalas and Novilla, 2010) have found strong associations between infant mortality and low birth weight (LBW). Although LBW is not a direct cause, the complications due to it (e.g. inability to maintain body temperature) account for 13.8 percent and 15.3 percent of infant deaths in the Philippines for the years 2006 and 2007, respectively. Also, these complications currently rank as the third leading cause of infant deaths both locally and globally (Reolalas and Novilla, 2010).

Aside from significant associations with infant mortality, LBW also has other negative effects particularly on physical and mental development of children. Barker (1997) has found that reduced fetal growth is strongly associated with many chronic conditions (e.g. cardiovascular disease, diabetes, obesity) in later life. Now known as the Barker's Hypothesis, it states that ?conditions in the maternal womb have a programming effect (fetal programming) on fetal physiology. For instance, when a fetus is deprived of adequate nutrient supply in the womb, it will develop a thrifty phenotype causing smaller body size and

lowered metabolic rate to name a few. In another study, LBW children are more likely to delay entry into school or attend special classes suggesting a direct link between birth weight and intelligence quotient (Corman and Chaikind, 1998). In the light of socioeconomic concerns, LBW babies result in higher economic costs for society such as higher health care costs and lower labor market payoffs. Even worse, socioeconomic inequality causes great disparity between LBW outcomes (Lewit et al., 1995).

Cheung et al, (2000) analysed the effect of early postnatal growth on motor development in Pakistani infants using the generalized lognormal model. The results showed that both fatal and early postnatal growth over a broad spectrum may affect infant motor development. It was not just the babies who were very small at birth that suffered. Birth length appeared to be more influential than other anthropometric indicators.

Francesca et al (2007) developed a measurement error model with counterfactual variables that address the scientific questions for the birth weight mortality case study. Their approach integrated Bayesian methods and data argumentation with counterfactual model and principal stratification. Francesca Dominici and her group first found that both Folic acid, Iron and vitamin A (F+I+A) and multiple nutrient and vitamin A (M+A) increase birth weight. However, the F+I+A increase birth weight mainly among LBW infants whereas M+A increase birth weight across the entire birth weight distribution compared to vitamin a only. The F+I+A reduce the risk of infant mortality, whereas the M+A slightly increases the risk of early infant mortality, especially among the larger infants.

The primary determinants of birth weight are gestation period and prenatal growth rate, while secondary factors consist of genetics and maternal behavior during pregnancy. External influences can be classified as environmental and socioeconomic factors such as educational attainment and wealth status. Many literatures and discussions on birth weight focus on prenatal care and micronutrient supplementation of the mother during her pregnancy. The

Department of Health (DOH) defined prenatal care as the use of health care during pregnancy, which includes screening for health conditions, providing therapeutic interventions, and educating women about safe child birth. Micronutrientsarecommonlyfoundinmanyironsupplementsbecauseofconstant concern about high prevalence of maternal iron deficiency (Allen and Gillespie, 2001). Prenatal care quality is considered as an essential indicator for maternal and infant health status. Lavado et al. (2010) has found that 96.16 percent of mothers had prenatal care but only 49.51 percent can be considered as good quality care. For the past decades, micronutrient supplementation during pregnancy had also earned great amount of interest in research in relation to birth weight. While there are several micronutrients (e.g. Zinc, Vitamin A, Calcium, Iodine) being associated with positive outcomes, the most important are Iron and Folic Acid.

The United Nations Children's Fund (UNICEF) defined LBW babies as newborns weighing less than 2,500 grams with the measurement taken within the first hour of life. Globally, 15.50 percent of total live births in 2008 are of LBW classification. In the Philippines, 21.20 percent of live births in 2008 are classified as LBW babies which is the largest for the past 23 years. Currently, the country ranks as the 14th (out of 225 countries) with the highest incidence of LBW cases (WHO, 2012).

Measures of body size, especially BMI, are associated with arsenic metabolism biomarkers. The association may be related to adiposity, fat free mass or body size (Mathew, 2013).

Das et al, (2014), concluded in their paper that lognormal and gamma regression models (with non-constant variance) are much more effective than either traditional simple, multiple, logistic regression and log-Gaussian models (with constant variance), because they better fit the data. They also said that to reduce the infant mortality due to low birth weight, white mothers with lower age should be a non-smoker, free of hypertension, free of uterine irritability with

higher weight at last menstrual period and without any premature labor.

Das(2014)inherjournalsaidthattheimpactofbiochemicalparameters, personal characteristics, family history and dietary factors on human plasma glucose concentration are explained based on mathematical relationships. Her results also identified many additional casual factors that explain the mean and variance of plasma glucose concentration.

Neonatal death is a serious concern, both in the developing and in the developed worlds. While infant mortality rates have been decreasing steadily all over the world, changes in neonatal mortality rate have been much slower. One of the commonest causes of neonatal mortality in the world is prematurity and low birth weight (Kramer,1987,Rich-Edward et al,2003,Kramer et al,2005, Basu et al, 2008). Generally, it is recognized that low birth weight can be caused by many factors (Collins et al,1990, David et al,1987, Cole et al,2002). Because many questions and conflicts still remain, however, about which factors exert independent causal effects, as well as the magnitude of these effects, a critical assessment and meta-analysis of the medical literature published from 1970 to till the date were carried out.

Neonate low birth weight has long been a subject of clinical and epidemiological investigations and a target for public health intervention. Low birth weight is defined by WHO as a birth weight less than 2500 g (before 1976, the WHO definition was less than or equal to 2500 g), since below this value birth-weight-specific infant mortality begins to rise rapidly (Rich-Edward et al,2003). In particular, considerable attention has been focused on the causal determinants of birth weight, and especially of low birth weight (LBW), in order to identify potentially modifiable factors. Many researches have focused on factors with well-established direct causal impacts on intrauterine growth include infant sex, racial/ ethnic origin, maternal height, pre-pregnancy weight, paternal weight and height, maternal birth weight, parity, history of prior low-birth-weight infants, gestational weight gain and caloric intake, general morbidity and episodic

illness, malaria, cigarette smoking, alcohol consumption, and tobacco chewing (Cole et al,2002). Note that these factors were identified based on preliminary statistical methods such as frequency distribution, odds ratio, simple regression analysis, logisticregressionetc. Thesemethodsmaynotidentifythedeterminants correctly in medical systems, demography and quality engineering process, as the variance of the response may be non-constant, and the variance may have some relationship with the mean (Das et al,2011). Generally, the above methods identify insignificant factors as significant and vice versa (Das and Lee ,2009), which is a serious error in any data analysis.

The present study analyzes the relationship of neonate birth weight (response) to the mother?s lifestyle explanatory variables. It has been identified that the response is non- constant variance. Consequently, two models (mean and variance) are derived. This particular analysis identifies the following: Mean neonate birth weight is explained by the statistically significant factors, mother weight at last menstrual period, her race, smoking status during pregnancy, history of premature labor, history of hypertension and presence of uterine irritability. Mother weight at last menstrual period is positively associated with her neonate mean weight, indicating that if mother weight at last menstrual period increases, her neonate birth weight will increase. Mother race is negatively associated with her neonate birth weight. It indicates that neonate birth weight will be lower for black women than white. Mother smoking status during pregnancy is negatively associated with her neonate birth weight. This implies that higher smoking status of mother during pregnancy decreases her neonate birth weight. Mother history of premature labor is negatively associated with her neonate birth weight. It indicates that if the mother number of premature labor increases, her neonate birth weight will decrease. Mother history of hypertension and presence of uterine irritability are negatively associated with her neonate birth weight. This implies that if mother hypertension and presence of uterine irritability increase, her neonate birth weight will decrease. Variance of

neonate birth weight is positively associated (statistical significant) with mother age, her history of hypertension and presence of uterine irritability. Thus, the neonate birth weight variance will increase with the increased of mother age, her hypertension and presence of uterine irritability. Therefore, the neonate birth weight variance will be lower for a mother with lower age, without hypertension and uterine irritability.

Hosmer and Lemeshow (2000) studied that the mother's lifestyle characteristics on her neonate birth weight based on the data described in Results Section. Similar study has been done by many researchers (Kramer MS,1987,Rich-Edward JW et al,2003). To identify the appropriate model, the earlier investigators used logistic regression techniques. Hosmer and Lemeshow (2000) also noted that the variance of the response (neonate birth weight) was non-constant, and its distribution was non- normal. Therefore, the researchers used logistic regression techniques by changing the responses (neonate birth weight) 0 (= birth weight >2500 g) and 1 (= birth weight <2500 g). Original responses are neglected, consequently, early researchers might loose many important information. For heteroscedastic data, log-transformation is often recommended to stabilize the variance (Box GEP and Cos DR,1964).

In practice, though, the variance is not always stabilized by this method. For example, Myers et al. analyzed The Worsted Yarn Data (Myers et al,2002) using a usual (errors are uncorrelated and homoscedastic) second-order response surface design. Myers et al. (2002) treated the response (y = T) as the cycles to failure (T), and also noticed that the variance was non- constant and the analysis was inappropriate. Then using log transformation of the cycles to failure (i.e., y= lnT), the final data analysis had been done, and it was found that log model, overall, was an improvement over the original quadratic fit. The researchers noticed, however, that there was still some indication of inequality of variance.

Recently, Das and Lee (2009) showed that simple log transformation was insufficient to reduce the variance constant, and the investigators analyzed the

data using joint generalized model. Das and Lee (2009) found that many factors were significant and the log-normal distribution was more appropriate. For non-constant variance of response, classical regression technique gives inefficient analysis, often resulting in an error so that the significant factors are classified as insignificant. In addition, positive data are generally analyzed by log-normal and gamma models (Myers et al,2002). For instance, the analysis by Myers et al. (2002) missed many important factors. This fault is very serious in every data analysis. The present authors notice that the original data set is positive, variance of the response is non-constant, distribution is non-normal, and original responses are neglected.

CHAPTER 3

METHODOLOGY

3.1 Introduction

This chapter focuses on the type of data used in the study. The chapter also elaborate the statistical methods and tools used in the analysis of the data.

3.2 Data

The data used in this study was the 2014 Birth Records from the Komfo Anokye Teaching Hospital(KATH). KATH is the only teaching hospital situated in Kumasi, the capital city of Ashanti Region of Ghana. Data management and preparation were aided by medical literatures and by an expert obstetrician-gynecologist.

The data consisted of 1007 observations with a set of 18 variables from which the researchers obtained 13 variables that are essential to the study. The data shows the variables (determinants of birth weight) and their descriptions. These variables contain information on maternal characteristics (age of mother at birth of child, age of the mother at first birth, educational attainment, employment status of the mother and total children ever born), weight gain during pregnancy, height of the mother and birth outcome (birth weight in kilograms and gender of child). The birth weight of the child was used as the response variable and the rest as the covariates. The variables with missing values in the original data were removed, so there were no missing values in the data used.

3.3 Models Description

3.3.1 Quantile Regression

For any real-valued random variable Y, with cummulative distribution function;

$$F(y) = P_r(Y \leq y) \tag{3.1}$$

for any $\tau \in (0, 1)$ or $0 \leq \tau \leq 1$, the τ^{th} quantile of Y is defined as;

$$Q(\tau) = \inf[y : F(y) \geq \tau] \tag{3.2}$$

The median quantile is then $Q(\frac{1}{2})$, the first quartile is $Q(\frac{1}{4})$ and the first decile is $Q(\frac{1}{10})$. The quantile function provides a complete characterization of Y, just like the distribution of F. The quantiles can be written as solutions to the following optimization problems. For any $\tau \in (0, 1)$, define a piecewise linear "check function" also known as the loss function as;

$$p_\tau(u) = u(\tau - I\, u(\leq 0)) \tag{3.3}$$

where I(.) is the indicator function. Solution to the minimization problem is then;

$$\alpha b(\tau) = \text{argmin}_{u \in R} E[p_\tau(Y - u)] = \min_u (\tau - 1) \int_{-\infty}^{u} (y - u) dF_Y(y) + \tau \int_{u}^{\infty} (y - u) dF_Y(y). \tag{3.4}$$

Setting the derivative of the loss function to zero and letting q_τ be the solution is then;

$$(1 - \tau) \int_{-\infty}^{\tau} dF_Y(y) - \tau \int_{q_\tau}^{\infty} dF_Y(y) = 0$$

the equation then reduces to;

$$F_Y(q_\tau) - \tau = 0$$

implying that;

$$F_Y(q_\tau) = \tau \qquad (3.5)$$

Hence q_τ is the τ^{th} quantile of the random variable Y.

The sample analogue of Q(τ) is based on the random sample $y_1, \ldots y_n$ of Y. The τ^{th} quantile can then be identified, in a split of (3.4) above as any solution to;

$$\alpha b_\tau = q b_\tau = \underset{q \in R}{\operatorname{argmin}} \sum_{i=1}^{n} p_\tau(y_i - q), = \underset{q \in R}{\operatorname{argmin}} \left[(\tau - 1) \sum_{y_i < q_i} (y_i - q) + \tau \sum_{y_i \geq q} (y_i - q) \right] \qquad (3.6)$$

Let x_i, i = 1,...,n, be a k×1 vector of regressors, we can then write the equivalent of expression (1) as;

$$F_{u_\tau}(\tau - x_i \hat{\beta}_\tau / x_i) = p_r(y_i \leq \tau / x_i) \qquad (3.7)$$

which is essentially a different form derived from the more familiar;

$$y_i = x_i \hat{\beta}_\tau + \mu_{\tau i} \qquad (3.8)$$

where the distribution of the error term $\mu_{\tau i}$ is left unspecified, the only constraint being the usual quantile regression $Q_\tau(\mu_{\tau i}/x_i) = 0$. Using the analogy, the estimation of conditional mean functions as in;

$$\hat{\beta} = \underset{\beta \in R^K}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - x_i \beta)^2, \qquad (3.9)$$

the linear conditional quantile function;

$$Q_Y(\tau / X = x) = x \beta_\tau \qquad (3.10)$$

can be estimated by solving the equivalent of expression (3.10) for this case;

$$\hat{\beta}_\tau = \arg\min_{\beta \in R^k} \sum_{i=1}^{n} \rho_\tau(y_i - x_i'\beta) \qquad (3.11)$$

3.3.2 Properties of Quantile Regression

Quantileregressionhasthefollowingimportantpropertiesthatdistinguishitfrom a ordinary least square regression.

Equivariance

In many situation it is preferable to adjust the scale of original variables or reparametize a model so that its result has a more natural interpretation. Such changes should not affect our qualitative and quantitative conclusions based on the regression output. Invariance to a set of some elementary transformations of the model is called equivariance in this context. Koenker and Bassett (1978) formulated four equivariance properties of quantile regression. Once we denote the quantile regression estimate for a given $\tau \in (0, 1)$ and observations (y,X) by $\hat{\beta}(\tau; y,X)$, then for any $p \times p$ nonsingular matrix $A, \gamma \in R^p$ and $a > 0$ holds

- $[\hat{\beta}(\tau; ay,X) = a\hat{\beta}(\tau; y,X)]$
- $[\hat{\beta}(\tau; -ay,X) = a\hat{\beta}(1 - \tau; y,X)]$
- $\hat{\beta}(\tau; y + X\gamma, X) = \hat{\beta}(\tau; y,X) + \gamma$
- $\hat{\beta}(\tau; y, XA) = A^{-1}\hat{\beta}(\tau; y,X)$

This means, for example , that if we use as the measurement unit of y millimeters instead of meters, that is y multiplied by 1000, then our estimate scales appropriately:

$$\hat{\beta}(\tau; y[mm],X) = 1000.\hat{\beta}(\tau; y[m],X).$$

Invariance to Monotonic Transformations

Quantiles exhibit besides "usual" equivariance to monotone transformations. Let f(.) be a nondecreasing function on R.. then it immediately follows from the definition of the quantile function that for any random variable Y ;

$$Q_{f(Y)}(\tau) = fQ_Y(\tau).$$

In other words, the quantiles of the transformed random variable f(Y) are the transformed quantiles of the original variable Y . This is not the case of the conditional expectation;

$$Ef(Y) = 6\ f(EY)$$

unlessf(.) isalinearfunction. Thisiswhyacarefulchoiceoftransformationofthe dependent variable is so important in the various econometric models when the ordinary least squares method is applied (unfortunately, there is ususaly no guide which one is correct) We can illustrate the strength of equivariance with respect to monotone transformation on the so-called censoring models. We assume that there exist, for example, a simple linear regression model with i.i.d. errors;

$$y_i = x\,\beta_i + \varepsilon_i$$

where $i \in 1,....,n$, and the response variable y_i is unobservable for some reason. Instead we observe $ye_i = maxy, a_i$ where $a \in R$ is a censoring point. Because of censoring the standard least squares method is not consistent anymore (but properly formulated maximum likelihood estimator can be used). On the contrary, the quantile regression estimator, thanks to the equivariance to monotone transformations, does not run into such problems as noted by Powell (1986). Using

$$f(x) = maxx, a$$

we can write;

$$Q_{y_i}(\tau/x_i) = Q_{f(y)}(\tau/x_i) = fQ_{y_i}(\tau/x_i) = f(x\,\beta_i\,) = \max x\,\beta, a_i \qquad 0$$

Thus, we can simply estimate the unknown parameters by;

$$\hat{\beta}(\tau) = \arg\min_{\beta} \in R^{P} \sum_{i=1}^{n} \rho_\tau (y_i - \max x\,\beta, a_i \qquad).$$

Robustness

Sensitivity of an estimator to departures from its distributional assumptions is another important issue. The sample mean , being a superior estimate of the expectation under normality of the error distribution, can be adversely affected even by a single observation if it is sufficiently far from the rest of data points. On the other hand, the effect of such a distant observation on the sample median is bounded no matter how far the outlying observation is. This robustness of the median is, of course, outweighed by efficiency in some cases. Other quantiles enjoy similar properties (the effect of outlying observations on the τ^{th} sample quantile is bounded,given that the number of outliers is lower than,

$$n \min \tau, 1 - \tau$$

Quantile regression inherits these robustness properties since the minimized objective functions in the sample quantiles (3.6) and in the case of quantile regression (3.11) are the same. The only difference is that regression residuals;

$$r_i(\beta) = y_i - x\,\beta_i$$

. are used instead of deviations from mean;

$$y_i - \mu$$

. Therefore, quantile regression estimates are reliable in the presence of outlying observations that have large residuals.

Asymptotic Property

For $\tau \in (0, 1)$, under some regularity conditions, $\beta\hat{b}$ is asymptotically normal:

$$\sqrt{n}(\beta\hat{b}_\tau - \beta_\tau) \xrightarrow{d} N(0, \tau(1-\tau)D^{-1}\Omega_x D^{-1}),$$

where;

$$D = E(f_y(X\beta)XX^0)$$

and

$$\Omega_x = E(XX^0)$$

. Direct estimation of the asymptotic variance-covariance matrix is not always satisfactory. Inference for quantile regression parameters can be made with the regression rank-score tests or with the bootstrap methods.

3.3.3 Computational Aspect

Quantile regression has a convenient linear programming(LP) representation. Using (3.8) and (3.11), we can translate to matrix notation.

$$y_i = \sum_{j=1}^{k} x_{ij}\beta_{\tau j} + \mu_{\tau j} = \sum_{j=1}^{k} x_{ij}(\beta_{\tau j} - \beta_{\tau j}^1) + (\varepsilon_{\tau i}^2 - v_{\tau i}) \qquad (3.12)$$

with $\beta_{\tau j}^1$, $\beta_{\tau j}^2$, $\varepsilon_{\tau j}$ and $v_{\tau i}$, non-negative $(j = 1, \ldots, k, i = 1, \ldots, n)$. The matrix notation for primal LP problem is then;

$$\min_z c^0 z \, st. Az = y, z \geq 0 \qquad (3.13)$$

where;

$$A = (X, -X, 1_n, -1_n)$$

,

$$z = ((\beta^{1})^{0}, (\beta^{2})^{0}, u, v^{0})^{0\,0}$$

$$c = (0^{0}, 0^{0}, \tau l, (1-\tau)l)^{0}$$

further 1_n is then n dimensional identity matrix, 0^0 is a $k \times 1$ vector of zeros and l is $n \times 1$ vector of ones. The dual side of the LP is easy to expose now after having obtained (3.13).

$$\max_{w} w^{0} y \text{ st. } w^{0} y \leq c^{0} \qquad (3.14)$$

The duality theorem implies that the solutions exist for both formation if X is a full rank matrix. Further, the equilibrium theorem of LP guarantees the optimality of the solution.

3.3.4 Lognormal Regression Model

We consider a regression model where the expected value of a continuous lognormal respose variable Y is a linear function of the predictors $X_1, X_2, \ldots, X_p$;

$$\mu_Y = \beta_0 + \beta_1 X_1 + \ldots\ldots + \beta_p X_p \qquad (3.15)$$

The variance of Y depends on both the expected value of Y, μ_Y and the variance of $Z = \ln(Y)$, σ_Z^2; $Var(Y/X) = \sigma_{Y/X}^2 (\exp(\sigma_Z^2) - 1).\mu_Y^2$. Ordinary least squares regression (here denoted by LS_{lin}) can be used to obtain unbiased estimates

$\beta_0, \beta_1, \ldots, \beta_p$. However, the estimates provided by LS assume homoscedasticity, which is incorrect for lognormal variable. This incorrect variance assumption leads to incorrect statistical inferences.

In a situation with heteroscedasticity, weighted least squares regression (here denoted by WLS) can be used. WLS can account for the heteroscedasticity by weighing each observation, Y_i with the inverse of its variance, $W_i \propto \sigma_{Y_i}^{-2}$. For a lognormal distribution, the weight for Y_i is $W_i = \frac{1}{\mu_{Y_i}^2}$, where, LS_{lin} can provide estimates of μ_{Y_i}. Unlike LS_{lin}, WLS provides an estimate of the variance σ_Z^2.

When the response Y is log-normally distributed, data are often log-transformes, ln(Y) = Z, and a log-linear model is estimated:

$$\mu_{Z|X} = \delta_0 + \delta X_1 \quad _1 + ... + \delta X_p \quad _P \qquad (3.16)$$

where the expected value of Y is $\mu_{Y|X} = \exp(\mu_{Z|X} + \sigma_Z^2/2)$. Ordinary least squares regression on Z (here denoted by LS_{exp}) provides estimates of the relative effect $(\delta_0, \delta_1, \ldots, \delta_p)$ as well as an estimate of the variance σ_Z^2 but no estimates of the absolute effects. Thus, both (3.15) and (3.16) can be used to estimate $\mu_{Y|X}$ and σ_Z. The reason for including LS_{exp}, even if the linear model in (3.15) is assumed, is that LS_{exp} is commonly used for lognormal data.

The lognormal distribution is often approximated by the gamma distribution, with parameters μ(expected value) and v(scale parameter, $Var(Y) = \mu/v$.) A generalized linear model (GLM) with gamma distribution and the identity link (denoted GLM_G), provides estimates $\beta_0, \beta_1, \ldots, \beta_p$ and an estimate of σ_Z^2 can be found through the transformation $\ln(1/v + 1) = \sigma_Z^2$.

Another GLM that can be used to estimate the absolute effects is one with a normal distribution and the link function $\exp(*)$, applied to $Z = \ln(Y)$, here denoted GLM_N, such that;

$$\exp(\mu_{Z|X}) = \varphi_0 + \varphi X_1 \quad _1 + ... + \varphi X_{.p} \quad _p \qquad (3.17)$$

The expected value of Y is often found as $\mu_{Y|X} = \exp(\mu_{Z|X}) \cdot \exp(\sigma_Z^2/2)$.

The method GLM_N, does not, however,take into account the stochastic variation due to estimating σ_Z^2. Therefore we also used a maximum likelihood method (ML_{LN}), (Gustavsson et al,2012) and (Yurgens ,2004), based on the likelihood function of the lognormal distribution;

$$f_Y(y) = \frac{1}{y} \cdot \frac{1}{\sqrt{2\pi\sigma_Z^2}} \exp\left[-\frac{(\ln(y) - \mu_Z)^2}{2\sigma_Z^2}\right] \qquad (3.18)$$

where $\mu_Z = \ln(\beta_0 + \beta X_{1\ i1} + ... + \beta X_{p\ ip}) - \sigma^2_Z/2$. The estimates $\hat{\beta}_0, \hat{\beta}_1, ... \hat{\beta}_p$, and $\sigma^2_{\hat{b}Z}$ are found using iterations, for example the Newton-Rapson iteration used here(Jensen et al, 2013)

Confidence Intervals

For LS_{lin}, WLS, GLM_G and ML_{LN}, a 95 percent confidence interval for $\mu_{Y|X}$ is estimated as;

$$\mu_{Y|X} \pm Z_{\alpha/2} \sqrt{var(\mu b_{Y|X})}$$

where the sample-specific variance is estimated as;

$$var(\mu b_{Y|X}) = \sum_{i=1}^{p} x_i^2 \cdot var(\beta_i) + 2 x\ x_{0\ 1} \cdot cov(\hat{\beta}_0, \hat{\beta}_{0\ 1}) + ... + 2x_{p-1}x_p \cdot cov(\hat{\beta}_{p-1}, \hat{\beta}_p) \quad (3.19)$$

where $x_0 = 1$, $var(\beta b_i)$ and $cov(\beta_i, \beta b_j)$ are sample-specific estimates of the variance and the covariance (the sample-specific standard error is $se(\beta b_i) = \sqrt{var(\beta b_i)}$).

For GLM_N, a confidence interval is estimated as $(\exp(\mu_{Y|X}) \pm z_{\alpha/2} \sqrt{var(\exp(\mu b_{Y|X}))}) \cdot \exp(\sigma^2_Z/2)$, where the sample-specific variance of the linear estimator is estimated as;

$$var(\exp(\mu b_{Z|X})) = \sum_{i=1}^{p} x_i^2 \cdot var(\varphi_i) + 2 x\ x_{0\ 1} \cdot cov(\hat{\varphi}_0, \hat{\varphi}_{0\ 1}) + ... + 2x_{p-1}x_p \cdot cov(\hat{\varphi}_{p-1}, \hat{\varphi}_p)$$

$i=0$

(3.20)

For LS_{exp}, a confidence interval for $\mu_{Y|X}$ is estimated as;

$$\exp\left(\mu_{b_{Z|X}} + \frac{\sigma^2_{bZ}}{2} \pm t_{\alpha/2,(n-p-1)} \sqrt{\mathrm{var}(\mu_{b_{Z|X}})^2 + \frac{(\sigma^2_{bZ})^2}{2(n-p-1)}}\right)$$

, using the modified Cox method(Niwitpong,2013). The sample-specific variance is estimated as;

$$\mathrm{var}(\mu_{b_{Z|X}}) = \sum_{i=0}^{p} x_i^2 \cdot \mathrm{var}(\delta_i^b) + 2x_0 x_1 \cdot \mathrm{cov}(\delta_0^b, \delta_1^b) + \ldots + 2x_{p-1}x_p \cdot \mathrm{cov}(\delta_{p-1}^b, \delta_p^b) \quad (3.21)$$

where $x_0 = 1$, $\mathrm{var}(\delta_i^b)$ and $\mathrm{cov}(\delta_i^b, \delta_j^b)$ are the sample specific estimates of the

variance and the covariance.

Some Properties of Lognormal Distribution.

- The random Variable that is lognormally distributed, can only assign positive real values in it.
- Lognormal distribution of a random variable x has two parameters (mean and standard deviation) which are denoted by μ and σ respectively. Then, we can write x in the following way;

$$x = e^{\mu+\sigma z}$$

where z is referred as standard normal variable. The μ is the location parameter and the σ is the scale parameter.

- Coeefficient of variation of lognormal distribution is;

$$CV = \sqrt{e^{\sigma^2} - 1}$$

.

- Mode and Median: The mode is the global maximum of the probability density function. In particular, it solves the equation;

$$(\ln f)' = 0$$

$$mode(X) = e^{\mu-\sigma^2}$$

. The median is such a point where $F_X = \frac{1}{2}$

$$Med(X) = e^{\mu}$$

.

- Geometric Moments: The geometric mean of lognormal distribution is $GM(X) = e^{\mu}$, and the geometric standard deviation is $GSD(X) = e^{\sigma}$. By analogy with the arithmetic statistics, we can define a geometric variance, $GVar(X) = e^{\sigma^2}$, and a geometric coefficient of variation $GCV(X) = e^{\sigma} - 1$. Because the log-transformed variable $Y = \ln X$ is symmetric and quantiles are preserved under monotonic transformation, the geometric mean of a lognormal distribution is equal to its median, Med(X).

  $GM(X) < AM(X)$ and this is due to the $AM - GM$ inequality, and corresponds to the logarithm being convex down. In fact,

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2}$$

$$E(X) = e^{\mu} \cdot e^{\sigma^2}$$

$$E(X) = GM(X) \cdot \sqrt{GVar(X)}$$

  In finance, the term $e^{\frac{1}{2}\sigma^2}$ is sometimes interpreted as a *convexity correction*, from the point of view of stochastic calculus, this is the same correction term as *Ito's lemma for geometric Brownian motion*.

- Arithmetic Moments: The arithmetic mean, arithmetic variance and arithmetic standard deviation of a log-normally distributed variable X are given by;

$$E(X) = e^{\mu + \frac{1}{2}\sigma^2}$$

$$Var(X) = (e^{\sigma^2} - 1)e^{2\mu + \sigma^2}$$

$$Var(X) = (e^{\sigma^2} - 1)(E(X))^2$$

$$SD(X) = \sqrt{Var(X)}$$

p

$$= E(X) \quad e^{\sigma_2} - 1$$

Thelocation μ andthescale σ^2 parameterscanbeobtainedifthearithmetic

mean and the arithmetic variance are known; it is simpler if σ is computed first:

$$\mu = \ln(E(X)) - \frac{1}{2}\ln\left(1 + \frac{Var(X)}{(E(X))^2}\right)$$

$$= \ln(E(X)) - \frac{1}{2}\sigma^2$$

$$\sigma^2 = \ln\left(1 + \frac{Var(X)}{(E(X))^2}\right)$$

For any real or complex number s, the s^{th} moment of a log-normally distributed variable X is given by;

$$E(X^s) = e^{s\mu + \frac{1}{2}s^2\sigma^2}$$

A lognormal distribution is not uniquely determined by its moments $E(X^k)$ for $k \geq 1$, that is, there exist some other distributions with the same moments for all k. In fact there is a whole family of distribution with the same moments as the lognormal distribution.

Maximum Likelihood Estimation of Parameters.

Determination of the maximum likelihood estimators for lognormal distribution parameters μ and σ, follows the same procedure as the normal distribution. We can observe that;

$$f_L(x; \mu,\sigma) = \prod_{i=1}^{n}\left(\frac{1}{x_i}\right) f_N(\ln x; \mu,\sigma) \qquad (3.22)$$

where f_L is the probability density function of the lognormal distribution and f_N is that of a normal distribution. This implies that we can write the log-likelihood function as;

$$\grave{}_L(\mu,\sigma|x_1,\ldots,x_n) = -\sum_k \ln x_k + \grave{}_N(\mu,\sigma|\ln x_1,\ldots,\ln x_n) \qquad (3.23)$$

Since the first term is constant with regards to μ and σ, both logarithmic likelihood functions $\grave{}_L$ and $\grave{}_N$, reach their maximum with the same μ and σ.

Hence, using the formulas for normal distribution maximum likelihood parameter estimators and the equality above, we can deduce that for the lognormal distribution, it holds that;

$$\mu b = \frac{\overline{X \ln x}}{\overline{k}} \qquad (3.24)$$

and,

$$\sigma b = \sqrt{\frac{P}{n} {}_k(\ln x_k - \mu b)^2} \qquad (3.25)$$

3.3.5 Gamma Regression

The probability density of observing a particular value y_i, given the shape parameter α_i and scale parameter β_i is;

$$f(y) = \frac{1}{\Gamma(\alpha)\beta_i^{\alpha_i}} \cdot y_i^{(\alpha_i - 1)} e^{-(y_i/\beta_i)}, y_i \alpha_i \beta_i > 0 \qquad (3.26)$$

where, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$, $E(y_i) = \alpha \beta_i$ and $Var(y_i) = \alpha \beta_i^2$.

Regression with that gamma model is going to use input variables X_i and coefficients to make a prediction about the mean of y_i, but in actuality, we are really focused on the scale parameter β_i. Generally, we assume $\alpha_i = \alpha$,α_i is the same for all observations. Variation from case to case in $\mu_i = \beta \alpha_i$ is due simply to variation in β_i. The shape parameter is just a multiplier (which is equal to the inverse of the "dispersion parameter" φ that is defined for all distributions that are members of the exponential family).

Linkage Between Mean and Variance

The ratio of the mean to the variance is a constant; the same, no matter how large or small the mean is. As a result, when the expected value is small (near zero), the variance is small as well. Conversely, when the expected value is large, the observed scores are less predictable in absolute terms.

$$\frac{Var(y_i)}{E(y_i)} = \frac{\alpha \beta_i^{\ 2}}{\alpha \beta_i} = \beta_i$$

If the gamma variable has an expected value of 100, the variance has to be $100 \cdot \beta_i$. The so-called coefficient of variation, which is used in introductory statistics as a summary of variability, is the ratio of the standard deviation to mean. It is also constant;

$$CV = \frac{\sqrt{Var(y_i)}}{E(y_i)} = \frac{\sqrt{\alpha \beta_i^2}}{\alpha \beta_i} = \frac{\sqrt{\alpha_i}\beta_i}{\alpha \beta_i} = \frac{1}{\sqrt{\alpha_i}}$$

If the gamma variable expected value is 100, the standard deviation is $100/\sqrt{\alpha_i}$. The ratio Var/E depends on β_i but the StdDev/E depends on α_i. The relationship between mean and variance here is different than some other distributions, because it is "adjustable". In contrast the poison and the binomial distribution have no such turning parameters.

Gamma Model

The gamma regression model can be viewed as a class of Generalized Linear Model (GLM). In general GLM has three components;

- Random Component: This component specifies the conditional distributionof the response variable, Y_i,(for the ith of n independently sampled observations), given the values of the explainatory variables in the model. In Nelder and Wedderburn's original formulation, the dristribution of Y_i is a member of the *exponential family*, such as Gaussian(Normal), Binomial, Poison, Gamma, orInverse-Gaussianfamilies of distributions. Subsequent work, however, has extended GLM's to multivariate exponential families (such as multinomial distribution), to certain non-exponential families (such as the two parameter negative binomial distribution), and to some situations in which the distribution of Y_i is not specified completely.

- Systematic Component: This component is a linear function of the regressors (predictors). That is;

$$\eta = \alpha + \beta X_{i1} + \beta X_{i2} + ... + \beta X_{ik}.$$

As in the linear model, and in the logit and probit models, the regressors X_{ij} are pre-specified functions of the explanatory variables and therefore may include quantitative explanatory variables , transformation of qualitative explanatory variables, polynomial regressors, dummy regressors, interactions, and so on. Indeed one of the advantages of of GLM s^{0}is that the structure of the linear predictor is the familiar structure of a linear model.

- Link Function: The last component is a smooth and invertible linearizing linkfunction g(.), whichtransformstheexpectationoftheresponsevariable, $\mu \equiv E(Y_i)$, to the linear predictor:

$$g(\mu_i) = \eta_i = \alpha + \beta X_{1\ i1} + \beta X_{2\ i2} + ... + \beta X_{k\ ik}$$

Because the link function is invertible, we can also write,

$$\mu = g^{-1}(\eta_i) = g^{-1}(\alpha + \beta X_{1\ i1} + \beta X_{2\ i2} + ... + \beta X_{k\ ik})$$

and, thus, the GLM may be thought of as a linear model for a transformation of the expected response or the nonlinear regression model for the response. The inverse link $g^{-1}(.)$ is also called the *mean function*. The *identity link* simply returns its argument unaltered, $\eta_i = g(\mu_i) = \mu_i$, and thus $\mu_i = g^{-1}(\eta_i = \eta_i)$.

Gamma as a Member of Exponential Family

We treat this as a basis for GLM, by treating α as known feature, the same for all observations and β_i(the scale parameter) as the parameter of interest. Exponential family has the form;

$$\exp\left[\frac{y_i.\theta_i - c(\theta_i)}{\varphi} + h(y_i, \varphi)\right] \qquad (3.27)$$

Rearranging the density for the gamma as follows;

$$\exp[-\frac{y_i}{\beta_i}+(\alpha-1)\ln(y_i)-\ln(\beta_i)^{\alpha}-\ln[\Gamma(\alpha)]]$$

$$[\exp[-\frac{y_i}{\beta_i}+(\alpha-1)\ln(y_i)-\alpha\ln\beta-\ln[\Gamma(\alpha)]]$$

$$[\exp[-\frac{y_i}{\beta_i}-\alpha\ln\beta_i+(\alpha-1)\ln(y_i)-\ln[\Gamma(\alpha)]]$$

Here, the natural parameter is,

$$\theta_i=-\frac{1}{\alpha\beta_i}$$

consequently,

$$-\frac{1}{\beta_i}=\alpha\theta_i$$

and

$$\beta_i=-\frac{1}{\alpha\theta_i}$$

Using those findingsin the previous expression,

$$\exp[\alpha y\,\theta_{i\ i}-\alpha\ln(-\frac{1}{\alpha\theta_i})-\alpha\ln(\alpha)+(\alpha-1)\ln(y_i)-\ln[\Gamma(\alpha)]]$$

$$\exp[\alpha y\,\theta_{i\ i}-\alpha\ln(-\frac{\alpha}{\alpha\theta_i})+(\alpha-1)\ln(y_i)-\ln[\Gamma(\alpha)]]$$

$$\exp[\alpha y\,\theta_{i\ i}-\alpha\ln(-\frac{1}{\theta_i})+(\alpha-1)\ln(y_i)-\ln[\Gamma(\alpha)]]$$

$$\exp[\alpha(y\,\theta_{i\ i}-\ln(-\frac{1}{\theta_i}))+(\alpha-1)\ln(y_i)-\ln[\Gamma(\alpha)]]$$

That was quite a lot of work to find out that $\alpha=\frac{1}{\varphi}$ and that $c(\theta_i)=\ln(-)^{\frac{1}{\theta_i}}$ But if we re-arrange just one more time, we find the Gamma in the form of the exponential density;

$$\exp[\frac{y\,\theta_{i\ i}-\ln(-1/\theta_i)}{\varphi}+(\frac{1-\varphi}{\varphi})\ln(y_i)-\ln[\Gamma(\varphi^{-1})]] \qquad (3.28)$$

But $\mu_i = dc(\theta_i)/d\theta_i$, and so that implies the Gamma's μ_i is,

$$\frac{dc(\theta_i)}{d\theta_i} = \frac{d\ln(-1/\theta_i)}{d\theta_i} = -\frac{d\ln(\theta_i)}{d\theta_i} = -\frac{1}{\theta_i} = \alpha\beta_i$$

and that $V(\mu_i) = d^2c(\theta_i)/d\theta_i^2$ and so, in this case,

$$V(\mu_i) = \frac{d}{d\theta_i^2}(-1/\theta) = \frac{1}{\theta_i^2} = \mu^2 = (\alpha\beta_i)^2$$

The observed variance of y_i in GLM has two components, that is;

$$Var(y_i) = \varphi V_i(\mu_i)$$

. For the Gamma, we already know that $E(y_i) = \mu_i = \alpha\beta$ and $Var(y_i = \alpha\beta^2)$. The variance function is $V(\mu_i) = \mu_i^2 = \alpha^2\beta^2$, and the dispersion parameter φ_i must be equal to the reciprocal of the shape parameter $(1/\alpha)$. This implies that;

$$Var(y_i) = \varphi V_i(\mu) = \varphi_i \cdot \alpha\beta^2 = \alpha\beta^2$$

where,

$$\varphi_i = \frac{1}{\alpha}$$

Canonical Link

The canonical link for the GLM with a Gamma-distributed dependent variable is the reciprocal, $1/\mu_i$. this means that the expected value of the observed y_i, $(E(y_i) = \mu_i)$, is related to the input variables as;

$$\frac{1}{\mu_i} = \alpha + \beta x_{1i} \qquad (3.29)$$

which implies,

$$\mu_i = \frac{1}{\alpha + \beta x_{1i}} \qquad (3.30)$$

Properties of Gamma Distribution

The distribution has a zero lower bound and is unlimited on the right. It is positively skewed,the amount of skew depending inversely on the shape factor α. The mode of the distribution is at $\beta(\alpha - 1)$ if $\alpha > 1$ and at zero if $0 < \alpha \leq 1$. In the latter case, the distribution is J − shaped. For $\alpha = 1$, the distribution is exponential with ordinates $1/\beta$ at $y = 0$; for $\alpha < 1$ the ordinate at $y = 0$ is infinite. The Gamma distribution is closely related to the chi-square distribution, for $\chi/2$ is a gamma variate with $\alpha = \frac{1}{2}n$ and $\beta = 1$.
The moments about zero of the gamma distribution are given by the relation,

$$\mu_r' = \beta^r \alpha(\alpha + 1)....(\alpha + r - 1)$$

, from which it follows immediately that the mean is,

$$\mu_1' = \beta\alpha$$

. From the moments relationships the second, third and fourth moments about the mean are easily found to be,

$$\mu_2 = \sigma^2 = \beta^2 \alpha.....(i)$$

$$\mu_3 = 2\beta^3 \alpha.....(ii)$$

$$\mu_4 = 3\beta^4 \alpha(\alpha + 2)$$

Since the skewness statistic is $\sqrt{b_1} = \mu_3 / \sigma_3^3$ we have from (i) and (ii) that,

$$\sqrt{b_1} = \frac{2}{\sqrt{\alpha}}$$

. Hence, the skewness goes to zero with increasing α showing that the gamma distribution becomes symmetrical for large α;in fact, it may be shown that the

distribution approaches normality slowly as α increases.

Statistical Estimators

It has long been known that there are many ways to estimate the parameters in a statistical equation from a sample of data. Two of the more common methods are least squares and moments. It was found by Fisher (1941), that the various methods of estimation do not give equally good results in the sense that some estimates or statistics are more reliable than others.Clearly, the best estimates are those which have smaller variability. For example, in samples of 10 from normal population the mean (expected value) could be estimated by averaging the smallest and the largest value or it could be estimated by averaging all the observations. Obviously, the latter statistic using all the observation should be better than the one using only the two observations. In fact, it has been shown that the variability from sample to sample for sample size 10 as measured by the variance is twice as large as when the only estreme observations are averaged. The median, which is also an estimate of mean, has variance about one-third greater than the mean for a sample of 10. From this it may be inferred that if we use the mean range as an estimate, we in effect discard half of our data;if we use the median, we discard one-third of it.

Fisher (1941), made a remarkable contribution to the statistical analysis by developing a method of estimation originally due to Gauss which he called the method of maximum likelihood (M.L.). this method consist of maximising what he calls the likelihood or the product of the frquency functions of the sample. If $f(y; \beta,\alpha)$ is any frequency function, the likelihood is defined as,

$$M = \prod_{i=1}^{n} f(y_i; \beta,\alpha). \qquad (3.31)$$

where y_i is the ith value in a sample of n. To maximise this, it is simplest to take

logarithms before differentiating and setting to zero. this gives,

$$L = \sum_{i=1}^{n} \ln(y_i; \beta, \alpha). \qquad (3.32)$$

Differentiating partially with respect to β and α gives the M.L. differential equations,

$$\left.\begin{aligned} \frac{\partial L}{\partial \beta} &= 0 \\ \\ \frac{\partial L}{\partial \alpha} &= 0 \end{aligned}\right\} \qquad (3.33)$$

Solving these gives the M.L. estimates commonly written as βb and αb. The M.L. estimates have certain remarkable advantages not always possessed other estimates which will now be discussed.

In order to assess the quality of estimators in general, Fisher defined three desirable properties of statistics; consistency, efficiency, and sufficiency. These may be defined as follows:

- If an estimator or statistic is consistent, it converges in probability to its population or parameter value. This may be expressed as,

$$P(|T_n - \theta| <) > 1 - \eta; n > N. \qquad (3.34)$$

  T_n is an estimate of the parameter θ based on sample size n, and η are arbitrarily small quantities, and N is any integer. This means that $T_n = \theta$ when T_n is calculated from the whole population.

- A consistent estimate T_1 is said to be more efficient than another consistent estimate T_2 if $v(T_1) < v(T_2)$; that is the variance of T_1 is less than the variance of T_2. An estimate is said to be efficient if it has the smallest variance of a class of consistent estimates. The efficiency of an estimate is

defined as v(Tb)/v(T) where Tb is M.L. estimate.

- An estimate T is said to be sufficient if it exhausts all possible information

on θ from a sample of any size. If T_1 and T_2 are two different estimate of θ not functionally related, an estimate T_1 of θ is sufficient if the jiont distribution of T_1 and T_2 has the form,

$$f = f_1(T, \theta_1) f_2(T/T_1). \tag{3.35}$$

where f_1 is the frequency distribution of T_1 and f_2 is the distribution of T_2 given a sample value of T_1. Once T_1 is known, the probability of any range of values for T_2 is the same for all θ; hence, T_2 cannot give any information on θ which is not already available from T_1. Sufficiency is the most desirable property of an estimate, and such estimates are said to be optimum. The superiority of M.L. estimates was demonstrated by Fisher and others when they proved that M.L. estimates are consistent and efficient and if a sufficient estimate exists, it will be given by the M.L. method.

Maximum Likelihood Estimation of Parameters

Applying (3.32) to the gamma distribution equation (3.26) gives,

$$L = -n\alpha \ln \beta - n \ln \Gamma(\alpha) + (\alpha - 1) \sum \ln y - \frac{1}{\beta} \sum y \tag{3.36}$$

where the summation is over the n sample values. Differentiating as indicated in equation (3.33) we find the M.L. equations,

$$\bar{y}/\beta - b - \alpha b = 0 \tag{3.37}$$

$$\ln \beta b + \frac{\partial}{\partial \alpha b} \ln \Gamma(\alpha b) - \frac{1}{n} \sum \ln y = 0 \tag{3.38}$$

Since $\frac{\partial}{\partial \alpha b} \ln \Gamma(\alpha b)$ is a digamma function, $\psi(\alpha b)$, we may write (3.38) in the simplified form,

$$\ln \beta b + \psi(\alpha b) - \frac{1}{n} \sum \ln y = 0. \tag{3.39}$$

Taking logarithms of (3.37) and substituting for βb in (3.39) gives,

$$\ln \alpha b - \psi(\alpha b) = \ln \bar{y} - \frac{1}{n} \sum \ln y \qquad (3.40)$$

This equation is implicit in αb but may be solved with some difficulty using the Davis (1933) tables of the ψ-functions. Masuyama and Kuroiwa (1951) prepared tables of $\ln \alpha b - \psi(\alpha b)$ from tables of logarithms and tables of digamma functions. We developed the application of the gamma distribution precipitation before Masuyama and Kuroiwa's tables were available although, of course, we had also followed the equivalent procedure of using the Davis tables. To simplify the technique of fitting, we developed an approximation to $\ln \alpha b - \psi(\alpha b)$ as follows: Norlund (1924),

$$\psi(\alpha) = \ln \alpha - 1/(2\alpha) - \sum_{k=1}^{m} (-1)^{k-1} B_k/(2k\alpha^{2k}) + R_m. \qquad (3.41)$$

is an asymptotic expansion in which B_k are the Bernoulli numbers, $B_1 = 1/6$, $B_2 = 1/30$, etc, and R_m is the remainder after m terms. For $\alpha \geq 1$, we may write the inequality,

$$|R_m| < \frac{B_{m+1}}{(2m+2)\alpha^{2m+2}}. \qquad (3.42)$$

For only $m = 1$ and $\alpha = 1$, $|R_m| < 0.00833$ which is less than 1.5percent of the table value $\psi(1) = -0.57722$ given by Davis (1933). The approximation, of course, increases in accuracy with α. At $\alpha = 2$ it is within 0.1percent of table value. We are not, however, interested in approximating ψ but in approximating α. From (3.41) for $m = 1$ we find,

$$\psi(\alpha) = \ln \alpha - 1/(2\alpha) - 1/(12\alpha^2) \qquad (3.43)$$

Subtituting in (3.40) we find,

$$1 \quad X$$

$$12(\ln \bar{y} - \underset{n}{\quad} y)\hat{a}\hat{b}^2 - 6\hat{a}\hat{b} - 1 = 0. \tag{3.44}$$

$$\bar{y} - \frac{1}{n}\sum \ln y$$

Simplifying by letting $A = \ln \bar{y} - \frac{1}{n}\sum \ln y$ we have,

$$12A\alpha b^2 - 6\alpha b - 1 = 0, \tag{3.45}$$

which is a quadratic equation whose only pertinent root is,

$$\alpha b = \frac{1 + \sqrt{1 + 4A/3}}{4A}. \tag{3.46}$$

This together with equation (3.37) gives the M.L. estimates for the gamma distribution.

3.3.6 Model Evaluation and Selection

The effectiveness of each model can be evaluated by testing the significance of the coefficient of the covariates.

Wild Test

. For non-normal data, we can use the fact that $\hat{\beta} \sim N(\beta, \varphi(X'WX)^{-1})$ and use the z-test to test the significance of the coefficients.

Specifically, we test,

$$H_0 : \beta_j = 0$$

vrs

$$H_1 : \beta_j = 06$$

using the test statistic;

$$\beta b$$

j

$$Z_j = p \qquad (3.47)$$

$$\varphi(X\ WX_0\quad)_{-1}$$

which is asymptotically N(0, 1) under H_0.

Standard Error.

The estimate βb have the usual properties of the maximum likelihood estimators. In particular, β̂ is asymptotically N(β,i $^{-1}$), where i(β) = φ $^{-1}$ (X W X). Standard errors for β_j may therefore be calculated as a square roots of the diagonal elements of;

$$\text{cov}(\hat{\beta}) = \varphi(X' W X)^{-1}$$

in which $(X' W X)^{-1}$ is a by-product of the final Iterative Weighted Least Squares(IWLS) iteration. If φ is unknown, an estimate is required.

There are a practical difficulties in estimating the dispersion φ by maximum likelihood. Therefore it is usually estimated by *method of moments*. If β was known, an unbiased estimate of φ = α var$_i$ (Y)/v(μ_i) would be,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\alpha_i(y_i-\mu_i)^2}{v(\mu_i)}$$

Allowing for the fact that β must be estimated, we obtain,

$$\frac{1}{n-p}\sum_{i=1}^{n}\frac{\alpha_i(y_i-\mu_i)^2}{v(\mu_i)}$$

Model Selection.

The probability distributions of lognormal and gamma are tested to see which one fit the data well. Since the parameters of the distributions are obtained using the maximum likelihood, the criteria for choosing one distribution out of the two is also based on the values of the estimated maximum likelihood estimates, the larger the likelihood, the better the model (Fiete, 2005).

Checking Model Fit

. It is assumed that no model in the set of models is true; hence selection of the better approximating model is the main goal (Anderson and Burnham, 2004).

A distribution getting the higher log-likelihood is not sufficient evidence to show that it is the right distribution for the data. Therefore an assessment would be made on how good this distribution fit the data using the Maximum Likelihood Estimate, Q-Q Plots(Quantile Quantile Plot) and the A.I.C (Akaike Information Criterion), with a significance level of $\alpha = 0.05$.

The Quantile-Quantile (Q-Q) Plots

. The Q-Q plots are graphical techniques used to check whether or not a sampled data set could have come from some specific target distribution, that is, to determine how well a theoretical distribution models the set of sampled data provided. This study will use the Q-Q plots to check for the goodness of fit of the distribution that would be chosen for the data. The Q-Q plots is chosen because of their multiple functions while analysing data sets and also because of their advantages.

The first Q stands for the quantiles of the sampled data set and the second Q stands for the quantile of the distribution being checked whether the data fits. In this case, the Q-Q plots is a plot of the target population and quantile against the respective sample quantile. If the sample data follows the distribution suspected, then the quantiles from the sample data would lie close to where they might be expected and the points on the plot would straggle about the lie $y = x$.

Theoretically, in order to calculate the quantiles of the distribution, this target distribution must first be specified, i.e. its population mean and the standard deviation but in practise, the sample estimates are used, therefore sample mean and satandard deviation of the distribution were estimated to be same as the ones of the sampled data set.

One of the advantages of the Q-Q plots is that the sample sizes do not need to be equal. Another one is that many distributional aspect can be simultaneously tested, for example shifts in locations, shifts from scale, changes in symmetry and the presence of outliers. This is important because if the data set come from

populations whose distributions only differ by shift in location, the points should lie along a straight line that is displaced up or down from the 45-degree reference line.

The Akaike Information Criteria (A.I.C.)

The A.I.C. is a type of criteria used in selecting the best model for making inference from a sampled group of models. It is an estimation of kullback-leibler information or distance and attempts to select a good approximating model for inference based on the principle of parsimony (Anderson and Burnham, 2004). The criterion was derived based on the concept that truth is very complex and that no ”true model” exists. Therefore in A.I.C, the model with the smallest value of A.I.C is selected because this model is estimated to be closet to the unknown truth among the candidate models considered.

The A.I.C is a measure of fit that penalizes for the number of parameters p. It is defined as,

$$AIC = -2\ln(L) + 2p. \tag{3.48}$$

where ln(L) is the maximized log-likelihood and p is the number of parameters estimated.

A.I.C rewards goodness of fit (as assessed by the likelihood function), but it also includes a penalty that is an increasing function of the number of parameters estimated. The penalty discourages *overfitting* (increasing the number of parameters in the model almost always improves the goodness of the fit).

Suppose that the data is generated from an unknown process f. We consider two candidate models to represent f; g_1 and g_2. If we knew f, then we could find the information lost from g_1 to represent f by calculating the *kullback-leibler divergence*, $D_{KL}(f \; k \; g_1)$; similarly the information lost from using g_2 to represent f could be found by calculating $D_{KL}(f \; k \; g_2)$. We would then choose the model that minimised the information loss. We cannot choose with certainty, because we do not know f. Akaike (1974) showed, however, that we can estimate, via

A.I.C, how much more (or less) information is lost by g_1 than by g_2.

CHAPTER 4

RESULTS AND ANALYSES

4.1 Introduction

This chapter entails the preliminary analysis as well as the inferential analysis of the study. The first part of the analysis investigate whether the distribution of our dependent variable (birth weight of a baby) is truely positively skewed and the other part of the analysis try to answer our research objectives.

4.2 Preliminary Results

The table below gives the variables in the data.

Table 4.1: Varaibles in the Data

Variable	Description
MOTAGE	Age of the Mother
MSTAT	Marital Status of the Mother
FBAGE	First Birth Age of the Mother
SWATER	Source of Water
NCHILD	Number of Children
EDULEVEL	Educational Level of the Mother
emstatus	Employment Status of the Mother
GESAGE	Gestational Age
BWEIGHT	Weight of the Baby
MWEIGHT	Weight of the Mother
MHEIGHT	Height of the Mother
SEXB	Sex of the Baby

All the variables in the table above are independent variables except BWEIGHT which is the response variable.

4.2.1 Summary Statistics

The modelling process start with the computation of the summary statistics of the dependent variable(BWEIGHT).These are presented in the table 4.1 below. This summary was necessary in pointing out the salient features of the data. From table 4.1 , most of the statistics computed depended on the sample size, N. Any data that is skewed to the right(Positively skewed), the mean is the highest among the three central tendencies, followed by the median with the mode being the smallest. From the table 4.1 below, the mean(2.04972)is the highest among mean,median and mode. The figures of mean, median and mode suggested that the data(BWEIGHT) is positively skewed.

Table 4.2: Summary Statistics of Birth Weight of a Baby

N	Mean	Median	Mode	Standard Dev.	Skewness	Kurtosis
1007	2.04972	1.9	1.7	0.7968505	1.177815	1.329359

Intuitively, the skewness is a measure of symmetry. When the mean of a data is larger than the median of that same data, then the value of skewness must be positive. This positive value for skewness suggest that the data from which the value was estimated is a positively skewed data. From the table 4.1, the value for skewness is positive indicating that the BWEIGHT data is a positively skewed data. The value for kurtosis(1.329359) from the table 4.1 indicates a leptokurtic(peaked distribution), since the value is positive.

In the same concept, the histogram and a boxplot of the BWEIGHT data are plotted to identify the shape of the data

4.2.2 Interpretation of the Histogram and the Boxplot

The shape of the histogram indicates that the data is skewed to the right, since the bars at the right tail is shorter than those at the left. The histogram in figure 4.1 also has a normal curve superimposed on it.This curve shows the skewness of the BWEIGHT data. From the diagram, it can be seen that the original

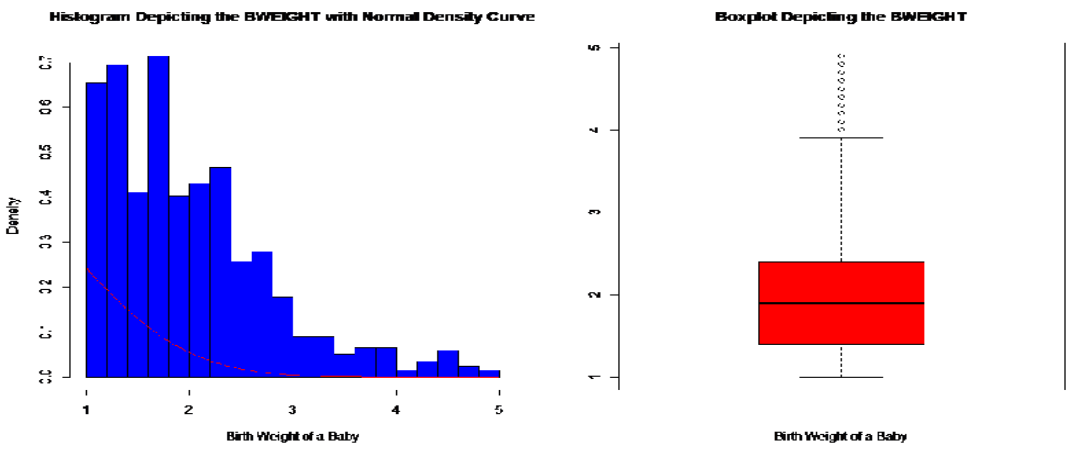

Figure 4.1: Histogram and Boxplot of the response variable(BWEIGHT)

BEWEIGHT data has a heavy right-hand tail. This means that the BWEIGHT data had few babies with very high weight while most of the babies were of low weight.

The boxplot on the other hand also depicts a right skewed data. This is because, from the figure, half of the observation (red box) fall within 1.5 – 2.5(observation between the first and the third quartile). The whisker to the right is far longer than the whisker to the left, indicating a right skewed data.

4.3 Further Results

4.3.1 Quantile Regression Analysis

Quantile regression models the relation between a set of independent variables(predictors) and a specific percentiles(quantiles) of the dependent variable(response variable). It specifies changes in the quantiles of the response. For example, the median regression(50^{th} percentile) of birth weight on mothers' characteristicsspecifiesthechangesinthemedianbirthweightasafunctionofthe predictors. The effect of mothers' age on median birth weight can be compared to its effect on other quantiles of birth weight.

In linear regression, the regression coefficient represents change in the response variable produced by a one unit increase in the predictor variable associated

with that coefficient. The quantile regression parameter estimates the change in specified quantile of a response variable produced by one unit change in the predictor variable. This allows comparing how some percentiles of the birth weight may be more affected by certain mother characteristics than other percentiles. This is reflected in the changes in size of the regression coefficient. Coefficient estimates for 5^{th}, 25^{th}, 50^{th}, 75^{th}, 95^{th} quantile regression coefficient estimates for birth weight are presented in the following table.

Table 4.3: Quantile Regression Coefficient Estimate

Variable	5th	25th	50th	75th	95th
Intercept	0.47292	0.437	1.19983	1.46745	5.27852
MOTAGE	0.00057	0.000029	-0.00131	-0.00385	-0.03194
MSTAT	-0.03028	-0.0577	-0.06615	-0.10135	-0.57572
FBAGE	-0.00127	0.000399	-0.00482	-0.00923	-0.04571
SWATER	-0.00164	0.0203	0.07274	0.04723	0.02985
NCHILD	0.00622	0.00581	-0.00149	-0.01064	-0.00703
EDULEVEL	0.00362	0.0452	0.01792	0.02840	0.20855
emstatus	-0.02349	0.0967	0.05801	0.10506	0.20964
GESAGE	0.00151	0.00184	0.00586	0.01787	0.01366
MWEIGHT	-0.00393	-0.00440	0.00314	0.00506	0.01271
MHEIGHT	0.56053	0.622	0.07938	0.06230	-0.47079
SEXB	0.02467	0.0813	0.10554	0.13641	-0.02304

The 5^{th} quantile of birth weight for babies born to mothers who had no employment is 0.02349 kilogram(from the table above) lower than babies born to mothers who had employment. However from the 25^{th} to 95^{th} quantile, the estimate rose from $9.64e-02$ to 0.20964 respectively. This indicates that babies born to mothers who had employment contribute $9.67e-02$(25^{th}), 0.05801(50^{th}), 0.10506(75^{th}) and 0.20964(95^{th}) kilograms to the weight of the baby.

The mothers' age(MOTAGE), from the table 4.3 above indicates from its coefficient that 0.00057(5^{th}) and $2.94e-05$(25^{th}) contribute to the weight of a baby. Nevertheless, from the 50^{th} to the 95^{th} quantile, the variable contributed less to the weight of the baby. The sex of the baby also had positive contribution

to the his/her weight but 0.02304 less at the 95^{th} quantile. Looking at the coefficient estimate of the variable MSTAT(marital status), it contributes negatively to the weight of the baby throughout the quantiles.

The educational level(EDULEVEL) is the only variable that contributes positively to the weight of the baby throughout the quantiles and marital status(MSTAT) is also the only variable that contributed negatively to the baby weight throughout the quantiles.

4.3.2 Model Equations

$$Q_{BWEIGHT}(0.05|X) = 0.47292 + 0.00057MONTAGE - 0.03028MSTAT - 0.00127FBAGE - 0.00164SWATER + 0.00622NCHILD + 0.00362EDULEVEL - 0.02349emstatus + 0.00151GESAGE - 0.00393MWEIGHT + 0.56053MHEIGHT + 0.02467SEXB \quad (4.1)$$

The above model measures the relationship that exist between the dependent variable(BWEIGHT) and the independent variables at the 5^{th} percentile. That is the data points within the first 5 percent of the data set. At this stage of the data, variables(MSTAT,FBAGE,SWATER,emstatus and MWEIGHT) contribute negatively to the baby weight whiles the others contribute positively. This means that when there is one unit increment in the predictor variables, the response variable(BEWEIGHT) will decrease by 0.03028, 0.00127, 0.00164, 0.02349, 0.00393 for MSTAT,FBAGE,SWATER,emstatus and MWEIGHT respectively.

$$Q_{BWEIGHT}(0.25|X) = 4.37e{-}01 + 2.94e{-}05MONTAGE - 5.77e{-}02MSTAT + 3.99e{-}04FBAGE + 2.03e{-}02SWATER + 5.81e{-}03NCHILD + 4.52e{-}02EDULEVEL + 9.67e{-}02emstatus + 1.84e{-}03GESAGE - 4.40e{-}03MWEIGHT + 6.22e{-}01MHEIGHT + 8.13e-02SEXB \quad (4.2)$$

This model also represent the relationship between the response and the predictor variables at the 25^{th} percentile. That is the first 25 percent of the data set. At this point, a unit increase in the predictor variables will cause the response variable to decrease by $5.77e-02, 4.40e-03$ for MSTAT and MWEIGHT respectively. However at this percentile only two variables contribute negatively to the weight of a baby compared with the 5^{th} percentile.

$$Q_{BWEIGHT}(0.5|X) = 1.19983 - 0.001315MONTAGE - 0.06615MSTAT - 0.00482FBAGE + 0.07272SWATER - 0.00149NCHILD + 0.01792EDULEVEL + 0.05801emstatus + 0.00586GESAGE + 0.00314MWEIGHT + 0.07938MHEIGHT + 0.10554SEXB \quad (4.3)$$

At the first 50 percent of the data, the relation above(4.3) represent the relationship between the response variable and covariates. Four predictor variables(MOTAGE,MSTAT,FBAGE and NCHILD) decrease the response variable with a unit increase in the predictors.

$$Q_{BWEIGHT}(0.75|X) = 1.46745 - 0.00385MONTAGE - 0.10135MSTAT$$

$$-0.00923FBAGE+0.04723SWATER-0.01064NCHILD+0.02840EDULEVEL$$
$$+0.10506emstatus+0.01787GESAGE+0.00506MWEIGHT+0.06230MHEIGHT$$
$$+0.13641SEXB \quad (4.4)$$

The model above(4.4) also measures the relationship of the covariates and the response variable at the 75^{th} percentile. Here the variables that decreased the response variable with a unit increase in the predictors at the 50^{th} percentile, also decrease the response variable at the 75^{th} percentile.

$$Q_{BWEIGHT}(0.95|X) = 1.46745 - 0.00385MONTAGE - 0.10135MSTAT$$

$$-0.00923FBAGE+0.04723SWATER-0.01064NCHILD+0.02840EDULEVEL$$
$$+0.10506emstatus+0.01787GESAGE+0.00506MWEIGHT+0.06230MHEIGHT$$
$$+0.13641SEXB \quad (4.5)$$

This model represent the relationship of the covariates and the response variable at the 95^{th} percentile.

4.3.3 Maximum Likelihood Estimates

Given any model, there exist a great deal of theories for making estimates of the model parameters based on the empirical data. In our case the birth weight(BWEIGHT) data was used to compute the maximum likelihood estimates of gamma and the lognormal distributions. The first step in fitting a model to BWEIGHT data is finding the parameter estimates of the particular statistical distribution. When the parameters of any distribution have been obtained using the BWEIGHT data , then literally, the statistical distribution has been fitted

to the BWEIGHT data. With regards to this work, table 4.4 below gives the parameters of the two distributions having been fitted to the BWEIGHT data. The μ was taken to be the value of the rate parameter and σ was taken to be the shape parameter value. The 4.4 table also shows the confidence interval within which the parameters lie at 5 percent level of significance. The parameters obtained were then used in the estimation of the log-likelihoods of the two distributions.

Table 4.4: Table 4.4 Estimation Results

Distribution	Log-Likelihood	AIC	μ	σ	Parameter Conf. Int.
Gamma	-1080.491	3.4656	3.6778	7.5387	$2.05 \leq \mu \leq 4.23$
Lognormal	-1060.953	2.3919	0.6499	0.3623	$0.13 \leq \mu \leq 1.14$

4.3.4 The Log-Likelihoods

The log-likelihood theory provides rigorous and omnibus inference methods if the model is given, that is, after the parameters of a distribution have been obtained. the log-likelihoods form the basis of the selection of the distribution that fits the data. It was the first tool used in the primary stage. Table 4.4 shows the computed log-likelihoods. R console was used to obtain the values. From the tabulated statistics, the lognormal distribution, with the log-likelihood value of −1060.953 has the higher log-likelihood value among the two distributions; hence the lognormal distribution was the better fit than the gamma distribution. With this, the lognormal distribution was selected as the statistical distribution that gave a relatively good fit for the BWEIGHT data as compared to the gamma distribution.

4.3.5 Goodness of Fit Test

This section is interested in the post Model selection fit to affirm the selected model. The central problem in analysis is which model to use for making inferences fro the data. The lognormal distribution emerged as having the higher

log-likelihood value than the gamma distribution but that could not have meant that it was a better statistical distribution to model the BWEIGHT data. With this argument, it was necessary to carry out a goodness of fit test in order to select the better statistical distribution that best fits the data. In this study, the goodness of fit test was done graphically and mathematically to affirm our decision. This called for plotting the Q-Q plots and computation of the Akaike's Information Criterion(AIC). Since it was necessary to ascertain how well the distribution fits the data, the AIC was estimated as presented in the table 4.4 above. Graphically the goodness of fit was established using the Q-Q plots of the two distributions as fitted on the BWEIGHT data.

Quantile-Quantile(Q-Q) Plot

The Q-Q plots for each of the two distributions were constructed using R console, andtheselectionwasbasedonacriticallookatthedatapointsandtheline y = x.

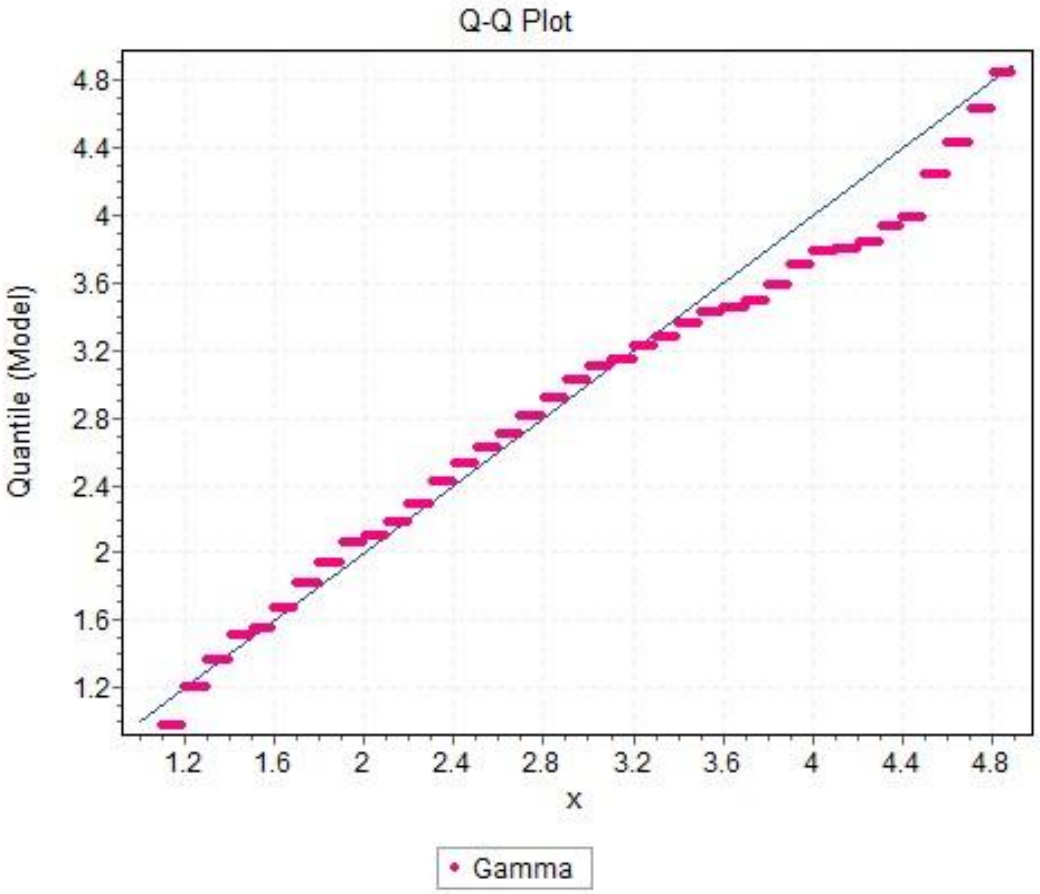

Figure 4.2: Q-Q plot for Gamma Distribution

The plot(fig:4.2) above shows the Q-Q plot of the gamma distribution. Looking at the data points and the line y = x, it is clear that the distribution is not a bad fit to the data. It is getting to the upper tail of the diagram that the points are a little bit away from the the line.

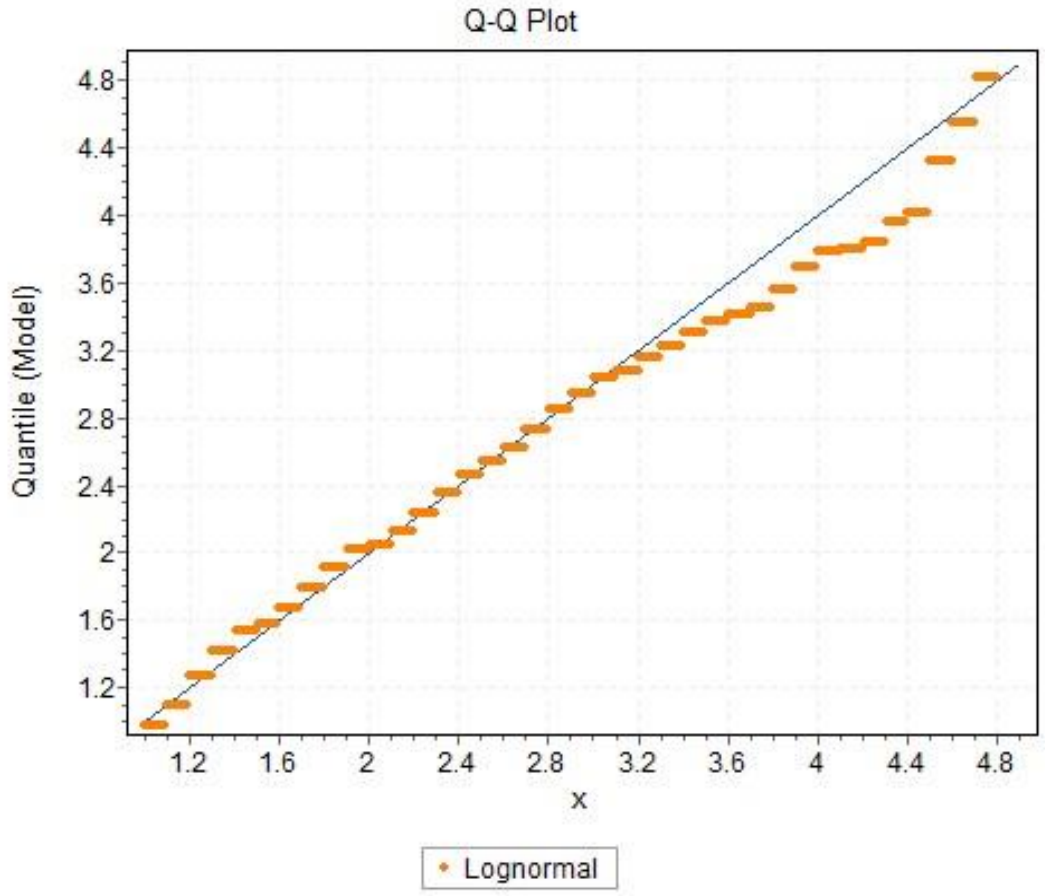

Figure 4.3: Q-Q plot for Lognormal Distribution

The plot(fig:4.3) above depicts the Q-Q plot of the lognormal distribution. The plot also shows that the line y = x passes through the data points except getting to the upper tail of the diagram, where some points are below the the line.
The two plots almost look the same especially getting to the upper tail of the diagram, but at the middle the line y = x passes through the data points well for lognormal distribution than the gamma ditribution. This affirms that the lognormal distribution fits the data better than the gamma distribution.

Density Curves

The figure below, fig(4.4), shows the density curve of the gamma distribution and fig(4.5) shows that of lognormal distribution.The curve for lognormal stretches to capture the mode of the distribution while that of gamma does not.

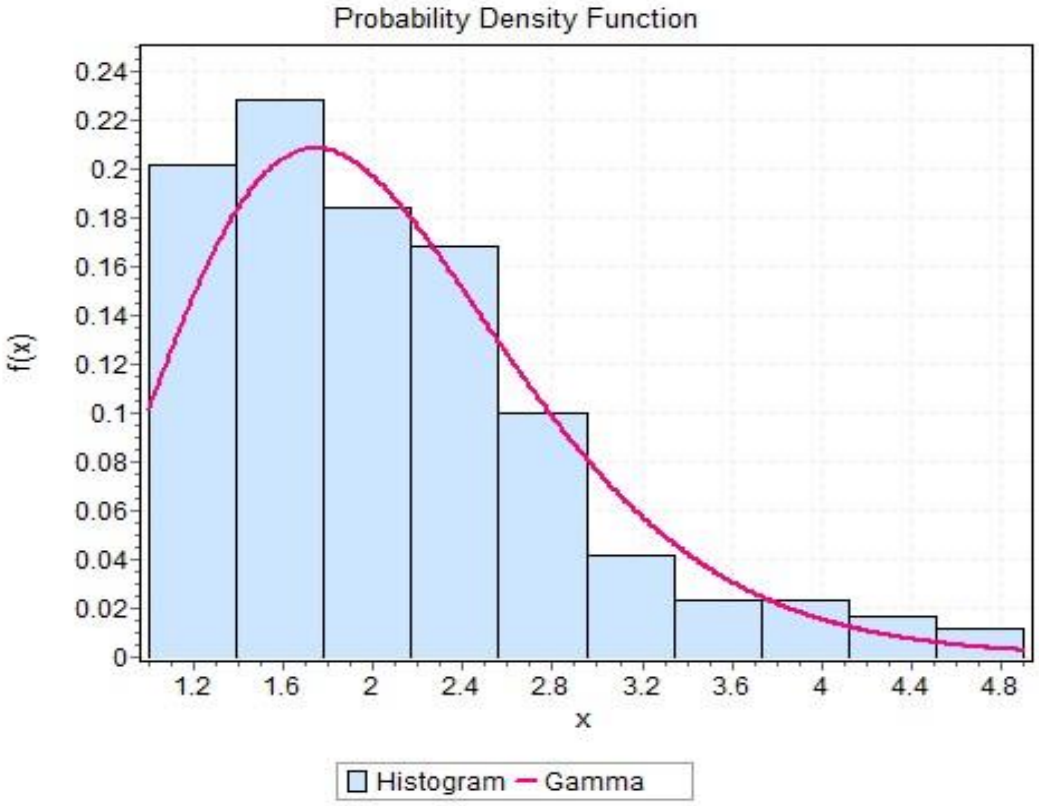

Figure 4.4: Probability Density Curve of Gamma Distribution

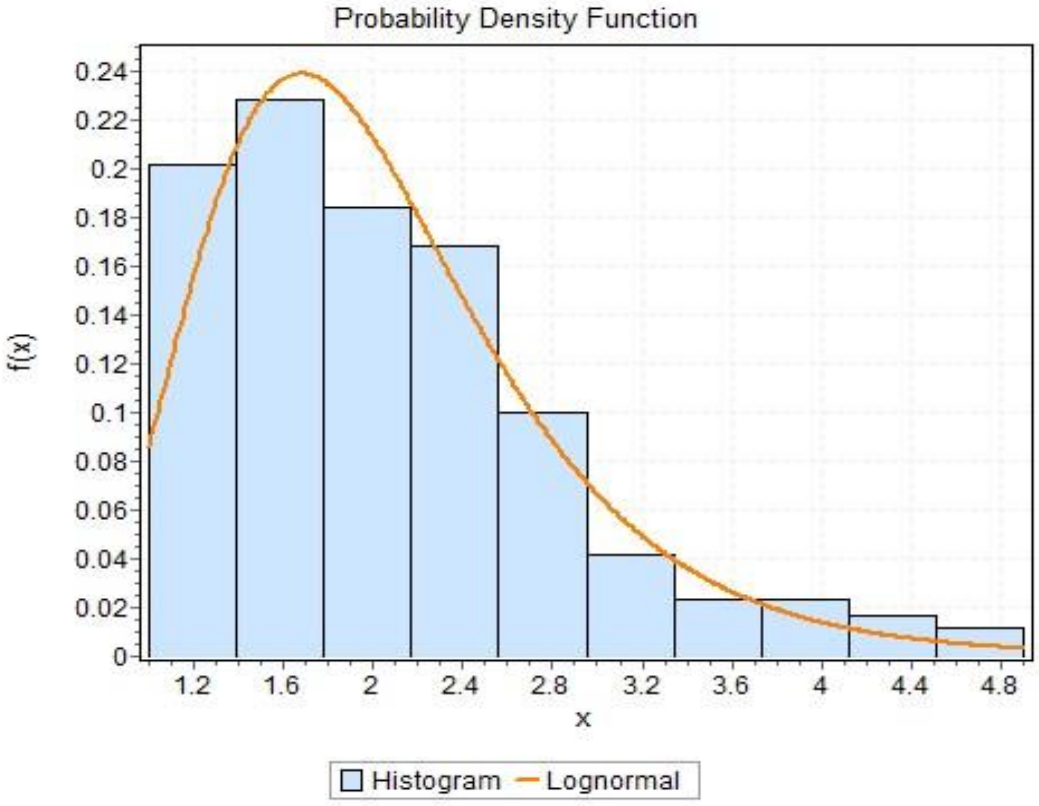

Figure 4.5: Probability Density Curve of Lognormal Distribution

4.3.6 The Akaike,s Information Criterion Interpretation

The criterion was derived based on the concept that truth is very complex and that no "true model" exists for any sampled data set. Therefore given the two statistical distributions, it was possible to estimate which distribution was close to the unknown true model.

The formulae for AIC was cited in chapter 3, AIC is used in selecting the model for making inferences for the BWEIGHT data. In this study was computed by STATA software and the results are tabulated in the table(4.4). In AIC, one should select the model with the smallest value of AIC.From table(4.4) the AIC value for lognormal is 2.3919 and that of the gamma is 3.4656. It was therefore concluded that the lognormal distribution was a better fit than the gamma distribution for the BWEIGHT data since it had the smaller value of AIC. That is, the lognormal distribution was estimated to be the closet to the unknown true distribution.

4.3.7 Gamma and Lognormal Regression Model

Table 4.5: Gamma and Lognormal Model Coefficient

Model	Lognormal	Gamma
Intercept	0.2466	0.4410
MWEIGHT	0.0005	0.0010
MOTAGE	-0.0009	-0.0010
FBAGE	-0.0027	-0.0050
SWATER	0.0111	0.0110
NCHILD	-0.0013	-0.0020
EDULEVEL	0.0160	0.0180
SEXB	0.0418	0.0390
emstatus	0.0311	0.0330
MSTAT	-0.0401	-.00460
GESAGE	0.0042	0.005
MHEIGHT	0.1469	0.0710

The table(4.5) above is the coefficient of the regressors in gamma and lognormal model. Looking critically at the coefficient of each variable, it is evident that the models are almost the same. Nevertheless, the lognormal looks better than the gamma.

4.3.8 Comparing Predictions

Themainobjectiveofthestudyistoknowthebestprocedurethatactuallyworks, not just one that has nice theory. On this data set, we can get predicted values for the quantiles of birth weight from quantile regression, gamma and lognormal model, and compare them to the actual weights. These are predicted values from the full model.

Table 4.6: Actual Versus Predicted

Quantile	Quantile Predict	Lognormal Predict	Gamma Predict	Actual
5th	1.32	1.28	1.24	1.30
25th	1.87	1.85	1.82	1.90
50th	2.41	2.34	2.33	2.40
75th	3.19	3.12	3.05	3.20
95th	4.13	4.01	3.98	4.10

The table(4.6) above represent the actual values of the birht weight against the predicted values. The quantile regression from the table has the best prediction. The lognormal and gamma predicted values look close, but the lognormal is slightly better than the gamma. We therefore conclude that the quantile regression is best regression analysis for our BWEGHT data. On the other hand, although lognormal and gamma models are among the most popular models used in the analysis of positively skewed data, the lognormal is better than the gamma with respect to our data.

CHAPTER 5

CONCLUSION

5.1 Introduction

This chapter entails the various conclusions drawn from the analysis in chapter four with regards to the various regression models. Recommendations about the whole study has also been summarized in this chapter.

5.2 Summary of Results

The focus of this study was to find the relationship that exist between the covariates and the response variable and come up with one statistical distribution(between lognormal and gamma) for the birth weight data and to test how well this statistical distribution fits the birth weight so that this distribution can be used for modeling the birth weight. In a very important sense, the study was not concern with the steps of modeling the data; instead, the study tries to model the information in the data to fit a particular distribution.

Therefore, an attempt was made to establish an appropriate statistical distributionthatbestfitsthebirthweightdatausingnumericalcomputationsand graphical implications using R software. From the analysis carried out coupled with the results displayed in table 4.4 it is revealed that the birth weight data for KATH can best be modeled using the log-normal distribution.

According to table 4.4, the log-normal distribution has the higher log-likelihood value of -1060.953, which implies that between lognormal and gamma statistical distributions, it stands a better chance in providing a good fit for the birth weight data. In figure 4.4 and 4.5 the P.d.f of the two distributions have been plotted

in comparison with the plot of the birth weight data. This figures illustrate that the log-normal distribution's shape matches the shape of the birth weight indicating that it can actually be used to model the birth weight data. To test whether the log-normal distribution provides a good fit to the birth weight data, the A.I.C is computed in the third column of table 4.4. With regard to the A.I.C, the lognormal distribution has the least A.I.C value of 2.3919 indicating that it provides a good fit for the birth weight data.

Q-Q plots for each of the two statistical distributions was plotted on figure 4.2 to figure 4.3 to graphically re-affirm the goodness of fit test computed by the A.I.C. Figure 4.3 shows that the Q-Q plot of the log-normal distribution provides better fit to the birth weight data as most points plot on the reference line and only a few points plot deviate from the line $y = x$.

Finally, table 4.6 shows the predicted values against its corresponding actual observations with the quantile, lognormal and gamma models. It was clear that the quantile regression gives the best predicted values, but lognormal was also better than gamma predicted values.

5.3 Conclusions

After carrying out each step in the regression modelling processes with diligence and accuracy, the study clearly indicates that the Lognormal distribution would provide a better fit to the BWEIGHT data than the Gamma distribution. The quantile regression also provided the stochastic relationship that exist between the covariates and the response variables at each quantile(percentile) and the best predicted values among the regression used in our study. Therefore if any researcher wants to model the birth weight of a child, the appropriate statistical distribution to use (between gamma and lognormal) to yield a reliable birth weight forcasts would be the lognrmal distribution.

Having tested the goodness of fit of the lognormal distribution both graphically using the Q-Q plots and mathematically using the A.I.C value, it is evident

that the steps followed in the modeling process is capable of yielding reliable results that can be used to make inferences useful for decision making in the general positively skewed data. This study has shown that the assumptions made before the analysis of birth weight data may greatly affect the final results as the assumptions made led to the choice a family of distribution consisting of the log-normal and gamma. Of which the log-normal distribution was proved to be capable of providing a good fit for the birth weight. From the modeling carried out using the R console, one can conclude that more right hand tailed statistical distributions would have been included in the study so as to increase the sample distributions used in the study for accuracy of findings. However the results of this study are dependent on a number of factors outside the modeling process. This means that the reseachers have to acknowledge these factors before using the results of this study in making future inferences. These factors are the factor that account for the birth weight of a child, that is, the independent variables and the scope of the data.

The future forecast of the birth weight data may change if some of the factors talked about changes in future. Despite the dependence of the results of this study, the analysis has yielded results which can be used to amend the problems faced by the hospitals and the pregnant women. In conclusion, the modeling process is an important step before any decision can be made with regard to future policies in the area that generate positively skewed data(that is , health, insurance, economics, education etc), therefore more effort must be dedicated to ensure that the process adopted yields accurate and reliable forecast.

5.4 Recommendations

For further studies, the following directions may be considered:

Include other relevant variables which are not considered in the study such as parental care of mothers during pregnancy and other nutritional intakes. Possible interaction effects among covariates of birth weight may also be taken into

consideration. Moreover, exploration of other model selection procedures (aside from stepwise selection) may help in understanding the relationships between variables considered.

Use a data set with larger sample size, especially at the extreme quantiles to allow stability of results. It is highly encouraged to use panel data to possibly control for some exogenous maternal characteristics (i.e. genetics, pregnancy history, etc.).

Do parallel studies for both the 2014 and the upcoming 2015 Birth Weight Records to compare results and look at the effectiveness of certain government policies regarding the improvement of maternal health through time.

Explore other statistical information from the results obtained from quantile regression,lognormal and gamma regression which are currently being studied in a wide variety of literature.

Thisinformationmaypertainnotonlytothetruedistributionofbirthweight(e.g. scale shift and skewness shift) but also to the relationships among its covariates (e.g. R-squared).

REFERENCES

Abrevaya, J., 2001, The effects of demographics and maternal behavior on the distribution of birth outcomes, *Empirical Economics* 26: 247-257

Aitchison, J. and Brown, J.A.C. (1957). " The Lognormal Distribution", Cambridge University Press, Cambridge, UK.

Allen, L. H., and Gillespie, S. R., 2001, What Works? A Review of the Efficacy and Effectiveness of Nutrition Interventions. *Geneva in collaboration with the Asian Development Bank, Manila: United Nations Administrative Committee on Coordination Subcommittee on Nutrition (ACC/SCN)*.

Anderson, D.R. and Burnham, K.P. (2004) "Multimodal Inference": Understanding AIC and BIC in Model Selection. *Social Methods* Res 33: 261-304.

Andreas B., Tosehke A.M., Fahrmeir L., Mansmann U. (2008). "Alternation regression model to assess increase in childhood BMI". *BMC. Med Res Methodol* 8:59.

Bandura, A. (2010). Self-efficacy. *In Corsini Encyclopedia of Psychology*. Wiley Online Library.

Barker, D. J. (2006). Adult consequences of fetal growth restriction. *Clinical obstetrics and gynecology* 49, 270–283.

Barker, D., 1997, Maternal nutrition, fetal nutrition, and disease in later life, Nutrition 3:807-813.

Basu S, Rathore P, Bhatia B.D. Predictors of mortality in very low birth weight neonates in children.

Behrman, R. and Butler, A. (2007). Preterm Birth: Causes, Consequences, and Prevention. National Academies Press, Washington, DC.

Benjamin, J.R. and Cornell, C.A. (1970) "Probability, Statistics and Decision for Civil Engineering", McGraw Hill, New York, NY.

Boan Health, "Chinese Manufacturer of Dietary Nutrition Supplement. www.boanhealth.com.

Bondell, H. D., Reich, B. J., and Wang, H. (2010). Noncrossing quantile regression curve

Box GEP, Cox DR. An analysis of transformations. J R *Stat Soc Series B Stat Methodol* 1964; 26: 211-252.

Buchinsky, M., 1998, Recent advances in quantile regression models: A practical guide for empirical research, *Journal of Human Resources* 33: 88-126

Chase H.C. Infant mortality and weight at birth: 1960 United States birth cohort. *Am J Public*

Health Nations Health 1969; 59:1618-1628.

Chase H.C. International comparison of perinatal and infant mortality: the United States and six west European countries. Washington, DC: U.S. Department of Health, Education, and Welfare, Public Health Service; 1967.

Cheung, Y.B., Yip, P. Karlberg, J., (2000) "Motility of Twins and Single-ton by Gestational Age, a Varying Coefficient Approach*". Am J Epidomiol*; pp. 1107-1116.

Cole C, Binney G, Casey P, Fiascone J, Hagadorn J, Kim C, et al. Criteria for determining disability in infants and children: low birth weight. *Evid Rep Technol Assess (Summ)* 2002; 70: 1-7.

Collins J.W, David R.J. The different effect of traditional risk factors on infant birthweight among blacks and whites in Chicago. *Am J Public Health 1990*; 80: 679-681.

Corman, H., and Chaikind, S. (1998), "The effect of low birth weight on the school performance and behavior of school-aged children*". Economics of Education Review*, 17: 307-316.

Crow, E.L. and Shinizu, K., Eds. (1998) "Lognormal Distribution, Theory and Application", Marcel Dekker, New York, NY.

Das R.N, Lee Y. Log normal versus gamma models for analyzing data from quality-improvement experiments. *Qual Eng* 2009; 21(1):79-87.

Das RN, Park JS. Discrepancy in regression estimates between Log-normal and Gamma: Some case studies. *J Appl Stat* 2012; 39(1): 97-111.

Das, R.N. and Park, J.S. (2014), "A Reinforced Randomized Block Design with Correlated Errors", Cummun, statist. *Theo Method* 43(1), pp. 191-209.

David R.J, Collins J.W. Differing birth weight among infants of U.S.-born blacks, African-born blacks, and U.S.-born whites. *N Eng J Med* 1997; 337: 209-214.

Davis, H.T., (1933), "Tables of Higher Mathematical Functions", Bloominton, IN: Principia Press.

Deininger, K., and Squire, L., (1996) "A New Data Set Measuring Income Inequality", *World Bank Economic Review*, vol. 10.

Firth D., (1988),Multiplicative errors: log-normal or gamma? *J R Stat Soc Series B Stat Methodol* ; 50:266-268.

Flegal, K.M., Graubard, B.I., Wilhamson, D.F. and Gail, M.H. (2005) "Excess Deaths Associated with Underweight, Overweight and Obesity" *Journal of the American Medical Association*, 293(15);1861-1867.

Francesca, D., Wendy, L., and Helen, R., (2007) "Editorial Academy of Marketing conference 2007", Marketing Theory into Practice Hosted by Kingston Business School, *Journal of Marketing* 23(5-6), pp. 387-393.

Fiete, S. (2005) COTOR challenge Ronnal 3, available at www.casact.org/cotor/Fiete.doc.

Fisher, R. A., (1941) "Average Excess and Average Effect of a Gene substitution" Annals of Eugenics.

Gardosi, J., Mongelli, M., Wilcox, M., and Chang, A. (1995). An adjustable fetal weight standard. *Ultrasound in Obstetrics & Gynecology* 6, 168–174.

Garite, T. J., Clark, R., and Thorp, J. A. (2004). Intrauterine growth restriction increases morbidity and mortality among premature neonates. *American journal of obstetrics and gynecology* 191, 481–487.

Gribble, O.M. (2013), "Body Composition and Arsenic Metabolism": A Cross-sectional Analysis in the Strong Heart Study, Environmental Health 12:107.

Gustavsson, A., Yuan, M., Fallmar. M. (2004) "Temporal Dissection of beta1-integrin Signaling a role for P130Cas-Crk in Fikopodia Formation". *J Biol Chem*; 272(22); pp. 22893 22901.

Hattis, D., and Burmaster D.E., (1994) "Assessment of Variability and uncertainty distribtion for Practical risk analysis. *Risk Analysis* 14: 713-730

Hogg, R. and Klugman, S. (1984), "Loss Distribution", Wiley, New York.

Hosmer R, Lemeshow J. Applied logistic regression. 2nd ed. New York: John Wiley & Sons Inc; 2000. India. *Singapore Med J* 2008; 49(7): 556-560.

Kenneth, B.,(2011) Linear Regression Model with Logarithm Transformation, Methodology Institude, London School of Economics, pp. 1-8.

Koenker, R. (2005). Quantile regression. Number 38. Cambridge university press.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives* 15, 143–156.

Koenker, R., 2005, Quantile Regression, Cambridge: Cambridge University Press.

Koenker, R., and Bassett, G., 1978, Regression quantiles. Econometrica 46: 33-50.

Kramer M.S. Determinants of low birth weight: methodological assessment and meta-analysis. Bull World Health Organ 1987; 65(5): 663-737.

Kumar P, Seshadri R. Neonatal morbidity and growth in very low birth-weight infants after multiple courses of Antenatal Steroids. *J Perinatol* 2005; 25: 698-702.

Lane, F.B. and Jerome, P.R. "Moelling Adverse Birth Outcomes via Confirmatory Factor Quantile Regression", Department of Statistical Science, Duke University, Durham NC 22708, USA.

Lavado, R. F., Lagrada, L. P., ULEP, V. T., and Tan, L. M., 2010, Who Provides Good Quality Prenatal Care in the Philippines, Makati: Philippine Institute for Development Studies.

Lewit, E. M., Barker, L. S., Corman, H., and Shiono, P. H., 1995,The direct costs of low birth weight, The Future of Children 5: 35-51.

Lorenz, J., Wooliever, D., Jetton, J., and Paneth, N. (1998). A quantitative review of mortality and developmental disability in extremely premature newborns. *Archives of Pediatrics and Adolescent Medicine* 152, 425–435.

Machado, J. A. F. and Silva, J. S. (2005). Quantiles for counts. *Journal of the American Statistical Association* 100, 1226–1237.

Marazzi, S., Blum, S., Hartman, R., Gunderson, D., Schreyer, M., Argraves, S., von Fliedner, V., Pytela, R., Ruegg, C. (1998), "Charaterization of Human Fibroleukin, a fibrinogen-like protein

ecreted by T lymphocytes, *J-Immun*.161:138-147.

McCullagh P, Nelder J.A. Generalized linear models. London: Chapman & Hall; 1989.

MedlinePlus, "USA Library of Medicine", www.medlineplus.com

Myers R.H, Montgomery D.C, Vining G.G. Generalized linear models with applications in engineering and the sciences. New York: John Wiley & Sons; 2002.

Myers, R.H., and Montgomery, D.C. (2002) "Response Surface Methodology", John Willey and Sons, New York.

Narchi, H., Skinner, A., and Williams, B. (2010). Small for gestational age neonates-are we missing some by only using standard population growth standards and does it matter? *Journal of Maternal-Fetal and Neonatal Medicine* 23, 48–54.

Norlund, N.E., (1924). "Volesungen uber Differenzenrechnung, Springer-verlag, Berlin.

Ott, W., (1990) "A physical explanation of the lognormality of pollutant concentrations-*Journal of Air and Waste management Association.* 40, 1378-1383

Ott, W., (1995). "Environmental Statistics and Data Analysis". Lewis, Boca Raton.

Puffer P.R, Serrano C.V. Patterns of mortality in childhood: report of the Inter-American

investigation of mortality in childhood. Washington, DC: Pan American Health Organization; 1973. 11.

Reich, B. J. and Smith, L. B. (2013). Bayesian quantile regression for censored data. Biometrics

Reich, B. J., Cooley, D., Foley, K. M., Napelenok, S., and Shaby, B. A. (2011). Extreme value analysis for evaluating ozone control strategies.

Reich, B. J., Fuentes, M., and Dunson, D. B. (2011). Bayesian spatial quantile regression. *Journal of the American Statistical Association* 106,.

Reolalas, A.T., and Novilla, M.M., (2010), "Newborn deaths in the Philippines", *11th National connection on Statistics (NCS)*, 6-7.

Rich-Edwards J.W, Buka S.L, Brennan R.T, Earls F. Diverging associations of maternal age with low birthweight for black and white mothers. *Int J Epidemiol* 2003; 32: 83-90.

Saugstad, L.F. Weight of all births and infant mortality. *Epidemiol Community Health* 1981;35:185-191.

Shah, N.R. and Bracken, M.D. (2000), "A Systematic Review and Meta-Analysis of Prospective Studies on the Association Between Maternal Cigarette Smoking and Preterm Delivery", *Am J Obstet Gynecol;* 182(2): pp. 465-472.

Sr´am, R. J., Binkov´a, B., Dejmek, J., and Bobak, M. (2005). Ambient air pollution and pregnancy outcomes: a review of the literature. *Environmental health perspectives* 113, 375.

Tabatabai, M.A. (1996): Sulfer In: Methods of soil analysis-Part3-Chemical Methods (Ed: Sparks D.L.) Madisson, Wilconsin, USA: *Soil Science Society of America, American Society of Agronomy*. Pp. 992-960.

Tokdar, S. and Kadane, J. B. (2011). Simultaneous linear quantile regression: A semi-parametric bayesian approach. Bayesian Analysis 6, 1–22.

Toschke, A.M., Kuchenhoff, H., Koletzko, B., Kries, R., (2005) Meal Frequency and Childhood Obesity; *Obes Res*:13;1932-8

Victiria-Feser, M.P. (2000), "A General Robust Approach to the Analysis of Income Distribution, Inequality and Poverty", *International Statistical Review* 68: 277-293.

Victoria-Feser, M.P. and Ronchetti, E., "Robust Estimation of Group Data", *Journal of American Statistics Association* 92(437): 333-340.

Victoria-Feser, M.P., and Ronchetti, E. (1994) "Robust Methods for Personal Income

Distribution Methods", *Cannadian Journal of Statistics* 22(2):247-258.

Wang, H. and Tsai, C.-L. (2009). Tail index regression. *Journal of the American Statistical Association* 104, 1233–1240.

Wang, H. J., Li, D., and He, X. (2012). Estimation of high conditional quantiles for heavy-tailed distributions. *Journal of the American Statistical Association* 107, 1453–1464.

World Health ORGANIZATION. World Health Statistics 2012. Geneva: WHO Press, 2012.

Yu, K. and Moyeed, R. A. (2001). Bayesian quantile regression. Statistics and Probability. Letters 54, 437 – 447.

Yurgens, Y. (2004), "Quantifying Environmental Inequality by Lognormal Regression Modelling of Accumulated Exposure". Chalmers University of Technology and Goteborg University; Goteborg, Sweden.

Zhou, J., Chang, H. H., and Fuentes, M. (2012). Estimating the health impact of climate change with calibrated climate model output. *Journal of agricultural, biological, and environmental statistics* 17, 377–394.

APPENDIX A

This is APPENDIX A

APPENDIX B

This is APPENDIX B