ROBUST REGRESSION AND MODEL DIAGNOSTICS: AN APPLICATION TO MATERNAL HAEMOGLOBIN DATA IN MALAWI

Master of Science (BIOSTATISTICS) THESIS

POTIPHAR MOSES DAMIANO

Bachelor of Science (Statistics and Mathematics) - University of Malawi

Submitted to the Department of Mathematical Sciences, Faculty of Science, in partial fulfilment of the requirements for the degree of Master of Science in Biostatistics

University of Malawi

DECLARATION

I, POTIPHAR MOSES DAMIANO hereby declare that this thesis/dissertation is my own original work and has not been presented/submitted to any other institution for a similar or any other degree award. Where somebody's work or results have been used, proper acknowledgements have been made.

_____

Signature

_____

Date

CERTIFICATE OF APPROVAL

The undersigned certifies that this thesis represents the student's own work and effort and it has been submitted with my approval.

Signature: _____ Date: _____

Tsirizani Kaombe, PhD
Supervisor

DEDICATION

This thesis is dedicated to my loving wife, Thandizo; my mother, Mrs. Moses and late Dad, Mr. Moses. May the spirit of Dad rest in eternal peace. Thank you all for everything! You are the best family in whole world. I love you with all my heart. May God bless you always.

ACKNOWLEDGEMENTS

First of all, I thank Lord Jesus Christ for the wonderful Blessings HE has given me: wisdom, intelligence, courage, peace, patience, hard work and strength in difficult situation throughout my study. I thank Him for HIS graces and love.

Very special thanks to wife (Thandizo Ngulinga) and kids (Hadassah and Comfort) for moral, physical and spiritual support, for unconditional love and care you have always given me. Special thanks to my mother, for your love, care and support, you really make me smile. Am always thankful for you, God bless you all.

I extend my gratitude to supervisor, Dr Tsirizani Kaombe for his constant support, constructive ideas, unconditional love, guidance and continuous encouragement throughout my Thesis work. Your comments that filled my drafts have really helped; this paper could not have come out like this without your help. God bless you abundantly.

I further extend my gratitude to Biostatistics classmates and workmates, you helped a lot in my academic life.

Abstract

The data on maternal anaemia is highly skewed in sub-Saharan Africa, with some women showing higher and others lower levels of Haemoglobin (Hb). A thorough analysis of maternal anaemia data is crucial for identifying effective strategies, but success depends on the choice of model and its ability to handle outliers. The study evaluated mean, quantile, and robust regression methods, along with diagnostic statistics, on maternal Hb data in Malawi. The analysis used simulations and real Hb data from the 2015-16 Malawi Demographic and Health Survey, calculated with STATA version 17. The simulation results revealed that in large sample sizes, outlier detection rates were similar across linear, quantile, and robust regression models. Further, all models showed similar accuracy without outliers. For datasets with outliers, robust and quantile regression (1st and 2nd quartiles) provided the most accurate estimates with smaller biases compared to linear and higher percentile models. The real data analysis showed that directions of estimates were similar across the models, but the linear, robust M- and MM-estimator models produced estimates with smallest standard errors. The estimated average Hb level for women was 13.7 g/dl. Residing in rural area, higher body mass index, having primary and secondary education were linked to high Hb levels. While older pregnancy, drinking from safe water sources, and living in a rich household were associated with low Hb levels. The model residuals detected considerable amount of outliers in the data, mostly they were women with extremely low Hb levels. Diverse statistical methods can strengthen evidence of maternal anaemia in sub-Saharan Africa, supporting the determination of effective interventions. Policymakers in Malawi should develop strategies to increase Hb levels in pregnant women, especially in their second and third trimesters, and other marginalized groups.

Contents

CHAPTER ONE

INTRODUCTION

1.1 Background to Maternal Anaemia

Anaemia is a blood disorder characterized by low concentration of haemoglobin, the protein responsible for carrying oxygen in red blood cells (Di Renzo et al., 2015; Meena et al., 2019). It is a significant public health problem in many developing countries, affecting about 571 million women of reproductive age and 32 million pregnant women worldwide (Stevens et al., 2022; Kassebaum et al., 2016; Chaparro & Suchdev, 2019; Pasricha & Moir-Meyer, 2023). The World Health Organization (WHO)-defined haemoglobin (Hb) cut-offs, specific to age, sex and pregnancy status, are most widely used to diagnose anemia (Ohuma et al., 2023). For example, the World Health Organisation (WHO) considers as anaemic the Hb level of less than 12 grams/decilitre in non-pregnant women aged 15-49 years, 11 grams/decilitre during the first and third trimesters of pregnancy, and 10.5 grams/decilitre in the second trimester of pregnancy (Kamruzzaman, 2021; Alem et al., 2023; Young et al., 2023).

Anaemia is often categorised based on its cause. Inadequate consumption of micronutrients, such as iron, folate, riboflavin, and vitamins A, B12, and C necessary for blood formation, is a common cause of nutritional anaemia (Ali et al., 2023; Shi et al., 2021). A condition known as nutritional iron deficiency (ID) is brought by inadequate dietary iron intake, increased iron demand, iron loss, and low iron bioavailability from staple foods. In global context, ID is considered the major contributor to the burden of WRA anemia (Ali et al., 2023). Additional factors contributing to anaemia include heavy men-

struation, growing children's and pregnant women's higher iron requiremnets. Furthermore, other factors are chronic infections (including HIV, TB, hookworm, and malaria), and disorders affecting the body's ability to absorb, transport, and store iron, such as haemoglobinopathies (Karami et al., 2022; Chaparro & Suchdev, 2019). About half of all female anemaia are caused by ID, which also serves as a precursor to iron deficiency anaemia (IDA), one of the main causes of Years Lived in a Disabled state (YLDs) (Chaparro & Suchdev, 2019). Anemia caused 52.0 million YLDs in 2021 (Kinyoki et al., 2021) and contributed to 58.6 million YLDs worldwide in 2019 (Kamruzzaman, 2021).

Globally, it is estimated that above a half billion Women of Reproductive Age (WRA) are anaemic representing about 33 percent of maternal women (Stevens et al., 2022; Pasricha & Moir-Meyer, 2023). The highest prevalence of maternal anaemia is in Lower and Middle-Income Countries (LMICs) with WRA and children having higher risk than adults (Kinyoki et al., 2021; Safiri et al., 2021; Pasricha & Moir-Meyer, 2023; Hasan et al., 2022; Moya et al., 2022). Implying that about one third of the women aged 15 to 49 years are anaemic worldwide.

Globally, West and Central Africa, and South Asia are the three regions that contribute the most to anaemia, affecting about 40 percent of maternal women (Alem et al., 2023). In a global context, there has been marginal progress on reduction of anaemia prevalence among WRA especially in LMICs (Chaparro & Suchdev, 2019). Maternal anaemia prevalence remained almost constant, from 31 percent in 2000 to 30 percent in 2019 (Young et al., 2023; Stevens et al., 2022). And according to Karami et al. (2022); Pasricha & Moir-Meyer (2023), in 2021, the prevalence of anaemia among WRA was 33.7 percent, compared to 11.3 percent for males. According to Chanimbe et al. (2023), in Malawi

the prevalence of maternal anaemia is at 29.8 percent. Implying that about one-third of Malawian maternal women are anaemic.

It is well known that anemia during pregnancy increases the risks of having miscarriages, intrauterine growth retardation, preterm births, still birth, babies with Low Birthweight (LBW), neonatal and maternal mortality (Ali et al., 2023; Young, 2018). In developing nations, anaemia is a major driver in maternal mortality and adverse pregnancy outcomes (Ali et al., 2023). According to published research, there is a direct correlation between anaemia and maternal mortality, with every 10 millilitre rise in haemoglobin causing a 30 percent reduction in maternal deaths (Ali et al., 2023; Black et al., 2013; Young, 2018)

Nutrition dificiency especially iron deficiency is recognized as crucial risk factor for anemia among WRA (Ali et al., 2023). There is high risk of iron deficiency due to pregnancy as the iron requirement triples due to the growth of the fetoplacental units and the increase in the number of maternal red blood cells (Shi et al., 2021; Ali et al., 2023). Iron deficiency accounts for at least 60 percent of anemia (Kassebaum et al., 2016). With an aim to control the effect of anemia, in 2020 WHO proposed distribution of iron suppliments to all WRA in regions with prevalence of above 20 percent (Ali et al., 2023). Because of this, pregnant women in the majority of low- and middle-income nations frequently take iron supplements to prevent and treat iron deficiency and anaemia during pregnancy. The current global control initiatives for maternal anaemia include provision of iron and folic acid supplements to women in regions having anaemia prevalence of above 20 percent (Ali et al., 2023).

Despite implementation of interventions aimed at reducing maternal anemia, burden of maternal anemia is still high in sub-saharan Africa, at 41.7 percent, which derails safe

motherhood campaign efforts in the region (Karami et al., 2022; Chaparro & Suchdev, 2019; Kassebaum et al., 2016). Effectively addressing anaemia in all its forms requires a firm understanding of the unique determinants of anaemia in a particular setting, including by subnational area (Kinyoki et al., 2021). Since anemia is significantly associated with morbidity and mortality, programs, strategies, and interventions targeted at lowering WRA anaemia have the potential to improve the general health outcomes of children as well as WRA. Therefore, studies that can bring in evidence on the determinants and drivers of WRA anemia reduction in LMICs can be supportive for development of appropriate interventions.

1.2 Statistical methods used to analyse maternal anaemia data

There are a number of studies in literature that employed regression analysis to determine the factors associated with maternal anaemia. For example, a study done by Alem et al. (2023) analysed data from the Demographic and Health Survey (DHS) in 46 LMICs during period of 2010 to 2021. The study involved 881,148 WRA with an aim of assessing the prevalence and factors associated with anaemia among WRA in LMICs. The proportions between pregnant and non-pregnant women were assessed using descriptive statistics. In order to determine the factors associated with anaemia in WRA, multilevel binary logistic regression was used.

The results from Alem et al. (2023) study found a high prevalence of 45.20 percent of anemia among pregnant women and 39.52 percent prevalence among non-pregnant women in LMICs. The study reported that these estimates were higher and far from the global target (less than or equal to 15.2 percent by 2025), comparable with previous studies (Kinyoki et al., 2021; Sun et al., 2021; Owais et al., 2021). The results further showed that Edu-

cation status, wealth status, family size, media exposure and residence were significantly factors associated with anaemia in both pregnant and non-pregnant women. The identified factors are similar to previous studies, for example a study in Etheopia (Geta et al., 2022), Pakistan (Ullah et al., 2019) and Nepal (Acharya et al., 2022). The study recommended global commitment and movement to reduce the prevalence of anaemia need to be revisited and redesigned for current circumstance.

Another study done by Sunuwar et al. (2020) analyzed DHS data between 2011 and 2016 from seven sampled Southern and Southeastern Asia countries (Bangladesh, Cambodia, India, Maldives, Myannar, Nepal and Timor-Leste). A total of 726,164 WRA were involved in this study with a purpose of identifying prevalence and factors associated with anaemia among WRA in seven selected South and Southeast Asian countries. Descriptive statistics of proportions among WRA were used to etimate prevalence. Multiple linear regression models were performed to identify the factors significantly associated with anaemia. The study reported multicollinearity among independent variables using variable inflation factors in order to prevent statistical bias.

The study by Sunuwar et al. (2020) reported overall WRA anemia prevalence of 52.5 percent, ranging from 22.7 percent in Timor-Leste to 63 percent in Maldives. Results from multiple logistics regression showed that age group, education status, wealth status, toilet type, water source, BMI and births in last five years are significant factors associated with anaemia. It suggested that young women (15-24 years), those with primary or no education, poorest wealth, without toilet facilities, not improved water source, underweight and with more than one child in last five years have significantly higher likelihood of anaemia.

Teshale et al. (2020) studied 101,524 WRA using DHS data conducted between 2008 and

2018 in ten eastern African countries with an aim of assessing prevalence and associated factors of anaemia among WRA in eastern Africa. The ten Eastern African countries involved in the study were Burundi, Ethiopia, Malawi, Mozambique, Rwanda, Tanzania, Uganda, Zimbabwe, Madagascar and Zambia. Descriptive statatics of proportions were used to report the unadjusted and adjusted prevalence of anaemia. Multilevel mixed-effects generalized linear model, using Poisson regression, was used to identify factors significantly associated with anaemia.

This study by Teshale et al. (2020) reported WRA anaemia prevalence of 34.85 percent in eastern Africa ranging from 19.23 percent in Rwanda to 53.98 percent in Mozambique. Multivariable level analysis showed that age, education, marital status, occupation, household wealth status, sex of household head, type of toilet facility, source of drinking water, ever had a terminated pregnancy, parity, household size, perception of distance from the health facility, pregnancy status and residence were significant determinants of anaemia among WRA. The results were consistent with other previous studies (Adamu et al., 2017; Soofi et al., 2017). The study recommended that with special attention on younger women, those with low socioeconomic status, unimproved toilet facility, unimproved drinking water source and pregnant women could reduce burden of anemia in WRA.

Talukder et al. (2022) analyzed DHS data for the period of 2017 to 2018 collected from Albania country located in Southern Europe with an aim of identifying the potential risk factors of anaemia among Albanian WRA. A total of 15,000 WRA were involved in the study. The study employed a quantile regression model to identify the determinants of anaemia.

The results from Talukder et al. (2022) study showed that women's education level, wealth

index, place of residence, contraceptive method use during pregnancy, BMI, and source of drinking water are the significant risk factors of anaemia among WRA. The results agreed with results from previous studies (Adamu et al., 2017; Shi et al., 2021)It was recommended from the study that effective strategies aiming at preventing and controlling anemia should focus on women living in the rural areas, underweight, not higher educated, not using contraceptives during pregnancy and drinking unsafe water.

A study conducted by Acharya et al. (2022) analysed Nepal country DHS data for 2006, 2011 and 2016 which involved a total of 23,149 WRA with an aim to assess trends of anaemia prevalence and determinants of anemia among WRA. Descriptive (frequencies and percentages), bivariate (cross-section with chi-square test), and multivariate analysis (binary logistics regression) were performed to address the study purpose. The results showed an inconsistent trends of anaemia prevalence among the survey years, with 36 percent in 2006, 35 percent in 2011 and 41 percent in 2016.

According to Acharya et al. (2022), age of women, place of living, wealth status, smoking habit, exposure to radio are significant predictors for having anaemia. The results were consistent with previous study (Teshale et al., 2020). The study recommended that the policymakers should re-evaluate and revise existing strategies of combating anemia as these seemed to be ineffective in reducing prevalence of anaemia.

1.3 Robust Regression Methods

George E.P. Box, a statistician, introduced the term "robustness" where robust techniques are those that are insensitive to the departures from the underlying assumptions (Grynovicki et al., 1983). Robust regression method is a technique used to analyze data

that is contaminated with outliers and minimize their impact on the coefficient estimates (Bary, 2017; Kalina, 2015; Ayinde et al., 2015; Ritschard & Antille, 1992; Denby & Mallows, 1977).

Ordinary least squares (OLS) and maximum likelihood (ML), widely used methods of estimating linear regression parameters, base their predictions on the assumptions such as normality and constant variance σ^2 of the response variable on the regression structure (Jajo & Hussain, 1989). Therefore, OLS and ML have the property of providing 'best' unbiased estimators when the error has a Gaussian distribution. However, it is recognized that outliers may have an unusually large influence on the OLS and ML estimators, outliers may push the line of best 'fit' too much in their direction (Jajo & Hussain, 1989; Adichie, 1967; Gray, 1989). The risks posed by the presence of outliers in OLS and ML estimations are currently, nevertheless, widely recognized (Rousseeuw & Leroy, 2005). Therefore, when there are outliers and extreme observations in the data set, OLS and ML methods produce inaccurate estimates as unusual observations are sensitive to these approaches (Rousseeuw & Leroy, 2005). In such data sets, using OLS method to estimate regression parameters may yield inaccurate conclusions.

In the past 50 years, OLS and ML alternatives, also referred to as "robust" regression techniques, have attracted more attention (Jajo, 2005; Andersen, 2008). These methods are mainly aimed to provide stable results in the presence of outliers (Jajo, 2005). Recent investigations have concentrated on robust approaches, all of which were inspired by the theories of Hampel (1974) (Bagheri et al., 2010). Modern robust regression techniques can be quite helpful in instances where the aim is to understand how a random variable y is related to a group of p predictor variables. One reason is that even one outlier among

the values or one unusual observation in the dataset can have a significant impact on the parameter estimation of a typical linear model using OLS and ML methods (Wilcox, 1996). Another reason is that modem robust methods can be much more efficient than OLS and ML estimation methods yet maintain good efficiency under the ideal conditions of normality and a homoscedastic error term (Kalina, 2015).

Robust regression methods are particularly well-suited for real-world datasets that may contain noise or anomalies, as evidenced by the literature (Kalina, 2015). These methods can be used as alternative, effective models to manage outliers and other deviations from the assumptions of classical OLS and ML estimation regression methods (Denby & Mallows, 1977; Kalina, 2015; Ritschard & Antille, 1992). The goal of robust analysis is to fit a regression model to the bulk of the data prior to identifying outliers as points with large residuals from the robust solution (Jajo & Hussain, 1989). These methods down weight the influence of outliers, giving more reliable estimates of the relationships between variables (Adichie, 1967).

Some commonly used robust regresstion estimators include M estimators, MM estimators, LTS estimators, and S estimators (Ayinde et al., 2015; Chen, 2002; Wilcox, 1996). These have been discussed in details in Chapter Two.

1.4 Diagnostic Statistics

A model's diagnostic statistics are a set of measures calculated to identify unusual or influential observations in the fitted model (Bagheri et al., 2010). They play a vital role in assessing the quality and fit of the regression model (Ayinde et al., 2015). These diagnostic tools identify potential issues such as influential observations or model misspecification,

allowing to make necessary adjustments for more accurate results (Ayinde et al., 2015; Gray, 1989). These techniques usually detect outliers that go masked in a residual-only analysis (Gray, 1989).

The fundamental techniques of Cook and Weisberg in 1982 were the beginning of a great deal of effort on various techniques to find these unusual points (Bagheri et al., 2010). Many strategies have been proposed during the last thirty decades to identify outliers; these procedures or methodologies are often referred to as diagnostics (Jajo, 2005). The commonly used diagnostic statistics include; Cooks' distance measure, The Welsch-Kuh distance (DFFITS) and DFBETAS (Ayinde et al., 2015; Kannan & Manoj, 2015; Türkan et al., 2012). These methods have been discussed in Chapter Two.

Huber in 1991 took on the task of clarifying the seemingly ambiguous relationship between robustness and diagnostics, which is often viewed as hostile (Jajo, 2005). It is believed that the two techniques to data analysis are complimentary and equally important. Both robustness and diagnostics look at the outliers' problem from different perspectives, and the more ambiguous is the problem, more vital it is to look at it from all angles (Jajo, 2005). Therefore, despite robust regression methods providing a remedy to fitting problem, the need for regression diagnostics remain as they often provide useful information (Gray, 1989).

The classical regression methods proposed the deletion of identified outliers prior to fitting model to the suitable dataset, but they did not address the question of how much deletion is permissible (Rousseeuw & Leroy, 1988). In case of too many outliers, this leads to deletion of more observations (yet not all influential) giving biased results that cannot be interpreted (Türkan et al., 2012). It is therefore, widely acknowledged that unusual

observation in regression analysis requires specific attention. Gray (1989) emphasized that unusual observations often provide useful information and need to be used in collaboration with familiar skills and knowledge of analysis. Thus, even if robust regression methods provide a remedy to fitting problem, the need for regression diagnostics remains (Ritschard & Antille, 1992). The detection of unusual observation is an important problem in model building, inference and analysis of a regression model (Ayinde et al., 2015). Therefore, the use of model diagnostics is necessary for both classical and robust regression methods.

1.5 Statistical research gaps in the analysis of maternal anaemia data

Understanding the factors associated with WRA prevalence of anaemia is fundamental to reduce the world burden of anaemia, which is a public health problem worldwide. Identifying and implementing strategies focusing on the determinants of anaemia among WRA is key to address the global challenge of anaemia which leads to high morbidity and mortality in WRA wordwide. Regression analysis is one of the important statistical tool widely used to identify determinants and drivers of anaemia in WRA. There are numerous studies in literature that reported the factors associated with maternal anaemia using regression analysis. However, some analytical methods have been prematurely carried out without exhausting all that was required to understand the data at hand.

For example, in a study by Alem et al. (2023) reviewed in section 1.2, multilevel binary logistics regression was applied to study risk fators of maternal aneamia. However, the study never performed the model diagnostic statistics to assess the quality and fit of the best regression model. This could have helped to identify influential observations

or model misspecification and eventually allowing for necessary adjustments for more accurate results (Ayinde et al., 2015; Rousseeuw & Leroy, 1988). Therefore, it is not known whether the dataset had influential points that warranted special attention for improvement of the model.

Another study by Sunuwar et al. (2020) also reviewed in section 1.2 fitted multiple logistics regression model to identify factors associated with maternal aneamia. However, this study did not perform the model diagnostic statistics which would have helped to evaluate the validity and reliability of the model. These statistics could have detected outliers and influential data points that could affect the model's accuracy for further improvement of the model.

The study by Teshale et al. (2020), also reviewed in section 1.2, used multilevel mixed-effects generalized linear model to study risk factors of maternal aeamia. Although Linear mixed models provides best unbiased prediction in analysis of sample surveys, designed experiments and data with repeated measurements, can be influenced by outlying observations (Sinha, 2004). Therefore, examination of the outliers on mixed effects and variance component parameter estimates using model diagnostic statistics was necessary for possible attention to unusual observations. However, this study ignored the model diagnostic assessment. To address the concern of having unusual observations in the data set which could have affected the accuracy of the results, the study could have incorporated Robust statistics and report on robust standard errors. By incorporating robust statistics in GLMMs, study could have obtained more accurate and reliable estimates of the fixed effects, even in presence of influential observations in data set (Koller, 2016; Yau & Kuk, 2002).

Another study by Talukder et al. (2022) reviewed in section 1.2 studied risk factors of maternal aneamia using quantile regression model, which provide better estimates than classical regression when the data have lot of outliers in the conditional distribution (Rodriguez & Yao, 2017; Waldmann, 2018). Despite robustness of the Quantile regression model (Waldmann, 2018), it is highly recommended that robust methods should go together with regression diagnostics as they provide useful information (Gray, 1989). However, this study did not report on the regression diagnostics for additional information about influential data points. Regression diagnostics incorporation could have given more insights about WRA anaemia data.

Despite several studies reporting that model diagnostic statistics and robust regression methods help to detect unusual observations in the linear fitted model (Ayinde et al., 2015; Ronchetti & Huber, 2009; Rousseeuw & Leroy, 2005), there is absence/limitations of studies that applied these methods on maternal anaemia data to observe their performance. Often as the case, these studies fit these models with an assumption that these methods are robust enough in presence of unusual observations. This may not be true. Therefore, it's unclear, though, if their application provide comparable quality in the estimates of risk factors of maternal anaemia. Thus, there is need to evaluate performance of robust regression techniques and model diagnostics statistics when applied to the same data.

1.5.1 Problem statement

The maternal anaemia data are highly skewed in low and middle income countries, with some women having extreme measurements, and thus some of the previous studies suggested using nonparametric quantile regression to analyse such data (Talukder et al.,

2022). However, less attention has been paid in previous studies in sub-Saharan Africa to accounting for the outlier Haemoglobin outcomes in the analysis of such data. This study therefore applies mean, quantile, and robust regression and their diagnostic statistics to study the unusual mothers to anaemia in Malawi. An outlier observation is one that appears to deviate markedly from other data points of the sample in which it occurs (Kaombe & Manda, 2023b). For the data that are contaminated with outliers, robust regression technique is known to achieve high accuracy of estimation (Andersen, 2008). Specifically, the research assess sensitivity and resistance to outlier observations among the mean, quantile and robust regression models when applied to both simulated data and the real maternal anaemia from the 2015-16 Malawi demographic and health survey. Ignoring the impact of outliers in regression estimation leads to biased conclusions from a study (Kaombe & Manda, 2023b,a; Kaombe, 2024).

1.6 Study objectives

1.6.1 General objective

- To assess performance of mean, quantile and robust regression methods and diagnostics statistics when analysing maternal anaemia data in Malawi

1.6.2 Specific objectives

1. To assess efficiency of estimates from mean, quantile and robust regression models using both simulations and real data

2. To assess sensitivity to outlier observations by mean, quantile and robust regression models of estimates from mean, quantile and robust regression methods using both

simulations and real data

3. To examine sensitivity to influential observations by mean, quantile and robust regression models of estimates from mean, quantile and robust regression methods using both simulations and real data

1.7 Significance of the study

Maternal anaemia remains one of the serious causes of maternal mortality in sub-Saharan Africa (Stevens et al., 2022; Owais et al., 2021; Kinyoki et al., 2021; Chaparro & Suchdev, 2019). Sub-Saharan Africa has average maternal anaemia prevalence of 41.7 percent, which hinders efforts to promote safe reproductive health (Karami et al., 2022; Chaparro & Suchdev, 2019; Kassebaum et al., 2016). Malawi, like other countries in sub-Saharan Africa, is also dealing with a high prevalence of maternal anaemia, currently at 29.8 percent (Chanimbe et al., 2023), which raises concerns about the country's ability to achieve the global target of 15.2 percent or less by 2025 (Kinyoki et al., 2021; Sun et al., 2021). Effectively addressing maternal anaemia requires a comprehensive analysis of the data to identify evidence-based strategies that can work (Kinyoki et al., 2021). The persistence of this health outcome in women means that additional interventions are required to reverse the trend.

This study is significant as it aligns with the Global Technical Strategy for Malaria 2016-2030, aiming to reduce malaria incidence and mortality by 2030, and supports Sustainable Development Goal 3, specifically Target 3.1, which focuses on reducing the global maternal mortality ratio to less than 70 per 100,000 live births by 2030. In Malawi, Presidential Initiative on Maternal Health and Safe Motherhood launched in 2012 emphasize improv-

ing maternal health services, which are crucial in combating malaria's impact on maternal health (Walsh et al., 2018). By addressing malaria, this study contributes to achieving these global and national health targets. Therefore, use of proper statistical analysis methods for such data will contribute in unearthing essential features of this health problem in the region. This will in turn invite right interventions and policies to deal with this health issue. This research will therefore be crucial in contributing evidence-based data on appropriate statistical techniques that could be applied to analyse maternal anaemia data, following thorough analyses using both simulations and actual applications on real data sets that will be undertaken.

1.8 Thesis structure

This thesis is structured as follows. In chapter two, an overview of diagnostic statistics and robust regression is presented. In chapter three, study methodology in data and the statistical methods that were applied and their justification. The methods section also presents the simulation design that was carried out to compare the robust regression and diagnostic statistics in this study. Chapter four, presents the results from both simulations and applications of the statistical methods involved. Finally, chapter five presents a unifying discussion of the findings, limitation and conclusion.

CHAPTER TWO

REVIEW OF MEAN, QUANTILE, AND ROBUST REGRESSION METHODS

2.1 Mean regression methods

Mean regression methods are statistical techniques used to model the relationship between a dependent variable and one or more explanatory variables by estimating the mean of the dependent variable based on the values of the independent variables (Sarstedt et al., 2019). The objective of using mean regression methods is to estimate a continuous normal response variable based on known variables. Mean regression is versatile and widely applicable in various fields where researchers seek to understand and quantify the relationship between variables (Sarstedt et al., 2019). Additionally, the investigator normally evaluates the estimated relationship statistical significance, or the degree of confidence that the true relationship is near to the estimated relationship. Some commonly used regression methods in the modellling of maternal anaemia includes; Linear model and generalized linear model

2.1.1 Linear model

Rousseeuw & Leroy (2005) provides the formal definition for linear regression model. The model assumes a linear relationship between the predictors and the response variable, applicable when the response variable is continuous and normally distributed.

Let y_i be the continuous random variables and X_i, for $i = 1, 2..., \rho$ be the ρ covariates,

then multiple linear regression model is given by:

$$Y_i = X_{ij}\beta_j + \epsilon_i, \tag{1}$$

where Y_i is the response measured on a ratio scale on the i-th subject, $X_{ij} = (1, X_{i1}, X_{i2}, ..., X_{ip})$ is a row vector of measurements for fixed proxy variables on the i-th subject, $\beta_j = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is a column vector of corresponding fixed effects of the variables X on Y. The term ϵ_i represents the measurement error for the outcome of the i-th individual. The responses Y_i from different measurements are assumed to be indendently and identically distributed (Peña & Slate, 2006). In addition, it is assumed that $\epsilon_i \sim N(0, \sigma^2)$. For this reason, the relationship between the covariates X and Y is on average of Y, i.e. $EY_i|X_{ij}$. The inference techniques have to be applied at each case to yield accurate case-wise predictions.

For the linear model in equation 1, the standard set of underlying assumptions as specified by numerous studies Peña & Slate (2006); Verran & Ferketich (1987); Poole & O'Farrell (1971) include; linearity, normality, nonrelatedness (autocorrelation), homoscedasticity (constant variance) and independent variables without measurement error

The first required assumption in the linear model is linearity, linear relationship between dependent and independent variables. This requires that the relationships between Y and each of the independent variables X_i are linear in the parameters of the specific functional form chosen. According to Peña & Slate (2006), the mathematical representation of the assumption is given by $\mu_i = EY_i|X = \beta_{(i)}x_i$, where X_i is the i-th row of X. This assumption is ascertained if the residuals show no evidence of departure from linearity, residual scatter plot is around zero (Verran & Ferketich, 1987; Sevier, 1957).

Homoscedasticity is another required assumption. Linear model assumes constant variance of the conditional distribution. Peña & Slate (2006) provide the formal mathematical representation of the assumption, $Var(Y_i|X) = \sigma^2$, where σ^2 is the standard deviation and $i = 1, 2, \ldots, n$. The residual analysis procedure is used to test this assumption. The assumption holds when the residual variance is equal at all points of the predicted dependent variable (Thompson, 1982; Verran & Ferketich, 1987). This can be ascertained by studying the pattern of the errors' scatter plot against the predicted values (Sevier, 1957)

Third assumption is uncorrelatedness, the values of μ are serially independent. This assumes that the values of the mean are independent of each other and their covariance is zero. Such that the error from one observation does not affect (or is independent) the error obtained from another observation. The assumption is formally defined by the mathematical representation, $cov(Y_i, Y_j|X) = 0$ for i not equal to j (Thompson, 1982). This is achieved if the residuals are independent.

Fourth assumption is Normality distribution of error term (residuals). According to Poole & O'Farrell (1971), this assumes that the dependent variable, $(Y_1, Y_2, \ldots, Y_n)|X$ has a normal conditional distribution. This assumption is achieved if the residuals (error term) are approximately normally distributed, $N(0, \sigma^2 I_n)$, I_n is identity matrix of size n by n (Verran & Ferketich, 1987). Testing the normality assumption in linear regression is essential to ensure the validity of statistical inferences drawn from the model.

According to numerous studies Peña & Slate (2006); Thompson (1982); Sevier (1957), residual analysis, Shapiro-Wilk Test and graphical methods are some common methods used to test the normality assumption. Residual analysis is one of the most common ways to test for normality in linear regression is by examining the distribution of residuals

(errors) from the regression model. You can create a histogram or a Q-Q plot of the residuals and compare it to a normal distribution. Shapiro-Wilk test is statistical test assesses whether a sample comes from a normally distributed population (Khatun et al., 2021). In linear regression, you can apply this test to the residuals to determine if they follow a normal distribution. Graphical methods are visualization tools such as a normal probability plot or a density plot which can help assess the normality assumption visually. These methods helps to determinate whether the normality assumption holds for the residuals in the model. If normality is violated, transformations or non-parametric regression techniques may be considered (Fox, 2002)

The fifth assumption requires that independent variables must be without measuremnet error. It assumes that each value of independent variables, X_i and dependent variable, Y is observed without measurement error. According to Poole & O'Farrell (1971), this assumption maybe partially relaxed to say that X_i must be without measurement error.

The simplest case is when the there is one explanatory variable in the model, and the model is considered as a simple linear regression

2.1.2 Generalized Linear Model (GLM)

Generalized Linear Models (GLMs) extend linear model to accommodate different types of response variables, for example, binary, count or exponential and categorical, that are nonnormal and has nonhomogeneous variance. Nelder and Wedderburn (1974) introduced the concept of generalised linear models (GLM), which McCullagh and Nelder (1989) went into great detail to examine (Myers & Montgomery, 1997). According to Dobson & Barnett (2018) and Myers & Montgomery (1997) GLM regression modelling is possible in cases where the responses are distributed as members of the exponential family, that

is, when;

$$f(y;\theta) = \exp[a(y)b(\theta) + c(\theta) + d(y)] \tag{2}$$

where y in equation 2 is response variable (observed data), θ is parameter of the distribution, $a(y)$ is function of y (normalizes the distribution), $b(\theta)$ is function of θ (relates to the natural parameter), $c(\theta)$ is function of θ (cumulant function) and $d(y)$ is function of y (ensures valid probability distribution).

Such that the joint distribution function is given by;

$$f(y_1, \ldots, y_n; \theta_1, \ldots, \theta_n) = \exp[\sum_{i=1}^{n} a(y_i)b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)] \tag{3}$$

where y_i in equation 3 is the i-th response variable (observed data), θ_i is the i-th parameter of the distribution, $a(y_i)$ is a function of the i-th response variable y_i, $b(\theta_i)$ is a function of the i-th parameter θ_i, $c(\theta_i)$ is a function of the i-th parameter θ_i and $d(y_i)$ is a function of the i-th response variable y_i.

GLMs allow for the specification of a link function and a distribution family appropriate for the response variable (Dobson & Barnett, 2018). Generalization has been due to the realisation that a wider class of distributions known as the exponential family of distributions has many of the "nice" properties of the Normal distribution.

Many well-known distributions belong to the exponential family which includes; the Poisson, Normal and Binomial distributions (Neuhaus & McCulloch, 2011). In GLMs, relationship between the response and explanatory variables need not be of the simple linear form. Some commonly used models includes; logistic, probit, poisson and survival models.

For y_i response variable and X_i, $i = 1, 2..., \rho$, ρ covariates, the Generalized Linear Model

(GLM) for y_i on the covariates $\mathbf{x}$ is given by:

$$g(\mu_i) = \eta = \mathbf{x}^T\beta = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{\rho-1} X_{\rho-1} \tag{4}$$

where g in equation 4 is a known monotone differentiable function, called the link function linking the mean, μ_i of the response y_i to the linear predictor η, $\mu = g^{-1}(\eta) = b'(\theta_i)$.

In case of binomial response, the link function g is the logit, given by;

$$logit(\mu_i) = \log(\frac{\mu_i}{1-\mu_i}) = \eta = \mathbf{x}'\beta \tag{5}$$

where η in equation 5 is the linear predictor, $\mathbf{x}$ is the vector of predictor variables (features), β is the vector of coefficients associated with the predictor variables.

This produces the model $\mu = \frac{1}{1+\exp(-\mathbf{x}'\beta)}$ which is a logistic model.

2.1.3 Generalized Linear Mixed Model (GLMM)

Generalized Linear Mixed Models (GLMMs) are statistical models that extend GLM by incorporating mixed models, models with both fixed and random effects, in the linear predictor η. In Linear models the regression coefficients are considered as fixed, unknown constants. However, in some scenario, when the observations are correlated, it is necessary to assume that some of the coefficients are random (Jiang & Nguyen, 2007). In Longitudinal data, the responses may not be necessarily normal. For example, in cases of binomial responses, GLMMs are applied to incorporate intra subject correlation of observations and the subject is modelled as random. Generally, a Generalized linear mixed model (GLMM) is fully specified by defining its response variable distribution, link function,

categorical and continuous fixed-effect predictors, and random effects, which indicate how certain model parameters vary at random in all groups.

According to Jiang & Nguyen (2007); Clayton (1996) generalized linear mixed effect model is denoted as

$$g(\mu_i) = \eta = X_i\beta + Z_i\alpha \tag{6}$$

where g in equation 6 is the link function, μ_i is the mean of the response, X_i is the design matrix for the mixed effects β and Z_i is the design matrix for random effects α.

2.2 Parameter Estimation Methods in Mean Regression

2.2.1 Ordinary Least Square (OLS) Estimation

OLS estimator, minimizing sum of squared residuals, $Q = \sum_{\epsilon_i^2}^{n} = \epsilon^T\epsilon = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) = Y^TY - 2(\hat{\beta})^TX^TY + (\hat{\beta})^TX^TX\hat{\beta}$ is most commonly used technique to estimate the linear regression parameters (Ayinde et al., 2015). OLS bases its predictions on the assumptions such as normality of the dependent variable on the regression structure (Bagheri et al., 2010; John & Nduka, 2009).

According to Lakshmi et al. (2021); Türkan et al. (2012), the formula for OLS estimates for coefficients, β and σ in linear model are expressed by;

$$\hat{\beta}_N = (X^TX)^{-1}X^TY \tag{7}$$

where X in equation 7 is the matrix of independent variables, Y is the vector of the dependent variable, X^T denotes the transpose of X and $(X^TX)^{-1}$ represents the inverse of the matrix product of X^T and X. The Standard Error of the Coefficients (SE) =

$\sqrt{(}\sigma * X^T X)^{-1})$

and formula for OLS estimates for variance, σ is expressed by

$$(\hat{\sigma}_N)^2 = \frac{1}{2\sigma^2} \sum (y_i - x_i \hat{\beta}_N)^2 = \epsilon^T \epsilon / (n - k) \tag{8}$$

where σ^2 in 8is the estimated variance of the errors, often referred to as the mean squared error (MSE), y_i are the observed values of the dependent variable, x_i are the observed values of the independent variable(s), $\hat{\beta}_N$ is the estimated coefficients from the OLS regression, ϵ^T is the vector of residuals (errors), calculated as the difference between the observed values and the predicted values from the regression model, n is the number of observations in the dataset and k is the number of estimated parameters (including the intercept) in the model.

The vector of fitted values is represented by;

$$\hat{Y_N} = X\hat{\beta}_N = X(X^T X)^{-1} X^T Y = HY \tag{9}$$

where $H = X(X^T X)^{-1} X^T$ in equation 9 is the vector of leverage measure, the influence of an individual data point on the model's parameter estimates.

The coefficients, $\hat{\beta}$ in linear model represent the change in the dependent variable for a one-unit change in the independent variable, holding all other variables constant. The residual standard deviation, σ, represents the average distance between the observed values of the dependent variable and the values predicted by the model. A lower residual standard deviation indicates that the model's predictions are closer to the actual data points, suggesting a better fit. In the context of regression coefficients, the standard

deviation, SE, measures the uncertainty or variability in the estimated coefficients. Larger standard deviations for the regression coefficients indicate that the estimates are less precise or more variable. For the leverage measure, H, a data point with high leverage has an independent variable value that is further away from the mean of the independent variables. High leverage points can exert significant influence on the parameter estimates in the regression model.

OLS estimation is not applicable to GLM and GLMM as these models violates the assumptions of normality and homogeinity of variance.

The method of OLS has nice property of providing best estimates under very general conditions. However, the estimates obtained are prone to gross errors in the presence of unusual observations called outliers (Adichie, 1967).

2.2.2 Maximum Likelihood (ML) estimation

Maximum Likelihood Estimation (MLE) involve several key functions that play a significant role in determining parameter estimates and assessing the goodness of fit of a model. These functions include; Maximum Likelihood function, log-likelihood, Fisher information and score function.

The Maximum Likelihood function, denoted as $L(\theta)$, represents the likelihood of observing the data given the model parameter θ, joint distribution and is formally given by;

$$L(\theta) = \prod f(X_i|\theta) \tag{10}$$

where $f(X_i|\theta$ is the probability density function (PDF) of the data point X_i given the parameter θ.

The log-likelihood function, denoted as $l(\theta)$, is the natural logarithm of the Maximum Likelihood function in equation 10 and is given by;

$$l(\theta) = \ln L(\theta) = \log(\prod f(X_i|\theta)) \tag{11}$$

The score function, also known as the gradient of equation 11 provides information about the direction in which the parameter should be updated to maximize the likelihood. Score function is formally defined by

$$Score(\theta) = \delta l(\theta) = \delta[\log L(\theta)] \tag{12}$$

where δ in equation 12 denote the gradient operator, first order derivative.

The Fisher Information ($I(\theta)$) measures the amount of information that the data provides about the parameter θ. It quantifies the expected curvature of the log-likelihood function around the true parameter value, represented by the formula;

$$I(\theta) = -E[\delta l(\theta)] \tag{13}$$

where E in equation 13 denotes the expectation operator.

These functions are fundamental in the MLE framework for estimating parameters and assessing the statistical properties of the model.

Linear Model ML estimation

For Linear Model, Maximum Likelihood (ML) Estimation is alternatively used to estimate the model parameters. The method leads to the same estimators for normal error regression model as those obtained from OLS Method(Dobson & Barnett, 2018). The Linear Model likelihood function is formally defined by;

$$L(\beta, \sigma^2) = \frac{1}{(2\pi\sigma^2)^(n/2)} \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - X_{ij}\beta_j)^2 \tag{14}$$

where Y_i in equation 14 is the response measured on a ratio scale on the i-th subject, $X_{ij} = (1, X_{i1}, X_{i2}, ..., X_{ip})$ is a row vector of measurements for fixed proxy variables on the i-th subject, $\beta_j = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is a column vector of corresponding fixed effects of the variables X on Y.

Maximizing the Score function, $Score(\beta, \sigma^2) = \delta L(\beta, \sigma^2)$ with respect to $\beta_0, \beta_1, \ldots, \beta_{\rho-1}$ leads to estimators for $b_0, b_1, \ldots, b_{\rho-1}$.

GLM ML estimation

For GLM, Maximum Likelihood (ML) Estimation or iterative algorithms are used to estimate the model parameters (Myers & Montgomery, 1997). The Likelihood function is given by;

$$L(\theta; y_1, \ldots, y_n) = \exp[\sum_{i=1}^{n} a(y_i)b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i)] \tag{15}$$

And log-likelihood function, drived by taking log of function 15, is given by;

$$I(\theta; y_1, \ldots, y_n) = \sum_{i=1}^{n} a(y_i)b(\theta_i) + \sum_{i=1}^{n} c(\theta_i) + \sum_{i=1}^{n} d(y_i) \tag{16}$$

In order to obtain ML estimate $\hat{\theta}$, first derivarive of equation 16 is equated to zero and solved, $E[\frac{\partial l}{\partial \theta}] = 0$ and variance is obtained from solving equation $E[\frac{\partial^2 l}{\partial \theta^2} + (\frac{\partial l}{\partial \theta})^2] = 0$ which are simplified to the following equations

$$E(y_i) = \mu_i = \frac{-c'(\theta)}{b'(\theta)} \tag{17}$$

and

$$Var(Y_i) = \frac{b''(\theta_i)c'(\theta_i) - c'"(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3} \tag{18}$$

these two equations, 17 and 18 are very useful as far as GLM estimation is a concern.

According to Dobson & Barnett (2008), to obtain the maximum likehood estimator for the parameter β_j which are related to Y_i's through $E(Y_i) = mu_i$ and $g(\mu_i) = x_i^T\beta$, chain rule for differentiation is used and is given by;

$$\frac{\partial l(\theta; y)}{\partial \beta_j} = U_j = \sum_{i=1}^{n}[\frac{\partial l_i}{\partial \theta_i}\frac{\partial l_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \beta_i}] = \sum_{i=1}^{n}[\frac{(y_i - \mu_i)}{var(Y_i)}x_{ij}(\frac{\partial \mu_i}{\partial \eta_i})] \tag{19}$$

Fisher information (information matrix) of equation 19 is given by

$$\tau_{jk} = E[U_j U_k] = \sum_{i=1}^{n}\frac{x_{ij}x_{ik}}{var(Y_i)(\frac{\partial \mu_i}{\partial \eta_i})^2} = X^T W X \tag{20}$$

where W in equation 20 is N by N diagonal matrix given by $w_{ii} = \frac{1}{var(Y_i)}(\frac{\partial \mu_i}{\partial \eta_i})^2$

Vector of estimates, b^m, of the parameters $\beta_{-i}, \ldots, \beta_p$ at m-th iteration (Dobson & Barnett, 2008) is given by

$$b^m = b^{m-1} + [\tau^{m-1}] + U^{m-1} = (X^T W X)^{-1}(X^T W z) \tag{21}$$

where z in equation 21 has elements $z_i = \sum_{i=1}^{n} x_{ik} b_k^{m-1} + (y_i - \mu_i)(\frac{\partial \eta_i}{\partial \mu_i})$

In both Linear Model and GLM, for a continuous explanatory variable X, it's coefficient, β_i, represents the change in the response corresponding to a change of one-unit in X. For categorical explanatory variables, there are parameters for the different levels of a factor.

GLMMs ML estimation

GLMMs estimate fixed effects (relationships between predictors and the outcome) and random effects (variance components) using methods like Restricted Maximum Likelihood (REML) (Jiang & Nguyen, 2007). For the Gaussian Mixed Models, the point likelihood function is given by

$$f(y) = \frac{1}{(2\pi)^{n/2}|V|^{1/2}} exp - \frac{1}{2}(y - X\beta)' V^{-1}(y - X\beta) \tag{22}$$

where $V = V(\theta)$ in equation 22, n is the dimension of y. And log-likelihood is given by:

$$l(\beta, \theta) = c - \frac{1}{2} log(|V|) - \frac{1}{2}(y - X\beta)' V^{-1}(y - X\beta) \tag{23}$$

and Score function, obtained by differentiating equation 23 is defined by:

$$\frac{\partial l}{\partial \beta} = X' V^{-1} y - X' V^{-1} X\beta \tag{24}$$

Maximizing equation 24, $\frac{\partial l}{\partial \beta} = 0$ and solving it simplifies to the ML estimates $\hat{\beta}$

$$\hat{\beta} = (X' \hat{V}^{-1} X)^{-1} X' \hat{V}^{-1} y \tag{25}$$

Fixed effects coefficients, given by equation 25, represent the average impact of predictors on the response, while random effects account for variability within groups.

2.3 Nonparametric regression methods

Nonparametric regression methods offer flexibility and robustness in modeling complex relationships and are valuable when standard parametric models may not adequately capture the underlying patterns in the data (Fox, 2002). Relationship between the response and explanatory variables does not necessarily to be linear as opposed to mean regression models. Furthermore, the relationship between the response and explanatory variables does not depend on any particular form of regression function (Čížek & Sadıkoğlu, 2020). They are suitable for various application in different fields because they offer a more detailed understanding of the relationships between variables. Commonly used nonparametric regression methods in the modelling of maternal anaemia includes; quantile regression model and generalized additive models (GAM)

2.3.1 Quantile regression model

Quantile regression (QR) is the statistical technique performed to estimate and provide inference about the conditional quantile functions, the function that describes the relationship between explanatory variables and the conditional quantile of a response variable without assuming a specific distribution (John & Nduka, 2009). It uses a general linear model to fit conditional quantiles of a response, providing information not available through mean regression methods. Quantile regression model assumes no parametric form for the conditional distribution of the response and no constant variance for the response, unlike least squares regression (Rodriguez & Yao, 2017). Therefore, Quantile regression is

more effective than classical methods for explaining relationships in circumstances where mean regression conditions, such as $E(\epsilon_i) = 0$, homoscedasticity $Var(\epsilon_i) = \sigma^2$, no auto-correlation $Cov(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$, normality assumptions are not met or the interest resides in the outer regions of the conditional distribution. It performs better than classical regression when the data is skewed as it minimizes the median than mean (Waldmann, 2018). QR offers advantages for various types of data, including independent, time-to-event, and longitudinal data (Huang et al., 2017)

Quantiles are commonly defined by ordering and sorting sample observations. Quantile regression as introduced by Koenker and Bassett in 1978 extend ideas of quantiles, τ's or percentile to estimation of conditional quantile functions models (Koenker & Hallock, 2001). In the model, quantiles of the condition function distribution of the response variable are expressed as function of observed covariates. Quantile regression extends the location shift model by determining the effect of factors on the shape and scale of the entire response distribution (Waldmann, 2018). The gap between quantile lines reflects whether the distribution is skewed to the right or left.

For response variable (Y) and it's distribution function $F(y) = \rho_\tau(Y \leq y)$, the τ-th, for $0 < \tau < 1$, quantile is defined as $Q(\tau) = inf(x : F(Y) \geq \theta)$

Quantile model for quantile level τ of the response is given by:

$$Q_\tau(Y_i) = X_{ij}\beta_j(\tau) + \epsilon_i(\tau) \tag{26}$$

where i in equation 26 is observation $1, ..., n$

One of the underlying assumption for the Quantile regression is heteroscedasticity, $V(\epsilon_i \neq$

$V(\epsilon_i)$ for all $i \neq j$. When a dataset has heteroscedasticity, OLS findings are no longer reliable (John & Nduka, 2009; Rodriguez & Yao, 2017). According to John & Nduka (2009), Quantile regression gives complete information about relationship between response and independent variables by posing the question of relationship between the response and the independent variables at any quantile of the conditional distribution function. Quantile Regression models can detect heterogeneous effects of covariates at different quantiles of the outcome and provide more robust and comprehensive estimates than mean regression, especially when the normality assumption is violated or outliers and long tails are present (John & Nduka, 2009; Huang et al., 2017).

According to Rodriguez & Yao (2017); Koenker (2005); Huang et al. (2017), there are various commonly used types of Quantile Regression which include; Lower QR (such as 25th percentile), Median Regression (50th percentile) and Upper QR (such as 75th percentile). Lower Quantile Regression (such as 10th, 25th percentile) estimates the relationship at lower quantiles of the response variable, providing insights into the lower end of the distribution. Median Regression (50th percentile) is the most commonly used type of quantile regression which estimates the relationship at the median of the response variable. In cases with asymmetries and heavy tails, the sample median (50th percentile) is a stronger indicator of centrality than the mean (Koenker, 2017). Upper Quantile Regression (such as 75th, 90th percentile) estimates the relationship at upper quantiles of the response variable, providing insights into the upper end of the distribution. Comprehensive understanding of the relationship between variables across the entire distribution of the response variable is gained by performing QR at different levels.

2.3.2 Generalized Additive Model

The model addresses the weakness of GLM which has a strictly linear predictor η. However, sometimes, the relationship between predictors and the response might be nonlinear, observations may be partially or temporarily correlated and also the covariates may not sufficiently describe individual heterogeneity. To address these difficulties, the linear predictor in GLM is replaced by the structured additive regression model (STAR) predictor. The model is defined by:

$$\eta_i = f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + f_{spat}(s_i) + \mu_i'\gamma \tag{27}$$

where f_j for $j = 1, 2, 3, \ldots, p$ in equation 27 are smooth functions expressing non linear relationship between the response variable and the continuous covariates, μ is the vector of the fixed effects, $f_{spat}(s_i)$ is partially correlated (random) effect of the location s_i where an observation pertains to.

2.4 Parameter estimation in nonparametric regression

2.4.1 Parameter estimation in Quantile Regression

The regression coefficients in the quantile model in Equation 26 are estimated by minimising a loss function called the check function, $\rho_\tau(r) = \tau max(r, 0) + (1 - \tau)max(-r, 0)$, $\tau \in (0, 1)$ (Rodriguez & Yao, 2017)

$$argmin_{\beta_0,\ldots,\beta_\rho}(\tau) \sum_{i=1}^{n} \rho_\tau \left[Y_i - X_{ij}\beta_j(\tau)\right]. \tag{28}$$

The minimization issue generates unique regression coefficients for each quantile level. The

median regression function is represented by $\tau = 0.5$, while the absolute value function is represented by $2\tau_{0.5}(r)$

In quantile regression, the estimated coefficients represent the change in the response variable at a specific quantile (Jamee et al., 2022). This provides more comprehensive understanding of the relationship between variables, especially when the relationship is not constant across different quantiles.

2.4.2 Parameter estimation in Generalized Additive Model

Parameter estimation in Generalized Additive Models (GAMs) involves estimating the smooth functions for each predictor variable while simultaneously estimating the parameters of the model. To prevent overfitting, models are estimated using penalised maximum likelihood estimation, such as maximising (Wood, 2004). Penalized maximum likelihood function is given by:

$$l(\eta) - \frac{1}{2}\sum_{j}\theta_j \int [f_j''(x)]^2 dx \tag{29}$$

where l in equation 29 is the log-likelihood of the linear predictor and the terms in the summation are measures of the wiggliness of the component functions of the GAM. Smoothing parameters (θ_i) determine the balance of fit and smoothness. The penalised likelihood is maximised using penalised iteratively reweighted least squares (P-IRLS).

2.5 Diagnostic statistics for mean regression methods

2.5.1 Outliers and leverage measures

Mean Regression diagnostics becomes necessary in regression analysis in order to detect the presence of outliers and influential points (Ayinde et al., 2015). An outlier is an

observation that appears to deviate markedly from other data points of the sample in which it occurs (Barnett et al., 1994). These outliers are frequently unrecognized because so much data is now processed by computers without sufficient monitoring. Such that many real world data set for which normal assumption are made, are skewed, heavy-tailed distribution due to presence of outliers (Chen, 2002; Koller, 2016).

Univariate data has unusual value for a single variable and much concern is an outlier in the dependent variable in the regression analysis (Kannan & Manoj, 2015). The widely used methods to identify outliers in univariate data are Box plots and scatter plots. According to Ritschard & Antille (1992); Chatterjee & Hadi (1986); Cook (1977, 2000), the classical approach to detection of outliers focuses on standardized least square residuals.

In mean regression, the simplest statistic for analysing outlier observations is the raw residual given by:

$$e_i = Y_i - \hat{Y}_i = Y_i - X_{ij}\hat{\beta}_j, \tag{30}$$

where $\hat{\beta}_j$ in equation 30 is the maximum likelihood estimator for the regression coefficients.

In Generalized Linear Model, we consider residuals that are approximately normally distributed. This provide more incisive investigation to consider first recipes for calculating of residuals $R(y_i, \theta_i)$ treating θ_i as known and the replacing θ_i by fitted values $\hat{\theta}_i = g(x_i'\hat{\beta})$ (Pierce & Schafer, 1986). There are two possibilities considered, linear and transformed residuals in GLM (Pierce & Schafer, 1986)

The linear residuals is denoted by:

$$R_L(y, \theta) = y - E_\theta(y)/SD_\theta(y) \tag{31}$$

where E and SD in equation 31 denote the mean and standard deviation and transformed linear residual is given by:

$$R_T(y, \theta) = t(y) - E_\theta[t(y])/SD_\theta[t(y)] \tag{32}$$

where $t(\dot{})$ in equation 32is specified tranformation, usually chosen based on particular distribution of y.

Residuals play a crucial role in assessing the quality and validity of a linear regression model. Firstly, by examining the distribution of residuals, helps to check if the normality assumption holds. Deviations from normality could indicate issues with the model assumptions. Secondly, residual plots helps to assess whether homoscedasticity assumption hold, the variance of the errors is consistent across all levels of the independent variables. Patterns in the residuals against fitted values may suggest heteroscedasticity. Thirdly, residuals helps to detect ouliers in the dataset which may affect model performance. Forthly, residual plots can also be used to assess how well the model fits the data. A pattern in the residuals may suggest that the model is missing important nonlinear relationships. Lastly, residuals are crucial for conducting hypothesis tests and calculating confidence intervals. They help assess the precision of the estimates and the significance of the predictor variables. Therefore, careful analysis and interpreting of the residuals helps to understand the strengths and weaknesses of the fitted linear regression model.

There are many methods for the detection of outliers in linear model, both graphical and analytical (Arimie et al., 2020). The graphical methods include Scatter graph, Boxplot, Williams graph, Rankit graph (or Q-Q Plot) and graph of predicted residuals. The analytical methods are predicted residuals, standardized residuals, studentized residuals

and Jack-knife residuals

The standardized (Studentized) residuals, $\epsilon_{S,i}$, used to detect outliers (Cook, 2000; Ritschard & Antille, 1992), it is given by:

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}}, \tag{33}$$

where $s = \sqrt{\frac{\sum_1^n e_i^2}{n-p}}$ in equation 33 is the estimate for σ, p is the number of regression parameters, $h_{ii} = X_i(X^TX)^{-1})X_i^T$ is the i-th row of the diagnonal of the hat-matrix, called leverage.

If an observation has a studentized residual that is larger than 2 (in absolute value) is regarded as an outlier. If the linear regression model is appropriate, with no outlying observations, each Studentized residual follows a t distribution with $n-p-1$ degrees of freedom. The standardized residuals of more than 3 potentially indicate outlier (Arimie et al., 2020)

The jacknife residuals is denoted by:

$$\epsilon_{J,i} = \hat{\epsilon_{S,i}}\sqrt{\frac{n-p-1}{n-p-\epsilon_{S,i}^2}} \tag{34}$$

where $\epsilon_{S,i}$ in equation 34 is the Studentized residuals. The jacknife residual examine the influence of individual point on the quadratic error of the prediction.

The predicted residual for observation i is defined as the residual for the i-th observation that results from dropping the i-th observation from the parameter estimates. Predicted residual is denoted by:

$$\epsilon_{P,i} = \frac{\epsilon_i}{1-h_i} \tag{35}$$

PRESS statistic is the sum of squares of predicted residual error, which have an ability to assess model's predictive ability (Arimie et al., 2020). In least square regression, PRESS is denoted by

$$PRESS = \sum_{1}^{n} \frac{\epsilon_i}{1 - h_i}^2 \tag{36}$$

where ϵ_i in equation 36 is residual and h_i is leverage value for the i-th observation. The predictive power of the model increases with decreasing PRESS value (Arimie et al., 2020).

In regression analysis, the concept of leverage is employed to identify the observation(s) that deviate from the corresponding mean covariate values (Chaku & Donev, n.d.). It therefore, doesn't have a very significant effect on the outcome of the models fitting but have greater potential to pool the regression line. If the leverage point falls outside the overall pattern, it can seem to be influential.

In Linear Model, the leverage measure is given by:

$$h_i = X_i(X^T X)^{-1})X_i^T \tag{37}$$

where h_i in equation 37 is i-th diagnonal element, interpreted as amount of Leverage or influence exerted by Y_i on $\hat{Y}_i$, h_i is large if $h_i \geq 2\frac{p}{n}$ where $p = \sum_1^n h_i$ (Gray, 1989).

The leverage value is related to the residual variance by $Var(e_i) = \sigma^2(1 - h_i)$. Implying that a high leverage point usually has a smaller residual value.

In GLM, the leverage formula takes into account the link function and the variance function of the model. Both linear models and GLMs use leverage to assess the influence of individual data points on the model fit.

2.5.2 Cook's distance measures

This measure is called Cook's distance and was proposed by Cook in 1977 (Cook, 1977). Cook's distance (D_i), (Cook, 1977, 2000), measures the distance between the estimates of the regression coefficients with the i-th observation $\hat{\beta}$

and without the i-th observation $\hat{\beta}_{-i}$ for the metric $\frac{1}{p\hat{\sigma}^2}(X^TX)$. Therefore, D_i is the aggregate influence measure of i-th deleted case on all fitted values. Such that D_i is defined by:

$$D_i = \frac{(\hat{\beta} - \hat{\beta}_{-i})^T(X^TX)(\hat{\beta} - \hat{\beta}_{-i})}{p\hat{\sigma}^2} = \frac{r_i^2}{p}\frac{h_i}{(1-h_i)}t_i \tag{38}$$

where $\hat{\beta}$ and $\hat{\beta}_{-i}$ in equation 38 respectively provide estimate on all n data points and the estimate obtained after the i-th observation is deleted. Cook suggests that D_i be compared to a central F distribution, $F(p, n-p)$. For example, if the percentile value is less than about 20 percent, the unit has little apparent influence on the regression coefficients (Oyeyemi et al., 2017). On the other hand, if the percentile value is near 50 percent or more, the influence is partially important (Ayinde et al., 2015). The i-th deleted case is considered influential if $D_i > 1$ (Cook, 1977).

2.5.3 The Welsch-Kuh distance (DFFITS)

Welsch and Kuh in 1977, Welsch and Peters in 1978, and Belsley, Kuh, and Welsch in 1980 suggested using $\hat{\sigma}_i^2$ as an estimate of σ^2 (Chatterjee & Hadi, 1986) and called the impact of i-th observation on the i-th predicted value by scaling the change in prediction at x_i when the i-th observation is omitted (Ayinde et al., 2015) $DFFITS_i$. DFFITS diagnostic combines the information in the leverage h_i, and the Studentized residual e_i.

The influence that case i has on the fitted value $\hat{Y}_i$ is defined by:

$$DFFITS_i = \frac{|\hat{Y} - \hat{Y_{-i}}|}{\hat{\sigma}_{-i}\sqrt{h_i}} = \frac{|X_i^T(\hat{\beta} - \hat{\beta_{-i}})|}{\hat{\sigma}_{-i}\sqrt{h_i}} = |t_i|\sqrt{\frac{h_i}{1-h_i}} \tag{39}$$

where $t_i = \frac{\epsilon_i}{\hat{\sigma}\sqrt{(1-h_i)}}$ in equation 39 is i-th studentized residual (also called externaal studentized residual).

It is recommended that a $|DFFITS| \geq 2\sqrt{\frac{p}{n}}$ requires attention for large data set and if DFFITS is greater than 1 for small to medium data set (Chatterjee & Hadi, 2009; Türkan et al., 2012), where p is the number of independent variable and $\frac{p}{n}$ is the mean leverage.

2.5.4 DFBETAS

DFBETA is used to determine the changes in parameters of the new regression equation produced after removing the ith observation from the dataset (Belsley et al., 2005). Thus, DFBETAS measures influence of i-th case on each regression coefficients, b_k. DFBETAS statistic is defined by:

$$DFBETAS_i = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{\sqrt{MSE_{(i)}C_{jj}}}, \tag{40}$$

where C_{jj} in equation 40 is the j-th diagonal element of the quantity $(X^TX)^{-1}$, and $MSE_{(i)}$ the mean square error estimate obtained after deleting the i-th case in the regression model fitting, is used to estimate the error term variance, σ^2. The i-th case's significant influence on the k-th regression coefficient is indicated by large absolute value of $(DFBETAS)_{k(i)}$. The value with higher BEFBETAS is considered an outlier. DFBETAS is considered large if it is greater than 1 for small data set or $\frac{2}{\sqrt{n}}$ for large data set (Ayinde et al., 2015). When the sample size is big, DFBETAS has limited sensitivity in outlier detection, but it is most effective in small sample sizes and outlier percentages

(Oyeyemi et al., 2017).

While a larger DFBETA value indicates an outlier, DFBETA values calculated from observations decrease proportionally as the number of observations increases (Bahadir et al., 2014). DFBETAS is considered large if is greater than 1 (small data) or $\frac{2}{\sqrt{n}}$ (large data).

2.6 Robust Regression methods

2.6.1 Maximum Likelihood Type Estimation (M-estimator)

It was introduced by Huber (1973) and is a commonly generally used method due to it's simplicity, both computationally and theoretically (Ayinde et al., 2015). It is robust when the outliers are in the response direction (y-direction) (Chen, 2002).

The M-estimator's goal is to minimise a function of the errors (loss function), ρ rather than the sum of squared errors, goal of OLS (Ayinde et al., 2015). The objective function of the M-estimate is:

$$Min \sum_{i=1}^{n} \rho(\frac{e_i}{s}) = Min \sum_{i=1}^{n} \rho(\frac{Y_i - X_i\beta}{s}) \tag{41}$$

where s in function 41 is estimate of scale often formed from linear combination of the residuals.

A reasonable ρ should have the properties: $\rho(e) \geq 0$, $\rho(0) = 0, \rho(e) = \rho(-e)$, and $\rho(e_i) \geq \rho(e_i^T)$ for $|e_i| = |e_i^T|$. Minima solution associated with equation 41 is obtained by taking Gauss-Newton iterations, helped by R ROSEPACK package: $\sum_{i=1}^{n}(\phi)(\frac{Y_i - X_i\beta_i}{s})X_i$ where ϕ is a derivative of ρ.

In general, the Huber M-estimator outperforms OLS regression when outliers are located

along the y-axis rather than when outliers are located along the x-axis (Kim & Li, 2023).

2.6.2 Schweppe's Estimators (S-estimator)

S-estimator is a high breakdown value method introduced by Rousseeuw and Yohai in 1984 (Chen, 2002). S-estimator which is derived from a scale statistic in an implicit way (Rousseeuw & Hubert, 2018), corresponding to s(θ) where s(θ) is a certain type of robust M-estimate of the scale of the residuals. Tukey's weight function was suggested and is defined as ρ(x)

S-estimator is defined by minimization of dispersion of residuals: Minimize $S(e_1(\theta), ..., e_n(\theta)$, defined as solution of

$$\frac{1}{n}\sum_{1}^{n}\rho(\frac{e_i}{s}) = K \tag{42}$$

where $s(\theta)$ in equation 42 is a type of robust M-Estimate of scale of residuals, K is a constant and $\rho(\frac{e_i}{s})$ is the residual function, k = 1.547, is a common choice. This S-estimator resists contamination of up to 50 percent of outliers; it is said to have a breakdown point of 50 percent (Verardi & Croux, 2009). Unfortunately, this S-estimator has a Gaussian efficiency of only 28.7 percent (Verardi & Croux, 2009).

2.6.3 Least Trimmed Squares (LTS) estimator

LTS estimation is a high breakdown value method introduced by Rousseeuw in 1984 (Ayinde et al., 2015). The breakdown value expresses the percentage of contamination that a process can tolerate without losing its resilience (Chen, 2002). LTS eliminates possible outliers by running the analysis on trimmed or winsorized distributions (Yaffee, 2002). Distributions that have their outliers trimmed prior to the analysis are sometimes called trimmed means procedures. According to Rousseeuw, the LTS procedure is more

efficient than the S or M procedure (Yaffee, 2002).

LTS estimator minimizes the sum of trimmed squared residuals and is given by:

$$\hat{\beta}_{LTS} = Min \sum_{i=1}^{n} e_i^2 \tag{43}$$

such that $e^2_{(1)} \leq e^2_{(2)} \ldots \leq e^2_{(n)}$ in equation 43 are the ordered squares residuals and h is defined in the range $\frac{n}{2} + 1 \leq h \leq \frac{3n+p+1}{4}$, with n and p being sample size and number of parameters respectively. The largest squared residuals are excluded from the summation in this method.

The previously proposed LTS algorithms grows too much with size of the data hence the proposition of the new algorithm called FAST-LTS (Rousseeuw & Van, 2006). For small data sets FAST-LTS typically finds the exact LTS, whereas for larger data sets it gives more accurate results than existing algorithms for LTS and is faster by orders of magnitude (Rousseeuw & Van, 2006).

Despite the limitation of relative efficiency of 37 percent and low convergence rate, LMS estimators can highly influence the calculation of the much more efficient MM estimators by providing initial estimates of the residuals (Bagheri et al., 2010).

2.6.4 MM Estimators

In the S-estimator, if k = 5.182, the Gaussian efficiency rises to 96.6 percent, but the breakdown point drops to 10 percent and to cope with this, Yohai (1987) introduced MM-estimators that combine a high breakdown point and a high efficiency (Yohai, 1987; Verardi & Croux, 2009). This is a special type of M-estimator (Yohai, 1987). It combines high breakdown value estimation and M estimation (Chen, 2002). They concurrently

possess the following qualities: When the mistakes are distributed normally and their breakdown point is 0.5, they are both (i) very effective and (ii) highly efficient (Chen, 2002; Ayinde et al., 2015). It was among the first robust estimators to have these two properties simultaneously (Ayinde et al., 2015). A three-stage process is used to define the MM-estimates (Ayinde et al., 2015). In the first stage an initial regression estimate is computed which is consistent robust and with high breakdown-point but not necessarily efficient. The residuals from the initial estimate are used to compute an M-estimate of the errors scale in the second step. Finally, in the third stage, a correct redescending psi-function-based M-estimate of the regression parameters is computed (Yohai, 1987). MM-estimator $\hat{\beta}$ defined as a solution to:

$$\sum_{i=1}^{n} x_{ij}(\phi_1)(\frac{y_i - x_i\beta_i}{s_n})x_i \tag{44}$$

where j=1,2,...,p, $\phi_1(\mu) = \frac{\partial\rho_1(\mu)}{\partial\mu}$

2.7 Robust diagnostic statistics measures

The diagnostics which are based on the mean regression estimates are not efficient and cannot detect correctly swamping and masking effects (Türkan et al., 2012). Outlier swamping effect happens where non outliers are made to appear to be outliers while masking effect happens where outliers conceal one another (Jajo, 2005). Robust regression is an appropriate substitute for the OLS and ML when there are influential observations (Bagheri et al., 2010). Therefore, Robust version of diagnostics were proposed to identify outliers. To create a diagnostic tool for outlier detection that may be resistant to masking or swamping, some studies suggested divide the robust residuals (residuals from robust

fit) in the numerator by the robust scale estimate in the denominator (Jajo, 2005). This technology, which is advertised as being simple to use and robust in its application, may detect single or several outliers without experiencing masking or swamping issues.

According to Rousseeuw & Hubert (2011); Iglewicz & Martinez (1982), the residual for the robust estimators is given by;

$$Z_{R,i} = \frac{y_i - median_{j=1,\ldots,n}(\hat{y_j})}{median_{i=1,\ldots,n}|y_i - median_{j=1,\ldots,n}(\hat{y_j})|} \tag{45}$$

The outlying observation has robust score, $Z_{R,i}$ of greater than 2.

It is proposed to use the Huber-M estimator of β instead of $\hat{\beta}$, which is the least square estimator, and the robust scale estimate of σ instead of $\hat{\sigma}$ which is the least square estimator in OLS/ML Cook's distance, DFFITS and DFBETAS to obtain a robust versions (Türkan et al., 2012). Therefore, the Robust version of Cook's Distance, RD_i, DFFITS, $RDFFITS_i$ and DFBETAS, $RDFBETAS_i$ can be defined as follows:

Therefore, the Robust version of Cook's Distance, RD_i is defined by:

$$RD_i = \frac{(\hat{\beta}_r - \hat{\beta}_{r(-i)})^T (X^T X)(\hat{\beta}_r - \hat{\beta}_{r(-i)})}{p\hat{\sigma}_r^2} \tag{46}$$

where $\hat{\beta}_r$ in 46 is the robust estimation of β and $\hat{\sigma}_r^2$ the robust scale estimation of σ.

Robust DFFITS, $RDFFITS_i$ is given by:

$$RDFFITS_i = \frac{|X_i^T(\hat{\beta}_r - \hat{\beta_{r(-i)}})|}{\hat{\sigma}_{r(-i)}\sqrt{h_i}} \tag{47}$$

where h_i in equation 47 is the i-th diagonal element of hat matrix and $\hat{\sigma}_{r(-i)}$ the robust scale estimation of σ calculated from the data set without i-th observation.

And Robust DFBETAS, $RDFBETAS_i$ is defined by:

$$(RDFBETAS)_{k(i)} = \frac{b_r - b_{r(i)}}{\sqrt{MSE_i C_{kk}}} \tag{48}$$

where $k = 0, 1, 2, ..., p-1$ in equation 48

2.8 Model goodness of fit measures

Standard Errors (SE) and Bias measures were used. Lower values of bias and standard errors indicate a better fit. Standard Error, $SE_{\hat{\beta}}$ and Bias are given by;

$$SE_{\hat{\beta}} = \frac{\sigma}{\sqrt{(n)}} \tag{49}$$

$$Bias(\beta) = E(\hat{\beta}) - \beta \tag{50}$$

where n in equation 49 and 50 is the sample size.

2.9 Application of mean, quantile, and robust regression methods and diagnostic statistics to real life data sets

Numerous studies (Notapiri et al., 2022; Doganer et al., 2021; Ayinde et al., 2015; Atkinson, 1982) used robust regression methods and diagnostic statistics to real life data set.

In order to overcome outlier problem, Notapiri et al. (2022) used robust regression with

S-estimator to model crime rate in Indonesia during the COVID-19 pandemic. The study used 7 steps to obtain S-estimator, this approach involved a detailed iterative process to obtain accurate estimates despite the presence of outliers, highlighting the robustness of their method in handling real-world data complexities. Firstly, estimated $\hat{\beta}$ using OLS. Secondly, computed residual $\epsilon_i = y_i - \hat{y}_i$. Thirdly, computed $\hat{\sigma_s} = \frac{median|\epsilon_i - median(\epsilon_i)|}{0.6745}$ for the first iteration, $\hat{\sigma_s} = \sqrt{\frac{1}{nK}\sum_{i=1}^{n} w_i \epsilon_i^2}$ for the next iteration, with K=0.1995. Forthly, $mu_i = \frac{\epsilon_i}{\hat{\sigma_s}}$. Fifthly, weighted value w_i computed using Tukey's bisquare (tuning constant c=1.548). Then the estimation of $\hat{\beta_s}$ using WLS with weighted w_i. The steps 2 to 5 were repeated to obtain a convergent value of $\hat{\beta_s}$. The study did not report on the software used for data analysis. The Regression assumption tests were performed and results reported that the data was not normally distributed. The study reported that crime rate in Indonesia during the COVID-19 pandemic was influenced by the unemployment rate, poverty rate, GRDP per capita, population density and human development index.

Similarly, Doganer et al. (2021) used M-estimator to investigate the effects of changing hormone levels in pregnancy on cognitive perception levels in pregnant women aged 18 to 40 years. A total of 84 individuals, 42 pregnant and 42 healthy non-pregnant women (as control group) enrolled in the study. The study used Shapiro-Wilk test to test for the normality assumption in linear model. Robust regression analysis, M-estimator, was used for model estimations. Data analysis was performed in IBM SPSS and R 3.6.0 software. By using robust methods, the researchers were able to obtain reliable estimates that provided meaningful insights into the cognitive changes experienced by pregnant women. The study observed significantly lower cognitive scores in pregnant women compared to control group. The results were in agreement with previous studies. The study concluded that it is important to identify the responsible factors causing cognitive changes in pregnant

women and provide necessary support through out the period. This study demonstrated the utility of robust regression in medical research, where data often contain outliers due to biological variability.

Ayinde et al. (2015) compared the performance of OLS and robust M-stimator with robust MM, S and LTS estimators to determine the most efficient estimator. The models were fitted to three real life data sets; Longley data, Scottish Hills data and Hussein data and performance of diagnostic measures was compared. The study did not report the software used in data analysis neither did it report how the analysis was performed. The study reported diagnostics measures based on OLS do not give reliable estimates as compared to robust estimators, with MM estimators being more effective. In the study suggested that the performance of the robust version of the influential statistic is largely dependent on the root mean square error. Furthermore, it was reported that Cook's distance and DFFITS detected almost similar influential data points than DFBETAS across OLS, M, MM, S and LTS estimators. Their findings underscored the superiority of robust estimators in providing reliable results when dealing with datasets containing outliers. This study reinforced the importance of choosing appropriate statistical methods to ensure the accuracy and reliability of research findings in diverse fields.

There are limited studies in literature that used linear regression model to study risk factors of maternal anaemia. When modelling maternal anaemia, it can be applied when the response variable is haemoglobin level, continuous variable rather than when the haemoglobin levels are categorized. Most research in this area has favored logistic regression models due to their suitability for binary or categorical outcomes, such as the presence or absence of anemia. For example, Alem et al. (2023), Sunuwar et al. (2020),

Acharya et al. (2022) all utilized logistic regression to identify risk factors for maternal anemia1. These studies typically focus on whether or not a woman is anemic, which is a binary outcome, making logistic regression an appropriate choice.

However, Pasricha et al. (2010) highlighted that use of haemoglobin level avoids categorization of haemoglobin level which has age and ethnic ambiquities, particularly in children. By modeling haemoglobin levels as a continuous variable, linear regression avoids the need to categorize haemoglobin levels, which can introduce ambiguities related to age and ethnicity. This approach allows for a more nuanced understanding of the factors influencing haemoglobin levels and can provide more precise estimates of the effects of various predictors.

The preference for logistic regression in many studies may be due to the ease of interpretation and the straightforward nature of binary outcomes. Nonetheless, linear regression models offer significant benefits in terms of detail and accuracy, especially when dealing with continuous data like haemoglobin levels. This methodological choice can lead to more comprehensive insights into the risk factors of maternal anemia, as it captures the full range of haemoglobin levels rather than reducing the data to a binary outcome.

CHAPTER THREE

MATERIALS AND METHODS

3.1 Statistical methods

3.1.1 Mean regression and estimation

For a continuous response variable Y_i measuring haemoglobin level of i-th woman and X_{ij} being her bio-demographic and socio-economic characteristics, where $i = 1, 2, ..., n$ and $j = 0, 1, 2, ..., p$, a mean regression model estimates the conditional mean of Y given a set of explanatory variables X_{ij} (Sarstedt et al., 2019). In reference to Equation 1, Y_i is the dependent Hb variable; $X_{ij} = (X_{i0}, X_{i1}, X_{i2}, ..., X_{ip})$ the row vector of a set of independent variables observed on the i-th woman, with $X_{i0} = 1$; $\beta_j = (\beta_0, \beta_1, \beta_2, ..., \beta_p)^T$ is a column vector of regression parameters; and ϵ_i the model's error term. The values Y_i are measured indendently and their variance is constant (Peña & Slate, 2006). Further, $\epsilon_i \sim N(0, \sigma^2)$. The linear combination of the variables X_{ij} directly describes the responsed Y_i. This is the reason the linear model in Equation 1 is called mean regression model (Sarstedt et al., 2019).

The maternal anaemia model of identifying determinants used can be given by

$$\begin{aligned} Hemoglobin - level = \beta_0 + \beta_1 * BMI + \beta_2 * Age + \beta_3 * distance + \beta_4 * Education \\ + \beta_5 * Residence + \beta_6 * Wealth - Index + \beta_7 * Gravidity \\ + \beta_8 * Current - preg - duration + \epsilon_i \end{aligned} \tag{51}$$

The ML/OLS estimation method was used to estimate the parameters, β and σ, in the model. The parameter β estimation formula is expressed by:

$$\hat{\beta_N} = (X^T X)^{-1} X^T Y \tag{52}$$

where $N = 0, ..., 8$

Based on the normality assumption for the error term, and hence the response Y, the regression parameters β of the linear model in Equation 1 are estimated using the maximum likelihood estimation technique given by:

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} exp\left[-\frac{1}{2\sigma^2}(Y_i - X_{ij}\beta_j)^2\right]. \tag{53}$$

The solutions are the values of β at the maximum turning point of the log-likelihood function that is obtained from the likelihood function in Equation 53. This is obtained by taking first partial derivatives of the log-likelihood function and equate the result to zero to solve for values of β. Such process gives the following maximum likelihood estimator:

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}, \tag{54}$$

where $\mathbf{X}$ is $n \times p$ design matrix and $\mathbf{y}$ is $n \times 1$ vector of responses. The estimation of β in Equation 1 can also be done through a ordinary least squares approach, where the the solutions are the ones that give the minimum of sum of squares of the model errors. The method yields similar estimates of β as those obtained through maximum likelihood approach given in Equation 54. The value $\hat{\beta}$ stands for amount of change in Y as a result of a unit increase in the value of X_j, holding other covariates constant. Both

maximum likelihood and least squares estimation methods extend to generalised linear models (Dobson & Barnett, 2008).

The interpretation of the parameters in the mean regression vary slighly depending on the type of predictor variable, continuous or categorical. For continuous predictor variable, coefficient β represents the change in the mean of the response variable for a one-unit change in that continuous variable, holding all other variables constant. For categorical predictor variable, the interpretation depends on how they are coded. Such that the coefficient β for each category represents the difference in the mean of the dependent variable compared to a reference category.

3.1.2 Quantile regression and estimation

When the ratio-scale data are skewed and the normality assumptions for the errors and responses do not hold true, then the nonparametric quantile regression model becomes an immediate choice to model the data (Fox, 2002; Koenker, 2017; Čížek & Sadıkoğlu, 2020). The relationship between X and Y is estimated without assuming any specific probability distribution for Y. The linear relationship between X and Y is estimated at particular quantile of Y, providing information not available through mean regression methods (Rodriguez & Yao, 2017). Becasue the quantile model in Equation 26 describes the regression relatioship at a chosen quantile of Y, it performs better than the mean model in Equation 1, when the data are skewed (Rodriguez & Yao, 2017; Waldmann, 2018). The quantile regression extends the location shift model by determining the effect of covariates on the shape and scale of the entire response distribution (Waldmann, 2018).

The regression coefficients in the quantile model in Equation 26 are estimated by minimising equation 28. The estimated coefficient from Equation 28 represents the change in

the response variable at a specific quantile corresponding to a unit change in the covariate (Jamee et al., 2022). The quantile regression provides more comprehensive understanding of the relationship between the explanatory and response variables, especially when the relationship is not constant across different quantiles (Rodriguez & Yao, 2017).

The model was fitted using STATA version 17.0. The package "qreg" was used to fit the quantile regression models for different levels, τ; 0.25 (25th), 0.5 (50th), 0.75 (75th) and 0.90 (90th).

In Quantile Regression, the interpretation of parameters for both continuous and categorical predictors variables is focused on estimating the conditional quantiles of the dependent variable. The parameters in quantile regression provide information about how the independent variables affect different quantile levels of the response variable distribution. The contionuous predictor variable, the coefficient β indicates how a one-unit change in the independent variable affects the conditional quantile of the dependent variable. In quantile regression, the interpretation is about the impact on a specific quantile of interest, as opposed to mean regression, where the coefficient represents the change in the mean. For categorical predictor variable, the coefficients β indicate the difference in the conditional quantile of the dependent variable compared to a reference category.

3.1.3 Robust linear regression and estimation

Robust regression for a linear model in Equation 1 generally refers to a set of model estimation techniques that relax the parametric assumptions of the model off the usual estimation techniques (Huber, 1973). The robust regression method outperforms MLE and OLS when outliers are located along the y-axis rather than the x-axis in the model (Chen, 2002). One of the methods called maximum likelihood type estimation (M-estimator) ap-

proaches the estimation by minimising a function of the errors called loss function, denoted by $\rho(.)$ rather than the sum of squared errors (Chen, 2002). Its objective function is given by:

$$min \sum_{i=1}^{n} \rho\left(\frac{e_i}{s}\right) = min \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_{ij}\beta_j}{s}\right), \tag{55}$$

where s is an estimate of scale, often estimated by median absolute deviation (MAD) of the residuals, i.e. $s = \frac{median|e_i - median(e_i)|}{0.6745}$. The loss function $\rho(.)$ has the following properties: $\rho(e) \geq 0$, $\rho(0) = 0, \rho(e) = \rho(-e)$, and $\rho(e_i) \geq \rho(e_i^T)$ for $|e_i| = |e_i^T|$ (Rousseeuw & Hubert, 2011). The solutions from Equation 55 are obtained using the Gauss-Newton iterations on the score function:

$$\sum_{i=1}^{n} (\phi)\left(\frac{Y_i - X_{ij}\beta_j}{s}\right) X_{ij} = 0, \tag{56}$$

where ϕ in Equation 56 is a partial derivative of ρ with respect to β (Rousseeuw & Hubert, 2018).

Another robust regression method used is the Schweppe's estimator (S-estimator), which is known to have a high-breakdown point and can withstand the influence of a large presence of outliers in regression parameter estimation (Chen, 2002). The S-estimator is derived from a scale statistic corresponding to residuals of M-estimator. For a set of residuals $e_1, e_2, ..., e_n$, the scale estimate min $s(e_1(\beta), e_2(\beta), ..., e_n(\beta))$ is the solution of:

$$min \sum_{i=1}^{n} \rho\left(\frac{Y_i - X_{ij}\beta_j}{s}\right), \quad and \quad \hat{\sigma}_s = \sqrt{(nK)^{-1} \sum_{i=1}^{n} w_i e_i^2}, \tag{57}$$

where $K = 0.199$ is the expectation value of $\rho(.)$ for a standard normal distribution, w_i is the weighting term, and the estimation proceeds using the score function as in Equation

56. The S-estimator in Equation 57 resists contamination of up to 50 percent of outliers, hence its breakdown point is 50 percent (Verardi & Croux, 2009).

Studies also use the least trimmed squares (LTS) estimator to flexibly estimate the regression parameters in linear models, which has also high-breakdown point (Rousseeuw & Hubert, 2018). The method eliminates regression parameters by running the analysis on trimmed or winsorized distributions without outliers (Rousseeuw & Hubert, 2011). The LTS-estimator is known to be more efficient than the S- or M-estimators (Rousseeuw & Hubert, 2018). The LTS-estimator is a solution that minimizes the sum of trimmed squared residuals, and it is given by:

$$\hat{\beta}_{LTS} = min \sum_{i=1}^{l} e^2_{(i)}, \tag{58}$$

where $e^2_{(1)} \leq e^2_{(2)} \ldots \leq e^2_{(n)}$ are the ranked squares residuals, $l = [n(1-\alpha)+1]$ is the number of observations included in the computation of the estimator, and α the proportion of trimming that is performed (Rousseeuw & Van, 2006). The largest squared residuals are excluded from the summation for being suspected as outliers.

Finally, an improved special type of the M-estimator called MM-estimator combines achieving high-breakdown point and high efficiency in the estimation (Yohai, 1987; Verardi & Croux, 2009; Chen, 2002). The method estimates the parameters using S-estimator which minimises the scale of the residual from the M-estimator and then proceed with M-estimation (Chen, 2002). It is one of the few robust estimators having the two properties simultaneously (Rousseeuw & Hubert, 2011). The MM-estimator $\hat{\beta}$ solutions are

obtained from the function:

$$\sum_{i=1}^{n} X_{ij}(\phi_1)\left(\frac{Y_i - X_{ij}\beta_j}{s_n}\right) = 0, \tag{59}$$

where $\phi_1(\beta) = \frac{\partial \rho_1(\beta)}{\partial \beta}$, with ρ from S-estimator. The MM-estimates from Equation 59 are obtained in a sequential manner, where an initial regression estimate is computed first to obtain consistent robust and high-breakdown point estimate, but which is not necessarily efficient. Then, from the initial estimate, M-estimates of the errors scale are computed in the second step. This is followed by computation of a correct redescending ϕ-function-based M-estimate of the regression parameters in the third stage (Yohai, 1987).

Parameter interpretation focuses on estimating the relationship between response and predictor variables while downweighting the impact of extreme values. Parameter estimates provide a robust estimation of the linear relationship between response and covariates by considering a subset of data points.

3.2 Outlier detection statistics for mean, quantile and robust regression methods

3.2.1 Analysis of outliers in mean regression

Anomolous data and outlier observations tend to distort and bias the conclusions from regression models, and need to be dealt with accordingly (Kaombe & Manda, 2023b,a; Kaombe, 2024). The raw residual given by equation 30 was employed in the study. The raw residual in Equation 30 measures the disagreement between the observed value of the response Y_i and the fitted value $\hat{Y}_i$ for the i-th subject (Kaombe, 2024). The larger the value of e_i, the poor the fit of the model to the i-th observation, and hence the higher

chances that the observation is an outlier in the model. While small values close to zero show high agreement between the fitted and observed values, hence a better fit (Kaombe & Manda, 2023b). If the responses Y_i were skewed, it becomes highly likely that the raw residual in Equation 30 will also be skewed. This affects assessments of outlier observations on both sides of the regression model. For this reason, studentised or standardised residual is often used to symmetricise the values of the residual and make the outlier assessment easier (Kaombe & Manda, 2023b). The standardised (Studentised) residual in Equation 33 is widely used to assess outliers. In most cases, the assessment of outlier observations is done graphically by plotting the residual in Equation 30 or 33. Boxplots of the residual can also help in analysing the unusual subjects. Other transformations of the raw residual exist in literature depending on the goal of analysis (Kaombe & Manda, 2023b).

Now, when outliers are detected in the model, various analyses follows. If the analyst is interested to know the source of outlierness so as to inform policy decisions from the data analysis or improve the data management, then the data back-inspection is done to further describe the outlier observations (Kaombe et al., 2023; Kaombe, 2024). If the researcher intends to improve the modelling, then influence analysis follows to estimate the impact of the outliers on regression coefficient estimates (Kaombe & Manda, 2023a). This is done using various statistics that analyse effect of deleting the outlier observation from the data. One such measure is the difference in beta standardised (DFBETAS) given by Equation 40. DFBETAS values calculated from observations decrease proportionally as the number of observations increases (Bahadir et al., 2014). Thus when the sample size is big, the DFBETAS in Equation 40 has limited sensitivity in influential points detection, but it is most effective in small sample sizes and outlier percentages (Oyeyemi et al., 2017). An observation with DFBETAS 40 that is greater than 1 for small dataset

or greater than $\frac{2}{\sqrt{n}}$ for large dataset is considered influential (Belsley et al., 2005).

3.2.2 Outlier analysis in quantile regression

A counterpart raw residual for a quantile linear model is given by:

$$e_{Q,i} = Q_\tau(Y_i) - \hat{Q}_{\tau|X}(Y_i), \tag{60}$$

where $\hat{Q}_{\tau|X}(Y_i)$ is predicted τ-th conditional quantile of the dependent variable Y_i at a specified quantile level τ given the covariates X, and $Q_\tau(Y_i)$ is the observed τ-th quantile of Y_i. Large values of the residual 60 correspond to outlier candidates. Similarly, the Studentized residual for the quantile model is given by:

$$r_{Q,i} = \frac{e_{Q,i}}{\hat{\sigma}\sqrt{1-h_{ii}}}, \tag{61}$$

where the quantity $\hat{\sigma}$ is as defined before, and $h_{ii} = X_i^T(X^T W_\tau X)^{-1} X_i$ is the leverage of i-th observation on the fitted value, where W_τ are the weights for each observation (determined by the robust estimator). Large Studentized residual in Equation 61 suggest potential outliers. Another useful transformation of the raw residual for quantile linear model is the jacknife residual given by:

$$J_{Q,i} = e_{S,i}\sqrt{\frac{n-p-1}{n-p-e_{Q,i}^2}}, \tag{62}$$

where $e_{Q,i}$ is the raw residual. The jacknife residual in Equation 62 examine the influence of individual point on the quadratic error of the prediction. The follow-up DFBETAS for assessment of influence is defined in a similar manner as in mean regression models.

3.2.3 Detecting outliers in robust regression model

The diagnostic statistics based on the mean regression estimates are limited in terms of dealing with swamping and masking effects (Türkan et al., 2012). Swamping effect means a good observation being wrongly identified as an outlier because of the presence of another clean subset of the data (Jajo, 2005). On the other hand, masking effect implies that an outlier is undetected because of the presence of another competing outlier (Jajo, 2005). Robust regression solves this by producing estimates with high-breakdown point (Rousseeuw & Hubert, 2011). As such, the robust regression diagnostics tend to be resistant to masking or swamping effects (Jajo, 2005). A standardised residual for the robust model is given by:

$$Z_{R,i} = \frac{0.6745(Y_i - \hat{Y}_i)}{median|(Y_i - \hat{Y}_i)_i - median((Y_i - \hat{Y}_i))|}, \tag{63}$$

where $\hat{Y}_i$ is a fitted value obtained from the robust regression method used and the denominator is the MAD of $(Y_i - \hat{Y}_i)$. The outlier observation has robust residual score, $Z_{R,i}$ in Equation 63 that is greater than 2 or less than minus 2 (Rousseeuw & Hubert, 2011; Türkan et al., 2012). The robust DFBETAS is defined and interpreted in a similar way as in mean regression.

3.3 Simulation scheme

A simulation study was carried out to analyse perfomance of the reviewed residuals for mean, quantile and robust regression methods in detecting outlier observations in a data

set. A linear model given below was used to generate the data:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \tag{64}$$

where $\epsilon_i \sim N(0,1)$, $X_{i1} \sim N(2.3, 0.5)$, $X_{i2} \sim N(8, 2.4)$, $\beta_0 = 2.1$, $\beta_1 = 0.7$ and $\beta_2 = 0.9$. Samples of size $n = 50$, and $n = 500$ observations were generated, and each sample was redrawn 100 times. The sample sizes were chosen to provide a comprehensive analysis of the performance of the residuals under different conditions. Larger sample sizes generally offer more reliable and robust statistical estimates. By including n=500, the study ensures that the results are not solely dependent on small sample behavior, which can sometimes be erratic or less reliable.

To ensure reproducibility, STATA command "set seed 12345 + simulation number" was used to set up and draw the data up to 100 simulations. Then perturbations were introduced to quarter of the observations in each data set generated by the model in Equation 64 as follows: where $\epsilon_i \sim N(-7.8, 22.1)$, $\beta_0 = 15$, $\beta_1 = 6$ and $\beta_2 = 10$ to observe the performance of the model in estimating the regression parameters. The final set of perturbations were introduced to only first five observations using the same parameters. The rest observations were generated based on model 64. This was done to assess the outlier detection ability by each model.

The three modelling methods: mean, quantile and robust regression were fitted to the data and their diagnostic statistics analysed. The efficiency of the three modelling methods was judged using bias of estimated parameters, calculated by:

$$bias(\hat{\beta}_j) = E(\hat{\beta}_j) - \beta_j, \tag{65}$$

where $E(\hat{\beta}_j)$ was the average of $\hat{\beta}_j$ in 100 simulations, β_j the original parameter value as in model 64, $j = 0, 1, 2$. An efficient model was the one with the bias in Equation 65 close to zero. The model efficiency was also assessed using effect sizes $\hat{\beta}$ and their standard errors, $\sqrt{var(\hat{\beta})}$. The smaller the standard error the more accurate the estimates from a particular model. The sensitivity of each model to outlier observations was judged by the number of times out of 100 simulations the model's outlier residual as per Section 2.5 detected the 5 individuals generated with perturbed data (Kaombe & Manda, 2023b). All the analyses were performed using STATA version 17 and code is given in Appendix 1.

3.4 Application to maternal anaemia data

The study further analyzed secondary maternal anaemia data, collected from the 2015-2016 Malawi Demographic Health Surveys (MDHS), inorder to compare the performance of the three modelling methods using real data. Demographic Health Surveys uses a cross-sectional study and cluster sampling to collect data from the individuals in the sample frame defined. The survey data was collected between 19th October 2015 and 17th February 2016. The 2015-16 MDHS is the fifth Demographic and Health Survey conducted in Malawi since 1992. Part of the purpose of the data collection was to provide an overview for monitoring maternal and child health, and to provde the nation's health experts with data they needed to carry out additional research on the subject. The data access permission was provided by Measure DHS Program through the website (https://dhsprogram.com/data/available-datasets.cfm).

The stratified two stage cluster sampling design was used. The Malawi National Statistical Office (NSO) provided the sampling frame for the 2015–16 MDHS, which was derived

from the 2008 Malawi Population and Housing Census (MPHC). The primary sampling units were the census standard enumeration areas (SEAs), and the secondary sampling units were the households. SEAs were stratified in terms of rural and urban areas which yielded to 56 sampling strata. In the first selection stage, 850 SEAs, comprising of 173 in urban and 677 in rural areas (stratum or SEAs), were selected using a probability proportional to the SEA size. In the second selection stage, fixed number of 30 households per urban cluster and 33 per rural cluster were selected with an equal probability systematic selection from the newly created household listing. A representative total sample of 27,516 households was selected for the 2015-2016 MDHS. The 2015-2016 MDHS data collection was by the questionnaire. There were four questionnaires; household, woman, men, and biomarker questionnaires. Computer-assisted personal interviewing (CAPI) data collection approach was used.

All women aged 15-49 who were either permanent residents of the selected households or visitors who stayed in the household the night before the survey were eligible to be interviewed. In the subsample of households selected for the male survey, anaemia testing was performed among eligible women who consented to being tested. Households that were successfully interviewed were 26361, yielding a response rate of 99 percent. Eligible women that were successfully interviewed were 24562, yielding a response rate of 98 percent.

Study used 21,935 reproductive women from 15-49 years who had haemoglobin level known to assess performance of robust regression methods and diagnostic statistics in linear models. In this study, a woman's body mass index, her education, place of redicence, fertility rate, wealth, duration of pregnancy (if pregnanct), distance to health facility,

and age were used as covariates to describe Hb levels. The mean, quantile, and robust regression models presented in Section 2.2 were fitted on the data and efficiency of each model and its sensistivity to outliers analysed. Data cleaning and analysis were done using STATA 17.0, the code is provided in Appendix 1.

CHAPTER FOUR

RESULTS

4.1 Introduction

This chapter contains results and interpretation of the simulation study and maternal anaemia data observing the performance of the three methods interms of efficiency (model estimates quality) and effectiveness (outlier and influential points detection). In these analyses, the level of significance was 0.05. Hence, all null hypotheses were rejected if the p-value of the test was less than 0.05. This section begins by presenting results of the simulation study. The study further analyzed maternal anaemia data from the Malawi DHS 2015-16 for 21,935 reproductive women aged 15-49 who had known haemoglobin levels.

4.2 Simulation results

4.2.1 Simulation results on estimates and standard errors of each model

The results in Table 1 indicated that in unperturbed data, all the three models showed relatively similar and accurate estimates with small standard errors. However, for perturbed data, the conventional linear model with maximum likelihood or least squares estimation showed significantly higher standard errors, on average, indicating high sensitivity to outlier observations in the data. In contrast, the robust methods like M-, MM-, S-, and LTS- estimators maintained more stable estimates with lower standard errors even in the presence of outliers. This demonstrated that the robust regression methods were more resistant to the influence of the outlier observations in the parameter estimates than the

conventional linear model and the quantile model. Quantile regression models (Q25, Q50, Q75, Q90) also showed varying degrees of robustness, with higher quantiles being more affected by perturbations and the 25th percentile model having smallest standard errors. Further, much reduced standard errors were observed in large samples of 500 observations for each model, while the trend of estimates remained similar between the models.

Table 1: Average parameter estimates and standard errors in 100 simulations for the robust, quantile, and mean regression models, with and without perturbations.

		Unperturbed data			Perturbed data		
n	Model	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$
50	LM-MLE	2.08 (0.85)	0.72 (0.29)	0.89 (0.12)	2.78 (35.90)	3.10 (12.5)	3.03 (2.53)
	Q25	1.41 (1.30)	0.70 (0.45)	0.91 (0.10)	1.75 (1.51)	0.70 (0.53)	0.90 (0.11)
	Q50	2.02 (1.10)	0.71 (0.40)	0.90 (0.14)	2.43 (41.35)	0.75 (14.48)	0.90 (2.91)
	Q75	2.66 (1.37)	0.75 (0.45)	0.90 (0.14)	11.82 (101.3)	4.56 (35.88)	4.45 (7.52)
	Q90	3.21 (1.37)	0.71 (0.48)	0.90 (0.15)	18.29 (65.53)	6.23 (23.04)	9.10 (5.06)
	Robust M	2.08 (0.84)	0.70 (0.32)	0.90 (0.11)	2.96 (2.21)	0.80 (0.74)	0.90 (0.16)
	Robust MM	2.03 (0.96)	0.68 (0.33)	0.91 (0.11)	2.15 (0.96)	0.69 (0.33)	0.90 (0.08)
	Robust S	1.98 (1.24)	0.75 (0.39)	0.90 (0.12)	2.09 (1.34)	0.71 (0.48)	0.90 (0.09)
	Robust LTS	1.99 ()	0.72 ()	0.92 ()	2.19 ()	0.69 ()	0.89 ()
500	LM-MLE	2.07 (0.28)	0.71 (0.12)	0.90 (0.02)	0.84 (10.86)	2.89 (3.77)	3.30 (0.78)
	Q25	1.40 (0.34)	0.71 (0.12)	0.9 (0.03)	1.58 (0.42)	0.72 (0.21)	0.91 (0.03)
	Q50	2.10 (0.33)	0.68 (0.12)	0.90 (0.03)	2.50 (0.50)	0.70 (0.18)	0.90 (0.04)
	Q75	2.71 (0.36)	0.7 (0.13)	0.90 (0.03)	-0.41 (52.42)	4.56 (18.24)	4.28 (3.88)
	Q90	3.30 (0.45)	0.71 (0.16)	0.91 (0.05)	10.33 (18.91)	7.18 (6.64)	10.01 (1.37)
	Robust M	2.07 (0.26)	0.72 (0.11)	0.90 (0.02)	2.75 (0.57)	0.72 (0.20)	0.90 (0.04)
	Robust MM	2.07 (0.28)	0.70 (0.10)	0.90 (0.02)	2.07 (0.32)	0.70 (0.12)	0.89 (0.04)
	Robust S	2.09 (0.52)	0.68 (0.18)	0.90 (0.08)	2.06 (0.41)	0.71 (0.14)	0.89 (0.04)
	Robust LTS	2.12 ()	0.68 ()	0.90 ()	2.11 ()	0.69 ()	0.90 ()

4.2.2 Bias of regression coefficient estimates from each model

The bias results in Table 2 showed that in unperturbed data, all the models had relatively similar and low bias of estimations. However, when the outliers were introduced in the sample, the robust regression methods of all types and the first and second quartile (Q25 and Q50) models produced best estimates with smallest bias. The biases were large in

linear and 75th and 90th percentile models for the data that contained outliers.

Table 2: Bias of estimation for regression parameters using the robust, quantile, and mean models based on 100 simulations with and without perturbations.

		Unperturbed data			Perturbed data		
n	Model	$bias(\hat{\beta}_0)$	$bias(\hat{\beta}_1)$	$bias(\hat{\beta}_2)$	$bias(\hat{\beta}_0)$	$bias(\hat{\beta}_1)$	$bias(\hat{\beta}_2)$
50	LM-MLE	-0.02	0.02	-0.01	0.68	2.40	2.13
	Q_{25}	-0.69	0.00	0.01	-0.35	0.00	0.00
	Q_{50}	-0.08	0.01	0.01	0.33	0.05	0.00
	Q_{75}	0.56	0.05	0.00	9.72	3.86	3.55
	Q_{90}	1.11	0.01	0.00	16.19	5.53	8.20
	Robust-M	-0.02	0.00	0.00	0.86	0.10	0.00
	Robust-S	-0.12	0.05	0.00	-0.01	0.01	0.00
	Robust-MM	-0.07	-0.02	0.01	0.05	0.01	0.00
	Robust-LTS	-0.11	0.02	0.02	0.09	-0.01	-0.01
500	LM-MLE	-0.03	0.01	0.00	1.26	2.19	2.40
	Q_{25}	-0.70	0.01	0.00	-0.52	0.02	0.01
	Q_{50}	0.00	-0.02	0.00	0.40	0.00	0.00
	Q_{75}	0.61	0.00	0.00	-2.51	3.86	3.38
	Q_{90}	1.20	0.01	0.01	8.23	6.48	9.11
	Robust-M	-0.03	0.02	0.00	0.65	0.02	0.00
	Robust-S	-0.01	-0.02	0.00	-0.04	0.01	-0.01
	Robust-MM	-0.03	0.00	0.00	-0.03	-0.01	-0.01
	Robust-LTS	0.02	-0.02	0.00	0.01	-0.01	0.00

4.2.3 Outlier detection by each model in 100 simulations with perturbed first 5 observations

The simulation results in Table 3 showed that the outlier residuals of all the models performed equally in detecting the five outlier observations in the data set with success rates close to 100%. There was one exception for the residual of the 90th percentile model in small sample sizes of 50 in which it had success rates of less than 15% for all the five outliers.

Table 3: Number of times out of 100 simulations an outlier observation has been detected by a residual of robust, quantile, and mean regression models, perturbed case.

		No. of times outlier is detected				
n	Model	*obs*.1	*obs*.2	*obs*.3	*obs*.4	*obs*.5
50	LM-MLE	98	98	95	97	97
	Q_{25}	100	100	100	100	100
	Q_{50}	100	100	100	100	100
	Q_{75}	100	100	100	100	100
	Q_{90}	13	12	11	11	13
	Robust-M	99	100	100	100	100
	Robust-S	99	100	100	100	100
	Robust-MM	99	100	100	100	100
	Robust-LTS	99	100	100	100	100
500	LM-MLE	100	100	100	99	99
	Q_{25}	100	100	100	99	99
	Q_{50}	100	100	100	99	99
	Q_{75}	100	100	100	99	99
	Q_{90}	99	100	100	99	99
	Robust-M	100	100	100	99	99
	Robust-S	100	100	100	99	99
	Robust-MM	100	100	100	99	99
	Robust-LTS	100	100	100	99	99

4.3 Maternal anaemia data results

subsectionMaternal anaemia reasults for the 2015-16 MDHS data This section presents the regression estimates and outlier residual results for each of the models reviewed in Section 2. The women data had average Haemoglobin level of 12.53 g/dl, and standard deviation of 1.74 g/dl. The Hb range was 23 g/dl minus 2 g/dl. The Hb data were skewed to the left, with a coefficient of -0.52. Thus, there were more Hb measurements below average than above it.

4.3.1 Regression model estimates results for the maternal anaemia data

The results in Table 4 showed that the directions of effect sizes were generally similar across all the models. However, the sizes of standard errors were smallest in the linear, 50th percentile, M- and MM- robust regression models. The standard errors were largest in the 25th percentile, 75th percentile, 90th percentile, and robust S-estimator models. The LTS- model does not process standard errors. Further, the model-based average Haemoglobin (Hb) level in women was 13.7 g/dl using M- and MM-estimator robust models, and 12.8 g/dl using the 25th percentile model. Thus, the 25th percentile model intercept was consistent with the raw data average Hb estimate.

The results also showed that staying in rural area increased Hb levels by 0.13 compared to urban area. Having primary and secondary education increased Hb levels by 0.24 and 0.19, respectively compared to no education. The Hb levels were not significantly different between Women with higher education and those with no education. Furthermore, a unit increase in age of pregnancy significantly reduced Hb levels by 0.31 g/dl. In addition, having normal, overweight, and obese body mass index increased Hb by 0.27, 0.42, and 0.62 g/dl, respectively. Women living in rich households had reduced Hb levels by 0.07 g/dl compared to those from poor household, but there was no difference in Hb between women from middle and poor households. It was also shown that women drinking from safe water sources had reduced Hb by 0.10 g/dl compared to those using unsafe sources. The fertility rate, distance from clinic, and age of a woman were not associated with Hb levels.

Table 4: Regression parameter estimates by the robust, quantile, and mean regression models from 2015-16 MDHS.

	Mean Reg	Quantile Reg			
Covraiate	$\hat{\beta}(SE, pv)$	$\hat{\beta}_{Q_{25}}(SE, pv)$	$\hat{\beta}_{Q_{50}}(SE, pv)$	$\hat{\beta}_{Q_{75}}(SE, pv)$	$\hat{\beta}_{Q_{90}}(SE, pv)$
Intercept	13.6 (0.14, <0.001)	12.8 (0.20, <0.001)	14.2 (0.16, <0.001)	14.5 (0.19, <0.001)	16.05 (0.19, <0.001)
Residence					
Urban*					
Rural	0.07 (0.04, 0.047)	0.07 (0.05, 0.215)	0.30 (0.04, <0.001)	0.06 (0.05, 0.260)	0.05 (0.05, 0.317)
Education					
None*					
Primary	0.23 (0.03, <0.001)	0.30 (0.04, <0.001)	0.20 (0.04, <0.001)	0.22 (0.04, <0.001)	0.25 (0.04, <0.001)
Secondary	0.16 (0.04, <0.001)	0.17 (0.07, 0.011)	0.20 (0.05, <0.001)	0.23 (0.06, <0.001)	-0.50 (0.15, 0.001)
Higher	-0.20 (0.11, 0.075)	0.07 (0.17, 0.688)	-0.00 (0.13, 1.00)	-0.26 (0.16, 0.101)	1.10 (0.97, 0.256)
Fertility Rate	-0.003 (0.01, 0.635)	-0.00 (0.01, 1.00)	-0.00 (0.01, 1.00)	0.01 (0.01, 0.243)	-0.05 (0.01, <0.001)
Pregnancy Dur	-0.29 (0.02, <0.001)	-0.30 (0.03, <0.001)	-0.40 (0.03, <0.001)	-0.27 (0.03, <0.001)	-0.40 (0.03, <0.001)
Clinic Distance					
Big problem*					
No problem	-0.04 (0.02, 0.072)	-0.07 (0.04, 0.063)	95.4 (0.03, 1.00)	0.08 (0.03, 0.017)	-0.05 (0.03, 0.134)
BMI					
Underweight*					
Normal	0.38 (0.05, <0.001)	0.30 (0.08, <0.001)	0.20 (0.06, 0.001)	0.27 (0.07, <0.001)	0.40 (0.07, <0.001)
Overweight	0.51 (0.05, <0.001)	0.47 (0.08, <0.001)	0.30 (0.06, <0.001)	0.38 (0.07, <0.00)	0.45 (0.07, <0.001)
Obese	0.52 (0.40, 0.199)	0.80 (0.09, <0.001)	0.48 (0.08, <0.001)	0.69 (0.58, 0.235)	0.65 (0.08, <0.001)
Wealth Index					
Poor*					
Middle	-0.03 (0.03, 0.328)	-0.07 (0.05, 0.151)	-0.10 (0.04, 0.007)	-0.08 (0.04, 0.083)	-0.15 (0.04, 0.001)
Rich	-0.09 (0.03, 0.002)	-0.10 (0.04, 0.020)	-0.00 (0.03, 1.00)	-0.05 (0.04, 0.196)	-0.05 (0.04, 0.210)
Age group					
15-24*					
25-49	-0.09 (0.04, 0.038)	-0.17 (0.06, 0.005)	-0.10 (0.05, 0.036)	-0.02 (0.06, 0.671)	0.20 (0.06, <0.001)
Water soucre					
Unsafe*					
Safe	-0.10 (0.03, 0.002)	-0.20 (0.05, <0.001)	-0.10 (0.04, 0.012)	-0.03 (0.05, 0.480)	-0.05 (0.04, 0.282)

	Mean Reg	Robust Reg			
Covraiate	$\hat{\beta}(SE, pv)$	$\hat{\beta}_M(SE, pv)$	$\hat{\beta}_S(SE, pv)$	$\hat{\beta}_{MM}(SE, pv)$	$\hat{\beta}_{LTS}(SE, pv)$
Intercept	13.6 (0.14, <0.001)	13.7 (0.16, 0.002)	15.0 (0.37, <0.001)	13.9 (0.19, <0.001)	14.7 ()
Residence					
Urban*					
Rural	0.07 (0.04, 0.047)	0.13 (0.04, 0.001)	0.43 (0.06, <0.001)	0.19 (0.04, <0.001)	0.71 ()
Education					
None*					
Primary	0.23 (0.03, <0.001)	0.24 (0.03, <0.001)	0.18 (0.04, <0.001)	0.23 (0.03, <0.001)	0.12 ()
Secondary	0.16 (0.04, <0.001)	0.19 (0.04, <0.001)	0.24 (0.07, <0.001)	0.22 (0.05, <0.001)	0.24 ()
Higher	-0.20 (0.11, 0.075)	-0.17 (0.11, 0.120)	-0.21 (0.13, 0.094)	-0.15 (0.11, 0.193)	-0.14 ()
Fertility Rate	-0.003 (0.01, 0.635)	-0.002 (0.01, 0.665)	0.01 (0.01, 0.184)	-0.001 (0.01, 0.858)	-0.001 ()
Pregnancy Dur	-0.29 (0.02, <0.001)	-0.31 (0.03, <0.001)	-0.56 (0.07, <0.001)	-0.34 (0.03, <0.001)	-0.60 ()
Clinic Distance					
Big problem*					
No problem	-0.04 (0.02, 0.072)	-0.001 (0.02, 0.973)	0.16 (0.04, <0.001)	0.05 (0.02, 0.062)	0.31 ()
BMI					
Underweight*					
Normal	0.38 (0.05, <0.001)	0.27 (0.05, <0.001)	0.08 (0.07, 0.274)	0.17 (0.05, 0.001)	0.49 ()
Overweight	0.51 (0.05, <0.001)	0.42 (0.05, <0.001)	0.19 (0.08, 0.012)	0.32 (0.05, <0.001)	0.41 ()
Obese	0.75 (0.06, <0.001)	0.62 (0.06, <0.001)	0.33 (0.08, <0.001)	0.51 (0.06, <0.001)	0.60 ()
Wealth Index					
Poor*					
Middle	-0.03 (0.03, 0.328)	-0.03 (0.03, 0.288)	0.06 (0.05, 0.212)	-0.02 (0.03, 0.525)	0.03 ()
Rich	-0.09 (0.03, 0.002)	-0.07 (0.03, 0.016)	0.12 (0.04, 0.004)	-0.02 (0.03, 0.426)	0.01 ()
Age group					
15-24*					
25-49	-0.09 (0.04, 0.038)	-0.07 (0.04, 0.052)	-0.09 (0.06, 0.131)	-0.06 (0.04, 0.155)	0.004 ()
Water soucre					
Unsafe*					
Safe	-0.10 (0.03, 0.002)	-0.10 (0.03, 0.002)	-0.12 (0.05, 0.007)	-0.10 (0.03, <0.001)	-0.23 ()

4.3.2 Assessment of outliers in the women Hb data

The box plots given in Figure 1 for the residuals of the applied models showed that all the methods consistently detected more outliers on the left side of the median value than the right. This confirmed that the Hb data were left-skewed, with some women in Malawi having extremely low Hb levels than the average (or being anaemic). This also explains why the 25th percentile model produced average Hb value that was consistent with the raw data estimate, as the model considered a group of women that were anaemic. The data inspection indicated that the models identified between 400 and 500 outliers in the data. Over half of the outliers were commonly detected by the models.

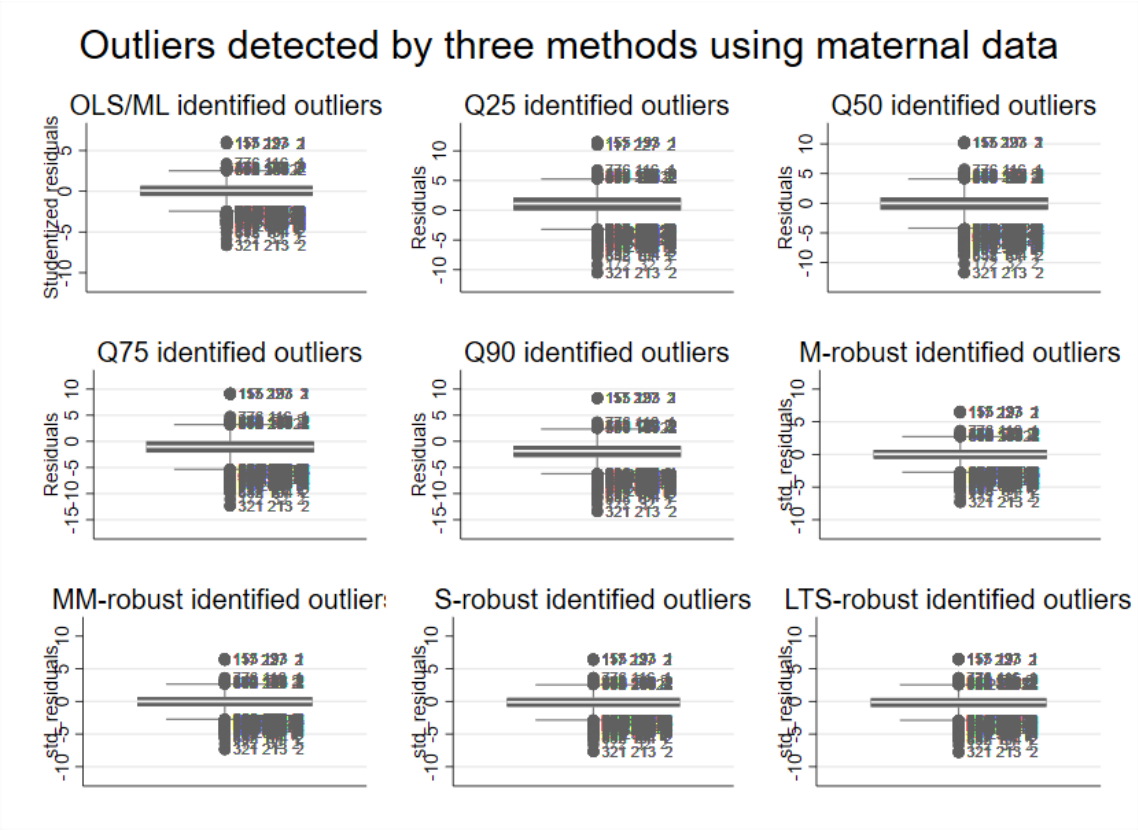

Figure 1: Plots of the outlier residual for each model using the women Hb data from 2015-16 MDHS.

A further inspection of the Hb data showed all the models detected between 400 and

500 outlier observations in the data, except the LTS which had over 1000 outliers, see Table 5. A considerable amount of the detected outliers were commonly identified by all the models. A large proportion of the outlier observations were those women who had extremely low Hb levels far from the normal range in the population.

Table 5: Exact common and uncommon outlier women ids and their residual signs detected by the diagnostics for robust, quantile, and mean regression models, 2015-16 MDHS.

Woman ID	LM-MLE	Q_{25}	Q_{50}	Q_{75}	Q_{90}	Robust-M	Robust-MM	Robust-S	Robust-LTS
102 240 2 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
321 213 2 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	+ve
321 213 2 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	+ve
321 213 2 3	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	+ve
397 163 2 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	+ve
397 163 2 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	+ve
479 185 2 1	+ve	+ve	+ve	+ve	+ve	+ve	+ve	+ve	-ve
479 185 2 2	+ve	+ve	+ve	+ve	+ve	+ve	+ve	+ve	-ve
479 185 2 3	+ve	+ve	+ve	+ve	+ve	+ve	+ve	+ve	-ve
612 89 1 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
612 89 1 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
612 89 1 3	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
612 89 1 4	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
638 203 2 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
638 203 2 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
638 203 2 3	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 3	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 4	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 5	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
681 287 1 6	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
743 63 2 1	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
743 63 2 2	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
743 63 2 3	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve	-ve
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total outliers	414	429	452	425	390	409	433	467	1133
Total -ve outliers	363	363	390	372	349	358	379	394	769
Total +ve outliers	51	66	62	53	41	51	54	73	364

CHAPTER FIVE

DISCUSSION, CONCLUSION AND RECOMMENDATION

5.1 Discussion

This study aimed to assess performance of mean, quantile, and robust regression methods in analysing correlates of women Haemoglobin levels in Malawi using simulations and real data applications. Through simulations, the study observed that each model's residual had the same capacity to detect the outliers, when they were present in a data set. This is the case since a residual statistic is defined within the assumptions framework of the respective model, hence it has to effectively track the unusual measurements in the model (Kaombe & Manda, 2023b; Kaombe, 2024). Further, it was shown that the linear, quantile and robust regression models performed with similar biases, in samples that had no outliers. But, in the presence of outlier observations, the robust regression methods and quantile model at 25th and 50th percentile produced best estimates that had smallest bias. The linear, 75th and 90th percentile models produced large bias estimates in data that had outliers. The presence of outliers in data usually skew the data, leading to violation of the normality assumptions of the linear model errors upon which the least squares and maximum likelihood estimation methods are based, hence causing the model to produce biased estimates (Sinha, 2004; Pérez et al., 2014). The robust and quantile regression methods bipass these strict assumptions to make the estimation through flexible nonparametric procedures that involve ranks of observations instead of their actual measurements or use a fraction of contaminated-free data to make estimates, and overcome

the impact of extremety of the measurements in the regression estimates (Mei Ling Huang & Tashnev, 2015; Geraci & Bottai, 2014; Yuen & Ortiz, 2017; Rousseeuw & Hubert, 2011). These results consolidates evidence that the supremacy of robust regression methods is in withstanding the impact of the outlier observation in the regression parameter estimates, and not in the detection of outliers themselves as observed in previous studies (Rousseeuw & Hubert, 2018, 2011; Santos, 2020).

When applied to women Haemoglobin data, the residuals for all the models reported considerable amount of outliers to the women's Hb data, most of whom were women who had extremely low Hb levels. This was consistent with the simulation results that showed that the diagnostic statistics for the three models had similar sensitivity to outlier observations in the data (Santos, 2020). The application showed that the directions of effect sizes were generally similar across the models. But, the linear, robust M-estimator, and MM-estimator models produced estimates with smallest standard errors. Again, these results were consistent with the simulation findings and reflected the ability of robust models to deal with outliers to get reliable estimates and the power of maximum likelihood-based estimates from linear models in large sample cases (Rousseeuw & Hubert, 2018).

The real data showed that residing in rural area, higher body mass index, having primary and secondary education was linked to high Hb levels. The body mass index is function of few other body mechanisms such as weight, height, fats, which are related to blood quantity in the body, which could be the reason this study, like others done previously, observed a positive association between women body mass index and Haemoglobin level (Mocking et al., 2018; Kamruzzaman, 2021). A woman's educational attainment is a

critical tool for nutritional awareness, and hence its link with Haemoglobin status of the woman (Adediran et al., 2011). The low likelihood of maternal anaemia in women from rural parts of Malawi is consistent with studies done in other low and middle income countries, such as South East Asia, but more research is needed to establish the reasons for the trend (Rahman et al., 2021).

In contrast, higher age of pregnancy, drinking from safe water sources, and living in a rich household were associated with low Hb levels. The low Hb levels in the second and third trimesters of pregnancy have been reported in many studies and it reflects the demand for more iron mineral by the growing baby (Ray et al., 2020; Churchill et al., 2019). Although drinking from borehole water is classified as safe, previous research observed that local communities in Malawi do not treat water from boreholes to make it safe for drinking (Mkwate et al., 2017). Over half of the women in this study used tube well or borehole as source of drinking water. This could reflect low Hb levels observed in women drinking from safe water sources, as the water might lack appropriate nitrates (Kothari et al., 2019; Westgard et al., 2021; Jana et al., 2022). The result of low Hb levels in women from rich household is uncommon as previous research established positive association between family wealth and Hb levels, since wealthier households could afford proper nutrition (Abate et al., 2021; Awoleye et al., 2022). These factors collectively highlight the complex interplay between socio-economic status, education, nutrition, and biological factors in determining maternal anaemia outcomes in Malawi. Addressing these issues requires a multifaceted approach.

5.2 Conclusion

This study evaluated the effectiveness and of robust, quantile, and mean regression models in managing outlier data related to haemoglobin levels in women in Malawi. Simulations revealed that all three models had similar outlier detection rates, except 90th quantile model in small sample sizes. However, robust and some lower quantile (25th and 50th) regression methods provided more accurate estimates in samples with outliers. When applied to women's haemoglobin data, the fixed effect estimates were consistent across models, with the linear, M-estimator, and MM-estimator models yielding the smallest standard errors in large samples. The average haemoglobin (Hb) level for women in Malawi was 12.8 g/dl, with a raw standard deviation of 1.74 g/dl. Women in rural areas, those with higher body mass index, and those with primary and secondary education had significantly higher Hb levels, while increased pregnancy age, drinking from safe water sources, and living in wealthy households were associated with lower Hb levels. These findings are supported by recent studies that emphasize the stability and reliability of robust regression methods in various data conditions

It was further observed that the women Haemoglobin data had a considerable amount of outliers (models detected a range of 400 to 500), whom the majority were women with extreme low Haemoglobin levels. Thus the Haemoglobin level data in Malawi were highly skewed to the left with more unusual values at the tip below te average. Such that using mean regression alone might not be sufficient. Quantile regression can provide insights across different points in the haemoglobin distribution, while robust regression can handle outliers effectively. Combining these methods offers a comprehensive approach to accurately model the non-linear relationship between predictors and haemoglobin levels.

The study generates strong evidence of the burden of maternal anaemia in Malawi based on outlier residuals of three different statistical modelling methods that were engaged in this study. Future research should consider using mixed-effects regression models to account for the clustering of women in their neighborhoods while analyzing outlier haemoglobin levels. This approach could offer a more comprehensive understanding of the factors influencing haemoglobin levels and help design more effective interventions. Overall, the study highlights the effectiveness of robust and quantile regression methods in handling outlier data and provides valuable insights into the factors affecting haemoglobin levels in women in Malawi.

5.3 Recommendations

The study recommends the use of robust regression methods to improve the modelling of women Haemoglobin data in Malawi. It also suggests implementing targeted interventions to boost haemoglobin levels, especially among expectant mothers in the second and third trimester and other outlier groups of women in society. These findings further suggest that triangulating a variety of statistical methods to analyse Haemoglobin data will help in concretising evidence of the burden of maternal anaemia in sub-Saharan Africa.

5.4 Study Limitation

The study analyzed maternal anaemia using Malawi's 2015-16 Demographic and Health Survey (DHS) data encountered two primary limitations. Firstly, missing values in critical variables (such as current pregnant duration, water source, and BMI). These missing data points may have introduced bias and affected the accuracy of statistical analyses. Secondly, due to a small sample size, generalizing the study findings to the entire Malawi

population is a challenge. Although mean, quantile, and robust regression methods effectively handle outliers and provide better estimates, in modelling of anaemia data might not be as clinically intuitive as logistic regression. This is so because mean, quantile, and robust regression use haemoglobin levels and hence fails to effectively captured the binary nature of anaemia diagnosis (i.e., anaemic vs. non-anaemic). However, quantile and robust models fitted well due to the skewness of the haemoglobin level data used to categorize the anaemia condition.

References

Abate, T. W., Getahun, B., Birhan, M. M., Aknaw, G. M., Belay, S. A., Demeke, D., ... Mengiste, Y. (2021). The urban–rural differential in the association between household wealth index and anemia among women in reproductive age in ethiopia, 2016. *BMC Women's Health*, *21*, 1–8.

Acharya, D., Adhikari, R., & Simkhada, P. (2022). Prevalence and determinants of anaemia among women aged 15-49 in nepal: A trend analysis from nepal demographic and health surveys from 2006 to 2016. *Asian Journal of Population Sciences*, *1*, 32–48.

Adamu, A. L., Crampin, A., Kayuni, N., Amberbir, A., Koole, O., Phiri, A., ... Fine, P. (2017). Prevalence and risk factors for anemia severity and type in malawian men and women: urban and rural differences. *Population health metrics*, *15*(1), 1–15.

Adediran, A., Gbadegesin, A., Adeyemo, T., Akinbami, A., Akanmu, A., Osunkalu, V., ... Oremosu, A. (2011). Haemoglobin and ferritin concentrations of pregnant women at term. *Obstetric medicine*, *4*(4), 152–155.

Adichie, J. N. (1967). Estimates of regression parameters based on rank tests. *The Annals of Mathematical Statistics*, *38*(3), 894–904.

Alem, A. Z., Efendi, F., McKenna, L., Felipe-Dimog, E. B., Chilot, D., Tonapa, S. I., ... Zainuri, A. (2023). Prevalence and factors associated with anemia in women of reproductive age across low-and middle-income countries based on national data. *Scientific Reports*, *13*(1), 20335.

Ali, S. A., Razzaq, S., Aziz, S., Allana, A., Ali, A. A., Naeem, S., ... Ur Rehman, F. (2023). Role of iron in the reduction of anemia among women of reproductive age in

low-middle income countries: insights from systematic review and meta-analysis. *BMC women's health*, *23*(1), 1–22.

Andersen, R. (2008). *Modern methods for robust regression* (No. 152). Sage.

Arimie, C. O., Harcourt, P., Harcourt, P., Harcourt, P., et al. (2020). Outlier detection and effects on modeling. *Open Access Library Journal*, *7*(09), 1.

Atkinson, A. C. (1982). Robust and diagnostic regression analyses. *Communications in Statistics-Theory and Methods*, *11*(22), 2559–2571.

Awoleye, A. F., Alawode, O. A., Chima, V., Okunlola, D. A., & Obiesie, S. (2022). Rural-urban differentials in the relationship between household wealth index and maternal anaemia status in nigeria. *Health Care for Women International*, 1–16.

Ayinde, K., Lukman, A. F., Arowolo, O., et al. (2015). Robust regression diagnostics of influential observations in linear regression model. *Open Journal of Statistics*, *5*(04), 273.

Bagheri, A., Midi, H., Ganjali, M., & Eftekhari, S. (2010). A comparison of various influential points diagnostic methods and robust regression approaches: Reanalysis of interstitial lung disease data. *Applied Mathematical Sciences*, *4*(28), 1367–1386.

Bahadir, B., İnci, H., & Karadavut, U. (2014). Determination of outlier in live-weight performance data of japanese quails (coturnix coturnix japonica) by dfbeta and dfbetas techniques. *Italian Journal of Animal Science*, *13*(1), 3113.

Barnett, V., Lewis, T., et al. (1994). *Outliers in statistical data* (Vol. 3) (No. 1). Wiley New York.

Bary, M. N. A. (2017). Robust regression diagnostic for detecting and solving multicollinearity and outlier problems: Applied study by using financial data. *Applied Mathematical Sciences*, *11*(13), 601–622.

Belsley, D. A., Kuh, E., & Welsch, R. E. (2005). *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons.

Black, R. E., Victora, C. G., Walker, S. P., Bhutta, Z. A., Christian, P., De Onis, M., ... others (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The lancet*, *382*(9890), 427–451.

Chaku, S. E., & Donev, A. (n.d.). An investigation into some results of glm modelling of some data sets. *Structure*, *2*, 2.

Chanimbe, B., Issah, A.-N., Mahama, A. B., Yeboah, D., Kpordoxah, M. R., Shehu, N., ... Boah, M. (2023). Access to basic sanitation facilities reduces the prevalence of anaemia among women of reproductive age in sub-saharan africa. *BMC Public Health*, *23*(1), 1999.

Chaparro, C. M., & Suchdev, P. S. (2019). Anemia epidemiology, pathophysiology, and etiology in low-and middle-income countries. *Annals of the new York Academy of Sciences*, *1450*(1), 15–31.

Chatterjee, S., & Hadi, A. S. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical science*, 379–393.

Chatterjee, S., & Hadi, A. S. (2009). *Sensitivity analysis in linear regression*. John Wiley & Sons.

Chen, C. (2002). Paper 265-27 robust regression and outlier detection with the robustreg procedure. In *Proceedings of the proceedings of the twenty-seventh annual sas users group international conference.*

Churchill, D., Nair, M., Stanworth, S. J., & Knight, M. (2019). The change in haemoglobin concentration between the first and third trimesters of pregnancy: a population study. *BMC pregnancy and childbirth*, *19*, 1–6.

Čížek, P., & Sadıkoğlu, S. (2020). Robust nonparametric regression: A review. *Wiley Interdisciplinary Reviews: Computational Statistics*, *12*(3), e1492.

Clayton, D. G. (1996). Generalized linear mixed models. *Markov chain Monte Carlo in practice*, *1*, 275–302.

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15–18.

Cook, R. D. (2000). Detection of influential observation in linear regression. *Technometrics*, *42*(1), 65–68.

Denby, L., & Mallows, C. L. (1977). Two diagnostic displays for robust regression analysis. *Technometrics*, *19*(1), 1–13.

Di Renzo, G. C., Spano, F., Giardina, I., Brillo, E., Clerici, G., & Roura, L. C. (2015). Iron deficiency anemia in pregnancy. *Women's Health*, *11*(6), 891–900.

Dobson, & Barnett. (2018). *An introduction to generalized linear models*. Chapman and Hall/CRC.

Dobson, & Barnett, A. (2008). An introduction to generalized linear models third edition introduction. *Ch Crc Text Stat Sci*, *77*(1).

Doganer, A., Tok, A., & Demirel, G. (2021). Prediction of factors affecting cognitive performance in pregnant women using robust regression methods. *Journal of Biostatistics and Epidemiology*.

Fox, J. (2002). Nonparametric regression. *Appendix to: An R and S-PLUS Companion to Applied Regression*, 1–7.

Geraci, M., & Bottai, M. (2014). Linear quantile mixed models. *Statistics and computing*, *24*, 461–479.

Geta, T. G., Gebremedhin, S., & Omigbodun, A. O. (2022). Prevalence and predictors of anemia among pregnant women in ethiopia: Systematic review and meta-analysis. *PloS one*, *17*(7), e0267005.

Gray, J. B. (1989). On the use of regression diagnostics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *38*(2), 97–105.

Grynovicki, J., Thomas, J., & MD, A. B. R. L. A. P. G. (1983). Robust regression: A diagnostic tool. *NTIS, SPRINGFIELD, VA, 1983, 33*.

Hasan, M. M., Magalhaes, R. J. S., Garnett, S. P., Fatima, Y., Tariqujjaman, M., Pervin, S., ... Mamun, A. A. (2022). Anaemia in women of reproductive age in low-and middle-income countries: progress towards the 2025 global nutrition target. *Bulletin of the World Health Organization*, *100*(3), 196.

Huang, Q., Zhang, H., Chen, J., & He, M. (2017). Quantile regression models and their applications: A review. *Journal of Biometrics & Biostatistics*, *8*(3), 1–6.

Huber, P. J. (1973). Robust regression: asymptotics, conjectures and monte carlo. *The annals of statistics*, 799–821.

Iglewicz, B., & Martinez, J. (1982). Outlier detection using robust measures of scale. *Journal of Statistical Computation and Simulation*, *15*(4), 285–293.

Jajo. (2005). A review of robust regression and diagnostic procedures in linear regression. *Acta Mathematicae Applicatae Sinica*, *21*, 209–224.

Jajo, & Hussain. (1989). Robust estimators in linear regression model. *Journal of Management and Economic*, *10*, 1–15.

Jamee, A. R., Sen, K. K., & Bari, W. (2022). Examining the influence of correlates on different quantile survival times: infant mortality in bangladesh. *BMC Public Health*, *22*(1), 1980.

Jana, A., Chattopadhyay, A., & Saha, U. R. (2022). Identifying risk factors in explaining women's anaemia in limited resource areas: evidence from west bengal of india and bangladesh. *BMC Public Health*, *22*(1), 1433.

Jiang, J., & Nguyen, T. (2007). *Linear and generalized linear mixed models and their applications* (Vol. 1). Springer.

John, O. O., & Nduka, E. C. (2009). Quantile regression analysis as a robust alternative to ordinary least squares. *Scientia Africana*, *8*(2), 61–65.

Kalina, J. (2015). Three contributions to robust regression diagnostics. *Journal of Applied Mathematics, Statistics and Informatics*, *11*(2), 69–78.

Kamruzzaman, M. (2021). Is bmi associated with anemia and hemoglobin level of women and children in bangladesh: A study with multiple statistical approaches. *PLoS One*, *16*(10), e0259116.

Kannan, K. S., & Manoj, K. (2015). Outlier detection in multivariate data. *Applied mathematical sciences*, *47*(9), 2317–2324.

Kaombe, T. M. (2024). A bivariate poisson regression to analyse impact of outlier women on correlation between female schooling and fertility in malawi. *BMC Women's Health*, *24*(1), 55.

Kaombe, T. M., Banda, J. C., Hamuza, G. A., & Muula, A. S. (2023). Bivariate logistic regression model diagnostics applied to analysis of outlier cancer patients with comorbid diabetes and hypertension in malawi. *Scientific Reports*, *13*(1), 8340.

Kaombe, T. M., & Manda, S. O. (2023a). Detecting influential data in multivariate survival models. *Communications in Statistics-Theory and Methods*, *52*(11), 3910–3926.

Kaombe, T. M., & Manda, S. O. (2023b). A novel outlier statistic in multivariate survival models and its application to identify unusual under-five mortality sub-districts in malawi. *Journal of Applied Statistics*, *50*(8), 1836–1852.

Karami, M., Chaleshgar, M., Salari, N., Akbari, H., & Mohammadi, M. (2022). Global prevalence of anemia in pregnant women: a comprehensive systematic review and meta-analysis. *Maternal and child health journal*, *26*(7), 1473–1487.

Kassebaum, N. J., Collaborators, G. . A., et al. (2016). The global burden of anemia. *Hematology/oncology clinics of North America*, *30*(2), 247–308.

Khatun, N., et al. (2021). Applications of normality test in statistical analysis. *Open journal of statistics*, *11*(01), 113.

Kim, J., & Li, J. C.-H. (2023). Which robust regression technique is appropriate under violated assumptions? a simulation study. *Methodology*, *19*(4), 323–347.

Kinyoki, D., Osgood-Zimmerman, A. E., Bhattacharjee, N. V., Kassebaum, N. J., & Hay, S. I. (2021). Anemia prevalence in women of reproductive age in low-and middle-income countries between 2000 and 2018. *Nature medicine*, *27*(10), 1761–1782.

Koenker, R. (2005). *Quantile regression* (Vol. 38). Cambridge University Press.

Koenker, R. (2017). Quantile regression: 40 years on. *Annual review of economics*, *9*, 155–176.

Koenker, R., & Hallock, K. F. (2001). Quantile regression. *Journal of economic perspectives*, *15*(4), 143–156.

Koller, M. (2016). robustlmm: an r package for robust estimation of linear mixed-effects models. *Journal of statistical software*, *75*, 1–24.

Kothari, M. T., Coile, A., Huestis, A., Pullum, T., Garrett, D., & Engmann, C. (2019). Exploring associations between water, sanitation, and anemia through 47 nationally representative demographic and health surveys. *Annals of the New York Academy of Sciences*, *1450*(1), 249–267.

Lakshmi, K., Mahaboob, B., Rajaiah, M., & Narayana, C. (2021). Ordinary least squares estimation of parameters of linear model. *J. Math. Comput. Sci.*, *11*(2), 2015–2030.

Meena, K., Tayal, D. K., Gupta, V., & Fatima, A. (2019). Using classification techniques for statistical analysis of anemia. *Artificial intelligence in medicine*, *94*, 138–152.

Mei Ling Huang, X. X., & Tashnev, D. (2015). A weighted linear quantile regression. *Journal of Statistical Computation and Simulation*, *85*(13), 2596–2618. doi: 10.1080/00949655.2014.938240

Mkwate, R. C., Chidya, R. C., & Wanda, E. M. (2017). Assessment of drinking water quality and rural household water treatment in balaka district, malawi. *Physics and Chemistry of the Earth, Parts a/b/c*, *100*, 353–362.

Mocking, M., Savitri, A. I., Uiterwaal, C. S., Amelia, D., Antwi, E., Baharuddin, M., ... Browne, J. L. (2018). Does body mass index early in pregnancy influence the risk of maternal anaemia? an observational study in indonesian and ghanaian women. *BMC Public Health*, *18*, 1–9.

Moya, E., Phiri, N., Choko, A. T., Mwangi, M. N., & Phiri, K. S. (2022). Effect of postpartum anaemia on maternal health-related quality of life: a systematic review and meta-analysis. *BMC Public Health*, *22*(1), 364.

Myers, R. H., & Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, *29*(3), 274–291.

Neuhaus, J., & McCulloch, C. (2011). Generalized linear models. *Wiley Interdisciplinary Reviews: Computational Statistics*, *3*(5), 407–413.

Notapiri, T., Toharudin, T., & Suparman, Y. (2022). Modeling of crime rate in indonesia during the covid-19 pandemic from a macroeconomic perspective: Using robust regression with s-estimator. *J. Math. Comput. Sci.*, *12*, Article–ID.

Ohuma, E. O., Jabin, N., Young, M. F., Epie, T., Martorell, R., Peña-Rosas, J. P., ... others (2023). Association between maternal haemoglobin concentrations and maternal

and neonatal outcomes: the prospective, observational, multinational, interbio-21st fetal study. *The Lancet Haematology*, *10*(9), e756–e766.

Owais, A., Merritt, C., Lee, C., & Bhutta, Z. A. (2021). Anemia among women of reproductive age: an overview of global burden, trends, determinants, and drivers of progress in low-and middle-income countries. *Nutrients*, *13*(8), 2745.

Oyeyemi, G., Oluwaseun, O., & Adeleke, M. (2017). Comparisons of some outlier detection methods in linear regression model. *Ilorin Journal of Science*, *4*(1), 130–138.

Pasricha, S.-R., Black, J., Muthayya, S., Shet, A., Bhat, V., Nagaraj, S., ... Shet, A. S. (2010). Determinants of anemia among young children in rural india. *Pediatrics*, *126*(1), e140–e149.

Pasricha, S.-R., & Moir-Meyer, G. (2023). Measuring the global burden of anaemia. *The Lancet Haematology*, *10*(9), e696–e697.

Peña, E. A., & Slate, E. H. (2006). Global validation of linear model assumptions. *Journal of the American Statistical Association*, *101*(473), 341–354.

Pérez, B., Molina, I., & Peña, D. (2014). Outlier detection and robust estimation in linear regression models with fixed group effects. *Journal of Statistical Computation and Simulation*, *84*(12), 2652–2669.

Pierce, D. A., & Schafer, D. W. (1986). Residuals in generalized linear models. *Journal of the American Statistical Association*, *81*(396), 977–986.

Poole, M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression model. *Transactions of the Institute of British Geographers*, 145–158.

Rahman, M. A., Rahman, M. S., Aziz Rahman, M., Szymlek-Gay, E. A., Uddin, R., & Islam, S. M. S. (2021). Prevalence of and factors associated with anaemia in women of reproductive age in bangladesh, maldives and nepal: Evidence from nationally-representative survey data. *Plos one*, *16*(1), e0245335.

Ray, J., Davidson, A., Berger, H., Dayan, N., & Park, A. (2020). Haemoglobin levels in early pregnancy and severe maternal morbidity: population-based cohort study. *BJOG: An International Journal of Obstetrics & Gynaecology*, *127*(9), 1154–1164.

Ritschard, G., & Antille, G. (1992). A robust look at the use of regression diagnostics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *41*(1), 41–53.

Rodriguez, R. N., & Yao, Y. (2017). Five things you should know about quantile regression. In *Proceedings of the sas global forum 2017 conference, orlando* (pp. 2–5).

Ronchetti, E. M., & Huber, P. J. (2009). *Robust statistics*. John Wiley & Sons Hoboken, NJ, USA.

Rousseeuw, & Hubert. (2011). Robust statistics for outlier detection. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, *1*(1), 73–79.

Rousseeuw, & Hubert. (2018). Anomaly detection by robust statistics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *8*(2), e1236.

Rousseeuw, & Leroy. (1988). A robust scale estimator based on the shortest half. *Statistica Neerlandica*, *42*(2), 103–116.

Rousseeuw, & Leroy. (2005). *Robust regression and outlier detection*. John wiley & sons.

Rousseeuw, & Van, K. (2006). Computing lts regression for large data sets. *Data mining and knowledge discovery*, *12*, 29–45.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of s-estimators. In *Robust and nonlinear time series analysis: Proceedings of a workshop organized by the sonderforschungsbereich 123 "stochastische mathematische modelle", heidelberg 1983* (pp. 256–272).

Safiri, S., Kolahi, A.-A., Noori, M., Nejadghaderi, S. A., Karamzad, N., Bragazzi, N. L., ... others (2021). Burden of anemia and its underlying causes in 204 countries and territories, 1990–2019: results from the global burden of disease study 2019. *Journal of hematology & oncology*, *14*(1), 1–16.

Santos, F. (2020). Modern methods for old data: An overview of some robust methods for outliers detection with applications in osteology. *Journal of Archaeological Science: Reports*, *32*, 102423.

Sarstedt, M., Mooi, E., Sarstedt, M., & Mooi, E. (2019). Regression analysis. *A concise guide to market research: The process, data, and methods using IBM SPSS Statistics*, 209–256.

Sevier, F. A. (1957). Testing the assumptions underlying multiple regression. *The Journal of Experimental Education*, *25*(4), 323–330.

Shi, G., Zhang, Z., Ma, L., Zhang, B., Dang, S., & Yan, H. (2021). Association between maternal iron supplementation and newborn birth weight: a quantile regression analysis. *Italian journal of pediatrics*, *47*(1), 133.

Sinha, S. K. (2004). Robust analysis of generalized linear mixed models. *Journal of the American Statistical Association*, *99*(466), 451–460.

Soofi, S., Khan, G. N., Sadiq, K., Ariff, S., Habib, A., Kureishy, S., ... others (2017). Prevalence and possible factors associated with anaemia, and vitamin b12 and folate deficiencies in women of reproductive age in pakistan: analysis of national-level secondary survey data. *BMJ open*, *7*(12).

Stevens, G. A., Paciorek, C. J., Flores-Urrutia, M. C., Borghi, E., Namaste, S., Wirth, J. P., ... others (2022). National, regional, and global estimates of anaemia by severity in women and children for 2000–19: a pooled analysis of population-representative data. *The Lancet Global Health*, *10*(5), e627–e639.

Sun, J., Wu, H., Zhao, M., Magnussen, C. G., & Xi, B. (2021). Prevalence and changes of anemia among young children and women in 47 low-and middle-income countries, 2000-2018. *EClinicalMedicine*, *41*.

Sunuwar, D. R., Singh, D. R., Chaudhary, N. K., Pradhan, P. M. S., Rai, P., & Tiwari, K. (2020). Prevalence and factors associated with anemia among women of reproductive age in seven south and southeast asian countries: Evidence from nationally representative surveys. *PloS one*, *15*(8), e0236449.

Talukder, A., Paul, N., Khan, Z. I., Ahammed, B., Haq, I., & Ali, M. (2022). Risk factors associated with anemia among women of reproductive age (15–49) in albania: A quantile regression analysis. *Clinical Epidemiology and Global Health*, *13*, 100948.

Teshale, A. B., Tesema, G. A., Worku, M. G., Yeshaw, Y., & Tessema, Z. T. (2020). Anemia and its associated factors among women of reproductive age in eastern africa: A multilevel mixed-effects generalized linear model. *Plos one*, *15*(9), e0238957.

Thompson, M. (1982). Regression methods in the comparison of accuracy. *Analyst*, *107*(1279), 1169–1180.

Türkan, S., ÇETİN, M. C., & TOKTAMIŞ, Ö. (2012). Outlier detection by regression diagnostics based on robust parameter estimates full text. *Hacettepe Journal of Mathematics and Statistics*, *41*(1), 147–155.

Ullah, A., Sohaib, M., Saeed, F., & Iqbal, S. (2019). Prevalence of anemia and associated risk factors among pregnant women in lahore, pakistan. *Women & health*, *59*(6), 660–671.

Verardi, V., & Croux, C. (2009). Robust regression in stata. *The Stata Journal*, *9*(3), 439–453.

Verran, J. A., & Ferketich, S. L. (1987). Testing linear model assumptions: Residual analysis. *Nursing Research*, *36*(2), 127–129.

Waldmann, E. (2018). Quantile regression: A short story on how and why. *Statistical Modelling*, *18*(3-4), 203–218.

Walsh, A., Matthews, A., Manda-Taylor, L., Brugha, R., Mwale, D., Phiri, T., & Byrne, E. (2018). The role of the traditional leader in implementing maternal, newborn and child health policy in malawi. *Health policy and planning*, *33*(8), 879–887.

Westgard, C. M., Orrego-Ferreyros, L. A., Calderón, L. F., & Rogers, A. M. (2021). Dietary intake, intestinal infection, and safe drinking water among children with anemia in peru: a cross-sectional analysis. *BMC nutrition*, *7*, 1–7.

Wilcox, R. R. (1996). A review of some recent developments in robust regression. *British Journal of Mathematical and Statistical Psychology*, *49*(2), 253–274.

Wood, S. N. (2004). Stable and efficient multiple smoothing parameter estimation for generalized additive models. *Journal of the American Statistical Association*, *99*(467), 673–686.

Yaffee, R. A. (2002). Robust regression analysis: some popular statistical package options. *Statistics, social science, and mapping group academic computing services information technology services*, 1–12.

Yau, K. K., & Kuk, A. Y. (2002). Robust estimation in generalized linear mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(1), 101–117.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *The Annals of statistics*, 642–656.

Young, M. F. (2018). Maternal anaemia and risk of mortality: a call for action. *The Lancet Global Health*, *6*(5), e479–e480.

Young, M. F., Oaks, B. M., Rogers, H. P., Tandon, S., Martorell, R., Dewey, K. G., & Wendt, A. S. (2023). Maternal low and high hemoglobin concentrations and associations with adverse maternal and infant health outcomes: an updated global systematic review and meta-analysis. *BMC Pregnancy and Childbirth*, *23*(1), 1–16.

Yuen, K.-V., & Ortiz, G. A. (2017). Outlier detection and robust regression for correlated data. *Computer Methods in Applied Mechanics and Engineering*, *313*, 632–646.

Appendices

Appendix 1: STATA Codes

```
*SIMULATING DATA: CASE OF UNPERTURBED DATA with sample size of 50 *
************************************************************
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Sim_unpertubed50"
clear all
set seed 12345 // Set a seed for reproducibility
forvalues i = 1/100 {
clear
set obs 50
gen group = `i'
gen id =_n
gen x1 = rnormal(2.3,0.5)
gen x2 = rnormal(8,2.4)
gen error = rnormal(0,1)
gen y = 2.1 + 0.7*x1 + 0.9*x2 + error
// Save each sample to a separate file
save sim_unpertubed50data`i'.dta, replace
}
* SIMULATING DATA - CASE OF PERTURBED first 5 observations of size 500 *
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Sim_pertubed500"
clear all
set seed 12345 // Set a seed for reproducibility
forvalues i = 1/100 {
93clear
```

```
set obs 500 // Set number of observations to 500
gen group = `i' // Create the group variable
gen id =_n // Generate normal data
gen x1 = rnormal(2.3,0.5)
gen x2 = rnormal(8,2.4)
gen error = rnormal(0,1)
gen y = 2.1 + 0.7*x1 + 0.9*x2 + error
// Introduce random outliers for the first 5 observations
set seed `= 12345 + `i" // Ensure reproducibility with varying seed
replace error = rnormal(-7.8,22.1) in 1/5 // introducing outlier in error term
// Recalculate y for the first 5 observations with outliers
replace y = 15 + 6*x1 + 10*x2 + error in 1/5 // b0, b1 and b2 perturbed to 15, 6 and 10
respectively
save sim_pertubed500data`i'.dta, replace // Save each sample to a separate file
}
* SIMULATING DATA - A CASE OF PERTURBED first 125 observations of size 500 *
**************************************************************************
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\half_sim_perturbed500"
clear all
set seed 12345 // Set a seed for reproducibility
forvalues i = 1/100 {
clear
set obs 500 // Set number of observations to 500
gen group = `i' // Create the group variable
gen id =_n // Generate normal data
gen x1 = rnormal(2.3,0.5)
gen x2 = rnormal(8,2.4)
9495
```

```
replace y = 15 + 6*x1 + 10*x2 + error in 1/125 // b0, b1 and b2 perturbed to 15, 6 and 10
respectivel
gen error = rnormal(0,1)
gen y = 2.1 + 0.7*x1 + 0.9*x2 + error
// Introduce random outliers for the first 250 observations
set seed `= 12345 + `i'' // Ensure reproducibility with varying seed
replace error = rnormal(-7.8,22.1) in 1/125 // introducing outlier in error term
// Recalculate y for the first 5 observations with outliers
y
save half_sim_pertubed500data`i'.dta, replace
// Save each sample to a separate file
}
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis"
*************************************************************************
* SIMULATION STUDY DATA ANALYSIS - pertubed first 5 observations for n=500
*************************************************************************
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Sim_pertubed500"
*A CASE OF LINEAR REGRESSION MODEL*****
capture log using Maternal_anaemiaThesisOLS500pertubed.log, replace
// Load the 100 replicated datasets
forval i = 1/100 {
use sim_pertubed500data`i'.dta, clear
regress y x1 x2 // Fit a linear regression model
di "Running regression for dataset number `i'" // Display the dataset number in the log
predict yhat
gen residuals = y - yhat
gen squared_residuals = residuals^2
```

```
sum squared_residuals, meanonly
display sqrt(r(mean)) // Calculate RMSEgen abs_residuals = abs(residuals
96
)
sum abs_residuals, detail
gen outlier = abs_residuals > r(p75) + 1.5 * (r(p75) - r(p25))
list group id if outlier
// List outliers
}
Capture log close
**A CASE OF ROBUST MODELS - LTS-estimator**********
capture log using Maternal_anaemiaThesisLTSRob500pertubed.log, replace
//Load the 100 replicated datasets
forval i = 1/100 {
use sim_pertubed500data`i'.dta, clear
robreg lts y x1 x2 // m-robust modeL
di "Running regression for dataset number `i'" // Display the dataset number in the log
predict yhat
gen residuals = y - yhat
gen squared_residuals = residuals^2
sum squared_residuals, meanonly
display sqrt(r(mean)) // Calculate RMSE
gen abs_residuals = abs(residuals)
sum abs_residuals, detail
gen outlier = abs_residuals > r(p75) + 1.5 * (r(p75) - r(p25))
list group id if outlier
}
capture log close
**********************************************************
* SIMULATED DATA ANALYSIS - Unpertubed for n=50
```

```
*******************************************************
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization (MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis\Sim_unpertubed50"**QR MODELS - 25th Quantile Regression**********
capture log using Maternal_anaemiaThesisQ25R50unpertubed.log, replace
//Load the 100 replicated datasets
forval i = 1/100 {
use sim_unpertubed50data`i'.dta, clear
qreg y x1 x2, quantile(0.25) // For 25th percentile
di "Running regression for dataset number `i'"
predict yhat
gen residuals = y - yhat
gen squared_residuals = residuals^2
sum squared_residuals, meanonly
display sqrt(r(mean))
gen abs_residuals = abs(residuals)
sum abs_residuals, detail
gen outlier = abs_residuals > r(p75) + 1.5 * (r(p75) - r(p25))
list group id if outlier
}
capture log close
* SIMULATION STUDY DATA ANALYSIS - Unpertubed for n=500
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization (MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis\Sim_unpertubed500"
**CASE OF ROBUST MODELS - MM-estimator**********
capture log using Maternal_anaemiaThesisMMRob500unpertubed.log, replace
//Load the 100 replicated datasets
forval i = 1/100 {
```

```
use sim_unpertubed500data`i'.dta, clear
robreg mm y x1 x2 // m-robust model
di "Running regression for dataset number `i'"
predict yhat
5gen residuals = y - yhat
gen squared_residuals = residuals^2
sum squared_residuals, meanonly
display sqrt(r(mean))
gen abs_residuals = abs(residuals)
sum abs_residuals, detail
gen outlier = abs_residuals > r(p75) + 1.5 * (r(p75) - r(p25))
list group id if outlier
}
capture log close
* SIMULATION STUDY DATA ANALYSIS - pertubed 25% (175 observations) in n=500
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\half_sim_perturbed500"
**CASE OF ROBUST MODELS - M-estimator**********
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\half_sim_perturbed500"
capture log using MRob500_half_pertubed.log, replace
//Load the 100 replicated datasets
forval i = 1/100 {
use half_sim_pertubed500data`i'.dta, clear
robreg m y x1 x2 // m-robust model
di "Running regression for dataset number `i'"
predict yhat
gen residuals = y - yhat
```

```
gen squared_residuals = residuals^2
sum squared_residuals, meanonly
display sqrt(r(mean))
}
6capture log close
* SIMULATION STUDY DATA ANALYSIS - pertubed 25% (13 observatiobs) in n=50
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\half_sim_perturbed50"
**CASE OF QR MODELS - 75th Quantile Regression**********
capture log using Q75R50_half_pertubed.log, replace
//Load the 100 replicated datasets
forval i = 1/100 {
use half_sim_pertubed50data`i'.dta, clear
qreg y x1 x2, quantile(0.75) // For 75th percentile
di "Running regression for dataset number `i'"
predict yhat
gen residuals = y - yhat
gen squared_residuals = residuals^2
sum squared_residuals, meanonly
display sqrt(r(mean))
}
capture log close
*******************************************************
***** MATERNAL ANAEMIA DATA CLEANING *
quietly{
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis"
use "dhs_data.DTA", clear
*Dropping respondents with missing hemoglobin data
```

```
drop if hemoglobin_level10 == 996 | hemoglobin_level == 995 | hemoglobin_level == 994 |
hemoglobin_level == . //45,441 observations deleted, 21,935 kept
summarize hemoglobin_level, detail
*Recategorization of variables
7gen hemoglobin_level = hemoglobin_level10/10 //converting g/l to g/dl
gen bmi2 = bmi/100
gen bmi_category = 0 if inrange(bmi2, 0, 18.5)
replace bmi_category = 1 if inrange(bmi2, 18.5, 22.99)
replace bmi_category = 2 if inrange(bmi2, 23.0, 27.49)
replace bmi_category = 3 if inrange(bmi2, 27.5, 90)
replace bmi_category = . if bmi==9998 | bmi==9996 | bmi==9995 | bmi==9994
gen contraceptive_use = 1 if contraceptive == 1
replace contraceptive_use = 2 if contraceptive == 2
replace contraceptive_use = 3 if contraceptive == 3 | contraceptive == 4
gen wealth_index = 1 if wealth_level == 1 | wealth_level == 2
replace wealth_index = 2 if wealth_level == 3
replace wealth_index = 3 if wealth_level == 4 | wealth_level == 5
gen age_category = 0 if inrange(age, 15, 24)
replace age_category = 1 if inrange(age, 25, 49)
gen water_source = 0 if water_of_source == 32 | water_of_source == 42 | water_of_source == 43
| water_of_source == 51 | water_of_source == 96 //unprotected well, unprotected spring, surface
water, rain water and other water sources
replace water_source = 1 if water_of_source == 21 | water_of_source == 11 | water_of_source ==
12 | water_of_source == 13 | water_of_source == 14 | water_of_source == 31 | water_of_source
== 41 //borehole/Tube well, piped, protected well and spring
replace water_source = . if water_of_source == 97
*Labelling created values
label define bmi_group 0 "underweight" 1 "normal_weight" 2 "overweight" 3 "obese", modify
label values bmi_category bmi_group
label define women_age_group 0 "15-24" 1 "25-49", modify
```

```
label values age_category women_age_group
label define h20source 0 "Unsafe water" 1 "Safe water"
label values water_source h20source
label define distancehc 1 "Big problem" 2 "No problem", modify
89
********Imputation of missing observations***************
codebook bmi_category age_category education_level distance_hc wealth_index water_source
residential_status contraceptive_use total_fertility_rate curr_preg_duration
***imputing missing values for continuous variables******
egen imp_curr_preg_duration = median(curr_preg_duration)
replace curr_preg_duration = imp_curr_preg_duration if curr_preg_duration==. // replacing
missing 20775 with median value = mdian(5)
***imputing missing values for categorical variables******
replace water_source = 1 if missing(water_source) //replacing 104 missing values with most
occuring category (Safe water)
replace bmi_category = 1 if missing(bmi_category) //replacing 58 missing values with most
occuring category (normal weight)
save "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\dhs_anaemia.dta", replace
}
**********************************************************************
*MATERNAL DATA ANALYSIS
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis"
capture log using DHS_Maternal_anaemiaThesismodel_fit.log, replace
*
CASE OF LM/MLE
use "dhs_anaemia.dta", clear
regress hemoglobin_level i.residential_status i.education_level total_fertility_rate
```

```
curr_preg_duration i.distance_hc i.bmi_category i.wealth_index i.age_category i.water_source
}
*Outlier detection
predict yhat
predict r, rstudent
// Identify outliers10
sum r, detail
display r(p75) + 1.5 * (r(p75) - r(p25)) //upper outlier cut off point
display r(p25) - 1.5 * (r(p75) - r(p25)) //lower outlier cut off point
gen outlier = r > r(p75) + 1.5 * (r(p75) - r(p25)) | r < r(p25) - 1.5 * (r(p75) - r(p25))
gen mle_out_sign = "+ve" if r > r(p75) + 1.5 * (r(p75) - r(p25))
replace mle_out_sign = "-ve" if r < r(p25) - 1.5 * (r(p75) - r(p25))
quietly {
graph box r, cwhisker marker(1, mlabel(caseid)) title(OLS/ML identified outliers)
graphregion(color(white))
graph save "Graph" "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource
Organization (MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Graph1_boxplot_OLS-Maternaldata.gph", replace
}
tab mle_out_sign //Outlier sign - MLE
quietly {
keep if outlier ==1
gen caseidbidx = caseid + " " + string(bidx)
save "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Anaemia_data\MLE_outliers.dta", replace
}
*
CASE OF QUANTILE REGRESSION MODELS
****Fitting 75th quantile regression model
```

```
use "dhs_anaemia.dta", clear
qreg hemoglobin_level i.residential_status i.education_level total_fertility_rate curr_preg_duration
i.distance_hc i.bmi_category i.wealth_index i.age_category i.water_source, quantile(75)
predict yhat_q75
predict residuals, resid
gen squaredr_q75 = (residuals)^2
summarize squaredr_q75, meanonly11
display sqrt(r(mean))
estimates store qreg75
// Identify outliers
sum residuals, detail
display r(p75) + 1.5 * (r(p75) - r(p25)) //upper outlier cut off point
display r(p25) - 1.5 * (r(p75) - r(p25)) //lower outlier cut off point
gen outlier = residuals > r(p75) + 1.5 * (r(p75) - r(p25)) | residuals < r(p25) - 1.5 * (r(p75) - r(p25))
gen q75_out_sign = "+ve" if residuals > r(p75) + 1.5 * (r(p75) - r(p25))
replace q75_out_sign = "-ve" if residuals < r(p25) - 1.5 * (r(p75) - r(p25))
***Graphing the outliers
quietly{
graph box residuals, cwhisker marker(1, mlabel(caseid)) title(Q75 identified outliers)
graphregion(color(white))
graph save "Graph" "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource
Organization (MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Graph1_boxplot_Q75-Maternaldata.gph", replace
}
tab q75_out_sign //Outlier sign - Q75
quietly {
keep if outlier ==1
gen caseidbidx = caseid + " " + string(bidx)
save "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
```

```
THESIS\Project\Data_analysis\Anaemia_data\Q75_outliers.dta", replace
}
*
CASE OF ROBUST REGRESSION MODELS
****Fitting mm robust regression model***
use "dhs_anaemia.dta", clear
robreg mm hemoglobin_level i.residential_status i.education_level total_fertility_rate
curr_preg_duration i.distance_hc i.bmi_category i.wealth_index i.age_category
i.water_sourcepredict yhat_mm
predict residuals, resid
gen squaredr_mm = (residuals)^2
summarize squaredr_mm, meanonly
egen median_residuals = median(residuals)
display sqrt(r(mean))
estimates store mmrob
// Identify outliers
gen MAD_value = abs(residuals - median_residuals)
sum MAD_value, detail
gen std_residuals = 0.6745*residuals/r(p50)
sum std_residuals, detail
display r(p75) + 1.5 * (r(p75) - r(p25)) //upper outlier cut off point
display r(p25) - 1.5 * (r(p75) - r(p25)) //lower outlier cut off point
gen outlier = std_residuals > r(p75) + 1.5 * (r(p75) - r(p25)) | std_residuals < r(p25) - 1.5 * (r(p75)
- r(p25))
gen mmrob_out_sign = "+ve" if std_residuals > r(p75) + 1.5 * (r(p75) - r(p25))
replace mmrob_out_sign = "-ve" if std_residuals < r(p25) - 1.5 * (r(p75) - r(p25))
*Outliers graghing
quietly{
graph box std_residuals , cwhisker marker(1, mlabel(caseid)) title(MM-robust identified outliers)
graphregion(color(white))
```

```
graph save "Graph" "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource
Organization (MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Graph1_boxplot_MM-Maternaldata.gph", replace
graph box std_residuals, cwhisker marker(1, mlabel(hemoglobin_level)) title(MM-robust identified
outliers)
}
tab mmrob_out_sign //Outlier sign - MM-robust
quietly {
12keep if outlier ==1
gen caseidbidx = caseid + " " + string(bidx)
save "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Anaemia_data\mmrob_outliers.dta", replace
}
quietly {
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Anaemia_data"
***Merging the outlier files****
use "LTSrob_outliers.dta", clear //use LTS robust outlier data
mmerge caseidbidx using "Srob_outliers.dta", missing(nomatch)
mmerge caseidbidx using "mmrob_outliers.dta", missing(nomatch)
mmerge caseidbidx using "mrob_outliers.dta", missing(nomatch)
mmerge caseidbidx using "q90_outliers.dta", missing(nomatch)
mmerge caseidbidx using "q75_outliers.dta", missing(nomatch)
mmerge caseidbidx using "q50_outliers.dta", missing(nomatch)
mmerge caseidbidx using "q25_outliers.dta", missing(nomatch)
mmerge caseidbidx using "mle_outliers.dta", missing(nomatch)
save "mw_anaemia_outliers.dta", replace
}
```

```
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Data_analysis\Anaemia_data"
use "mw_anaemia_outliers.dta", clear
////Outliers identified by each model
tab mle_out_sign
tab q25_out_sign
tab q50_out_sign
tab q75_out_sign
tab q90_out_sign
13tab mrob_out_sign
tab mmrob_out_sign
tab srob_out_sign
tab ltsrob_out_sign
///Common outliers analysis
gen common_outlier = 1 if !missing(mle_out_sign) & !missing(q25_out_sign)
& !missing(q50_out_sign) & !missing(q75_out_sign) & !missing(q90_out_sign)
& !missing(mrob_out_sign) & !missing(mmrob_out_sign) & !missing(srob_out_sign)
& !missing(ltsrob_out_sign)
//validating common outliers and checking signs
Preserve
drop if common_outlier==.
keep caseidbidx mle_out_sign q25_out_sign q50_out_sign q75_out_sign q90_out_sign
mrob_out_sign mmrob_out_sign srob_out_sign ltsrob_out_sign
order caseidbidx mle_out_sign q25_out_sign q50_out_sign q75_out_sign q90_out_sign
mrob_out_sign mmrob_out_sign srob_out_sign ltsrob_out_sign
export excel using "common_outliers.xlsx", firstrow(variables) replace
restore
cd "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Data_analysis"
```

```
**Combined Box plots for the outliers
quietly {
graph combine "Graph1_boxplot_OLS-Maternaldata.gph" "Graph1_boxplot_Q25-
Maternaldata.gph" "Graph1_boxplot_Q50-Maternaldata.gph" "Graph1_boxplot_Q75-
Maternaldata.gph" "Graph1_boxplot_Q90-Maternaldata.gph" "Graph1_boxplot_M-
Maternaldata.gph" "Graph1_boxplot_MM-Maternaldata.gph" "Graph1_boxplot_S-
Maternaldata.gph" "Graph1_boxplot_LTS-Maternaldata.gph", title("Outliers detected by three
methods using maternal data") graphregion(color(white))
14graph export "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource
Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1. THESIS\Project\Full
thesis\Graph1boxplotsmaternaldata.png", replace
graph export "C:\Users\User\OneDrive - Malawi AIDS Counselling and Resource Organization
(MACRO)\Desktop\Other\MSc. Biost _ 2022\1.
THESIS\Project\Manuscripts\BoxPlots_Resid.png", as(png) name("Graph") replace
}
capture log close
```