STATISTICAL MODELLING OF THE ASSOCIATION BETWEEN DIETARY DIVERSITY, DIETARY PATTERNS AND NON-COMMUNICABLE DISEASES IN NAMIBIA.

A DISSERTATION SUBMITTED IN FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN SCIENCE (APPLIED STATISTICS)

OF

THE UNIVERSITY OF NAMIBIA

BY

LAINA TULIPOMWENE MBONGO

201180405

JANUARY 2024

MAIN SUPERVISOR: Prof. Lawrence Kazembe (University of Namibia (UNAM), Department of Computing, Mathematics and Statistical Sciences)

CO-SUPERVISOR: Prof. Lillian Pazvakawambwa (University of Namibia (UNAM), Department of Computing, Mathematics and Statistical Sciences)

Abstract

Globalization coupled with urbanization has placed a significant pressure on the food systems of many developing countries. This has led to lifestyle changes that have become one of the most important influences on dietary patterns. The nutritional transition has affected the dietary pattern and nutrient intake greatly and has led to a rise in the purchases and consumption of processed and convenience foods. Analysis in nutritional epidemiology typically examined diseases in relation to a single or a few nutrients or foods. However, people do not eat isolated nutrients. Instead, they eat meals consisting of a variety of foods with complex combinations of nutrients. The high degree of inter-correlation among nutrients as well as among foods makes it difficult to attribute effects to single dietary components. Dietary patterns can influence health and the risk of developing chronic conditions. Therefore, to gain full understanding of the relationship between diet and the development of non-communicable diseases (NCD), it is desirable to use several methodological approaches.

The main objective of this study was to explore the linkages between dietary patterns, dietary diversity and prevalence of non-communicable diseases. Specifically, the study aimed at: (i) applying count models on dietary diversity in Namibia, (ii) using bivariate count modelling approach in analyzing convenience and non-convenience consumption food preference in Windhoek, (iii) applying copula joint modelling of food insecurity indicators with application to food insecurity prevalence (FIP), household dietary diversity score (HDDS) and months of inadequate household food provisioning (MIHFP), (iv) fitting multiple indicators-multiple-cause modelling to examine the relationship between foods consumed and non-communicable diseases. The analysis used two representative survey data, namely the AFSUN-HCP Household Food Security Baseline Survey (2016) and Namibian Household and Income Expenditure (NHIES) of 2015/2016.

The study focused on dietary diversity by using different count models. The household dietary diversity score presented a mean score of 6.5, suggesting a moderate diverse diet, with less consumption of food made from beans/lentils; eggs; fruits/vegetables and more consumption of starch food. Determinants for household dietary diversity included educational level, sex of head of household and main source of income (p-value <0.005). The study further used bivariate

modelling approaches to analyze the food consumption patterns. The results found that, whereas the consumption of food monthly was more on the non-convenience foods, the purchases of convenience was frequent on a weekly basis and in multiple food sources. Moreover, the study employed copula joint modelling of food security indicators. The findings show that AIC of the untruncated (conditional/marginal) Poisson regression model was lower and thus proved to fit the data better. The Frank Copula and Bivariate Normal Copula best fitted the data of establishing the relationship between HFIP and HDDS, and between HFIP and MIHFP respectively. Lastly, we analyzed multiple indicators-multiple causes examining the relationship between foods consumed and non-communicable disease. Principal Component Analysis (PCA) and Structural Equation Models (SEM) were used as data reduction methods to derive dietary patterns. Fruits, foods such as condiments/tea/coffee and potatoes, yams, cassava, or any foods made from roots and tubers accounted for majority of the variation.

The study concluded that the usage of appropriate methods for specific data types is very critical. Generalized Poisson Regression models through the usage copula approaches are best to analyze jointly two outcomes in order to test for significant relationships between high-level hierarchical effects (e.g., random effects). Specifically, the bivariate normal and the Frank Copula were found to fit the data best. The unique nature of the bivariate normal model is that it does not allow for a different dependence structure between the outcomes while the frank copula does not have tail dependence and it can model both positive and negative dependencies as the normal copula. SEM and PCA's were used as data reduction methods. Lastly, the study concludes that food and nutrition insecurity is a major threat to the development of the country and the study recommends for strengthened advocacy for consumption of healthy and diverse diets in the country in order to slow down and arrest proliferation of non-communicable diseases.

List of Papers Presented

1. **Mbongo L**., Kazembe L., and Pazvakawambwa L., Dietary Patterns and Linkages to Non-Communicable Disease in Namibia, presented at Academic council of the United Nations System (ACUNS): The UN and Africa: Progress towards achieving the Sustainable Development Goals (SDGs), Stellenbosch, Cape Town, 19-21 June 2019.
2. **Mbongo L**., Kazembe L., and Pazvakawambwa L**.,** Application of Count Models to Dietary Diversity in Namibia, presented at 2020 DELTAS Africa SSACAB Virtual Scientific Conference, 19-23 October 2020.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ACUNS	Academic Council of the United Nations System
AFSUN	African Food Security Urban Network
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
CMP	Conway-Maxwell Poisson
DASH	Dietary Approaches to Stop Hypertension
DELTAS	Developing Excellence in Leadership, Training and Science
DQI	Diet Quality Index
FAO	Food and Agricultural Organization
GJRM	Generalized Joint Regression Model
GLM	Generalized Linear Model
HEI	Healthy Eating Index
HDDS	Household Dietary Diversity Score
HFIAS	Household Food insecurity Access Scale
HFIP	Household Food Insecurity Prevalence
IDDS	Individual Dietary Diversity Score
LASSO	Least Absolute Shrinkage and Selection Operator
LMIC	Low- and Middle-Income Countries
MIHFP	Months of Inadequate Household Food Provisioning
MOHSS	Ministry of Health and Social Services
MSDPS	Mediterranean Style Dietary Pattern Score
NB	Negative Binomial
NCD	Non-Communicable Disease
NHIES	Namibia Household and Income Expenditure Survey
NSA	Namibia Statistics Agency
ONQI	Overall Nutritional Quality Index
PAHO	Pan American Health Organization
PCA	Principal Component Analysis
PIG	Poisson Inverse Gaussian
PLSR	Partial Least Square Regression
PSU	Primary Sampling Unit
RDA	Recommended Dietary Allowances
RMSEA	Root Mean Square Error of Approximation
RRR	Reduced Rank Regression
SDG	Sustainable Development Goals
SEM	Structural Equation Model
SPSS	Statistical Package for Social Sciences
SRMR	Standardized Root Mean Square Residual
SSACAB-SUSAN	Sub-Saharan Africa Consortium for Advanced Biostatistics and Sub-Saharan Africa Network
UCPS	University of Namibia Centre for Postgraduate Studies
UREC	University of Namibia, Research and Ethic Committee
WHO	World Health Organization

Acknowledgements

I feel indebted to give my sincere gratitude to my Supervisors Prof. L. Kazembe and Prof. L. Pazvakawambwa for their tremendous effort invested. I appreciate that despite having many commitments, they devoted valuable time to this study. I'm specifically grateful that they were able to push me during my struggles.

The authors had financial support from the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

I am thankful to the Namibia Statistics Agency for allowing me to gain access to the NHIES 2015/16 datasets and the Department of Statistics and Population Studies for granting me access to the AFSUN-HCP Household Food Security Baseline Survey (2016) for Windhoek. The usage of these datasets highly assisted me to achieve the objectives of this study.

Last but not least, I wish to thank the Almighty God for giving me strength, accompanying me this far, and granting me His endless mercies.

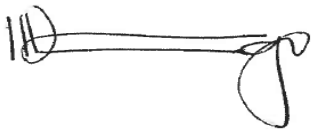
Dedication

This dissertation is dedicated first to my daughter, Felicity Marsha Nangolo. Thank you for your understanding and for sharing your time with my work.

Secondly, I wish to dedicate this thesis to my ever-loving family, your support has been my strength. To my mother, Lavinia (Stephanus) Mbongo and late father Thomas Abraham Mbongo, thank you for believing in us all and for giving us the best education despite the many circumstances.

Declarations

I, Laina Tulipomwene Mbongo, hereby declare that this study is my own work and is a true reflection of my research, and that this work, or any part thereof has not been submitted for a degree at any other institution.



25 January 2024

Laina Tulipomwene Mbongo

CHAPTER 1: INTRODUCTION

1.1. Background of the Study

Globalization driven by advances in transportation and telecommunications, and a positive political climate has created a global economy characterized by unprecedented levels of urbanization, more and bigger cities than ever before (Henderson, Cities & Development, 2010). In the low-income countries, urbanization is often associated with changes in food habits, westernization of dietary patterns, lower levels of physical activity, and overall changes in lifestyle (Kennedy & Shetty, 2004). Eating between main meals, buying street foods or processed foods and eating outside the home are frequent practices of urban citizens (Becquey, et al., 2010). Coupled with this, consumption of meat, fat, salt and sweetened products has increased (Becquey, et al., 2010) (Nickanor, 2014).

The rising occurrence of non-communicable diseases (NCDs), including cancer, hypertension, cardiovascular disease, and chronic obstructive pulmonary disease, poses an escalating public health challenge in the Global South (Abegunde, et al., 2007). The prevalence rates surpass those in numerous Global North countries, and the burden of these diseases has now exceeded that of extensively documented infectious ailments like HIV and AIDS and tuberculosis. According to the World Health Organization (WHO), there is a projected 17% increase in NCDs in Africa from 2013 to 2025. Noncommunicable diseases (NCDs) kill 41 million people each year, equivalent to 74% of all deaths globally. Each year, 17 million people die from NCD before age 70; 86% of these premature deaths occur in low- and middle-income countries. Of all NCD deaths, 77% are in low- and middle-income countries. Cardiovascular diseases account for most NCD deaths, or 17.9 million people annually, followed by cancers (9.3 million), chronic respiratory diseases (4.1

million), and diabetes (2.0 million including kidney disease deaths caused by diabetes) (WHO, 2023). NCDs threaten progress towards the 2030 Agenda for Sustainable Development, which includes a target of reducing the probability of death from any of the four main NCDs between ages 30 and 70 years by one third by 2030.

Studies of dietary patterns have become popular in nutritional epidemiology (Smith, Emmett, Newby & Northstone, 2011). Dietary patterns facilitate studies of the whole diet, recognizing that people consume foods in combination (Hox, Moerbeek & Van De Schoot, 2017). They are therefore said to complement traditional methods of examining diet-health relationships which look at individual foods or nutrients. Dietary patterns consider the complex interrelationships between different foods or nutrients as a whole, reflect individuals' actual dietary habits, and provide more information to indicate when many nutrients are associated with diseases. Moreover, these patterns exhibit greater consistency over time and exert a more significant impact on health outcomes compared to isolated nutrients (Zhao, et al., 2021). Nutrition is pivotal for the health and growth of every individual. A diet lacking in essential nutrients can lead to serious repercussions, contributing to health issues like obesity, cardiovascular diseases, type 2 diabetes, and cancer (Moe, et al., 2022).

Although the essential role of nutrition in the promotion and enhancement of the overall quality and span of life is widely recognized, many Namibians suffer from nutritional disorders that are due to an inadequate food intake, in terms of quality and quantity. These nutrition problems are related to diets, which are monotonous, deficient in food energy, and contain few foods that are rich in vitamins and minerals ((Ministry of Health and Social Services, 2013, (Mbongo, 2017)). Though urban dwellers have better access to health services, water and sanitation facilities and

education, non-communicable diseases (NCD's) have been on the rise in urban communities recently (Nickanor, 2014). The three major dietary patterns identified in Windhoek through PCA are starch-oil-sugar, meat-fish-dairy, and fruit-vegetable. Based on our findings, "starch-oil-sugars" and "meat-fish-dairy" dietary patterns are related to negative health outcomes, while "fruit-vegetables" are not (Kazembe, Nickanor, & Crush, 2021).

1.2. Problem Statement

Urban food systems in the southern hemisphere are undergoing major transformations (Von Braun , Meinzen-Dick, Rosegrant & Nin-Pratt, 2008). These changing food systems are intersecting with urban poverty to produce new forms of urban food and nutrition insecurity (Battersby, 2013). Namibia is witnessing an increasing burden of non-communicable diseases. Cardiovascular diseases (CDVs) are the most common NCDs in Namibia, accounting for 21% of mortality in 2012. Traditionally, analyses in nutritional epidemiology examined diseases in relation to a single or a few nutrients or foods. However, people do not eat isolated nutrients. Instead, they eat meals consisting of a variety of foods with complex combinations of nutrients. The high degree of inter-correlation among nutrients as well as among foods, makes it difficult to attribute effects to single dietary components. Dietary patterns can influence health and the risk of developing chronic conditions. Therefore, to gain full understanding of the relationship between diet and the development of NCDs, it is desirable to use several methodological approaches. One of the statistical exploratory methods that accomplishes pattern derivation is the Principal Component Analysis (PCA). Applied to food intake, PCA aims to explain the total variation in the intake of many foods or food groups in terms of a few linear functions called principal components,

examining the totality of the diet, and considering multiple components simultaneously as a dietary pattern.

Furthermore, the issues of chronic food insecurity, poverty, and malnutrition continue to be fundamental human welfare challenges in developing and developed countries (Suresh, Shailendra & Prabuddha, 2014). A diverse diet with foods from all food groups is necessary for population groups to meet their requirements for essential nutrients. Count data arise frequently in nutritional analyses, but regularly violate the equi-dispersion constraint imposed by the most popular distribution for analyzing these data, the Poisson distribution. A need for count model exploration is thus required. Several measurement issues still need to be addressed to improve diet assessment. Measurement error affects dietary pattern definitions and the assessment of food or nutrient intake.

1.3. Aim and Objectives of the Study

The overall aim of this study is to explore the linkages between dietary patterns, dietary diversity, and the prevalence of non-communicable diseases. Four study objectives are defined:

a) Apply count models to examine factors associated with Household dietary diversity in Namibia.
b) Employ a bivariate count modelling approach to analyze convenience and non-convenience consumption of food preferences in Windhoek.
c) Apply copula joint modelling of food insecurity indicators with application to food insecurity prevalence (FIP), household dietary diversity score (HDDS), and months of inadequate household food provisioning (MIHFP)
d) Fit multiple indicators-multiple-cause modelling to examine the relationship between foods consumed and non-communicable diseases.

1.4. Significance of the Study

The foundation of Namibia's health services has been the Health Policy Framework of 1998, aimed at the attainment of a level of health and social well-being by all Namibians, which will enable them to lead economically and socially productive lives. Namibia's Vision 2030 aspires to have a *"a healthy and food-secured nation in which all preventable, infectious and parasitic diseases are under secure control; people enjoy a high standard of living, good quality life and have access to quality education, health and other vital services. All of these translate into long life expectancy and sustainable population growth".* Additionally, in order to ensure that the health of Namibians is prioritized, Namibia has vested interest in achieving the targets of the UN 2030 Sustainable Development Goals (SDGs). Specifically, and of interest to this study is Goal 3 on ensuring healthy lives and promoting well-being for all at all ages. Thus, this study is essential in informing policies and programs about the health and well-being of Namibians, particularly by addressing health issues such as non-communicable diseases that are attributable to diet choices.

1.5. Limitation of the Study

Due to time and money constraints, this study was only restricted to secondary datasets. In-depth interviews (qualitative) to obtain explanations to specific behaviors and trends could not be conducted. Additionally, due to the complicated nature of the Poisson (bivariate) distribution, some applications in this study were limited. For future researchers, it is important that an inclusion of qualitative approaches form part of their studies as a learning outcome from the limits in this research. Furthermore, more models to handle bivariate and complex data need to be developed.

1.6. Delimitation of the Study

The study used a combination of surveys, including the Namibia Household and Income Expenditure Survey (NHIES) of 2015/16, which was used to model health outcomes (non-communicable diseases) and quality of life. The Food Insecurity Dataset of 2016/17 of Windhoek by the African Food Security Urban Network (AFSUN) was used to model food security and nutrition related aspects.

1.7. Structure of the Study

This research is structured into 7 chapters. Chapter 1 presents the background to food insecurity, dietary diversity and NCDs; the problem statement; objectives and the significance of the study. Chapter 2 is a literature review of the methods of dietary pattern analysis, namely a priori approaches, empirical methods, and hybrid methods. The application of correlational analysis in dietary patterns, path analysis and structural equation models, measurement errors in dietary assessment, and dietary diversity are also discussed. The goal of this chapter is to provide the reader with a general understanding of dietary diversity and NCDs, and to lay some groundwork to the concepts and statistical methods used in food security, dietary diversity, and NCDs that helped the development of the subsequent chapters.

In Chapter 3 count models are applied to dietary diversity in Namibia using cross-sectional survey data of the NHIES of 2015/2016. Chapter 4 focuses on the application of a bivariate count modelling approach in analyzing convenience and non-convenience consumption of food preferences using the Windhoek AFSUN data of 2016. Chapter 5 provides copula joint modelling of food insecurity indicators with applications to food insecurity prevalence, household dietary diversity score and months of inadequate household food provisioning. Chapter 6 shows how

multiple-indicator-multiple cause modelling is used to examine the relationship between foods consumed and NCDs. Chapter 7 revisits the primary objectives of the research in order to evaluate if they have been achieved. The chapter presents the conclusions and outlines recommendations for improvements and future studies. After Chapter 7, an appendix that includes all R-programmes used in this dissertation is provided. A bibliography of all references is given at the end of dissertation.

CHAPTER 2: LITERATURE REVIEW AND DATA OVERVIEW

2.1. Introduction

Dietary patterns are conceptualized and defined in many ways: as an exposure or a behavior, by numbers or labels, as univariate or multivariate constructs, as research-driven or data-driven, and as static or dynamic (Jill, Amy, Stephanie, & Susan, 2018). Dietary patterns can be defined as the quantities, proportions, variety, or combination of different foods and drinks in diets, and the frequency with which they are habitually consumed.

Dietary patterns can influence health and the risk of developing chronic conditions. For example, adherence to a Mediterranean-style dietary pattern, characterized as a diet emphasizing plant foods, especially legumes and nuts, fish, and olive oil, appears to promote health and reduce the risk of developing cardiovascular disease (Estruch et al., 2013). Consumption of a Western dietary pattern, characterized as a diet high in fat, sugar, and refined grains, and low in fruits and vegetables, is associated with obesity, diabetes mellitus, metabolic syndrome, and hypertension (Cordain et al., 2005).

Consumption of inadequate quantities and poor quality of foods by households results in nutrient deficiencies (Mirmiran, Azadbakht, Esmaillzadeh, & Azizi, 2014). According to Vakili et al. (2013), dietary diversity can be viewed as a proxy measure of food security. Dietary diversity has been estimated to have a greater potential of meeting nutrient requirements because no single food can have all nutrients (Labadarios, Steyn, & Nel., 2011).

The interest in Structural Equation Models (SEMs) is often on theoretical constructs, which are represented by the latent factors (Hox, Moerbeek, & Van De Schoot, 2017). SEM have the ability to combine both measurement and structural considerations. It integrates psychometric concepts

(i.e., measurement approaches) and the econometric ideas (structure approaches). Thus, this method has the ability to take into account measurement errors. As for the structure approaches in SEM, path analysis is applied to estimate the relationships among latent constructs. The ability to combine these two analyses is one of the advantages of SEM. By specifying and describing the plausible relationships between latent concepts and manifest variables, associated measurement errors and proposed structural relationships among latent structures in SEM can effectively estimate parameters simultaneously, which mirror the fact that the variables coexist in reality.

2.2. Methods of Dietary Pattern Analysis

Three categories of dietary pattern assessment methods exist: theoretical methods, empirical methods, and hybrid methods.

2.2.1. A Priori Approaches

Theoretical methods, also known as *apriori* methods, assess diet based on prior knowledge and scientific evidence such as the dietary guideline index (Thorpe, Milte, Crawford, & McNaughton , 2016). Dietary indices are the most common hypothesis-oriented approaches that evaluate the adherence of population intake to nutritional recommendations and represent a measure of "healthy" eating patterns and are known by various names, including diet quality indices or healthy eating indices. These indices or scores have certain advantages over data-driven dietary pattern approaches. They are based on existing knowledge of optimal dietary patterns and provide a clear nutritional benchmark. Consequently, diet indices may be easy to interpret and may therefore be more easily understood by the public (Thorpe et al., 2016).

Currently, there are 2 commonly used diet scores: The Healthy Eating Index (National Center for Health Statistics, 2020) and the Revised Diet Quality Index (Burggraf, Teuber, Brosig, & Meier, 2018). Both of these are based on the U.S. dietary guidelines and include both food and nutrient-based indicators. Although they have been adapted for use in other countries by altering the cut-offs, there has been little adaptation of the range of indicators to reflect the dietary guidelines in other countries or to focus solely on food-based indicators (National Center for Health Statistics, 2020).

The Healthy Eating Index (HEI)

The Healthy Eating Index (HEI) was developed by Kennedy, Ohls, Carlson, & Fleming (1995) to investigate American eating habits and their compliance with the dietary guidelines provided by the Recommended Dietary Allowances (RDA) of the U.S. Departments of Agriculture (USDA) and of Health and Human Services in 1980. The HEI is updated every five years.

The original Healthy Eating Index analyzes five food and nutrient groups, namely grains, vegetables, fruits, milk and dairy products, and meats, which receive a score of 0 to 10 according to the number of servings consumed from each group. Diet variety and some nutrients, such as total fat, saturated fat, cholesterol, and sodium, are also scored 0 to 10 points. Diet quality increases with the score as shown in Table 1 (1990 version), Table 2 (2005 version) and Table 3 (2010 version).

Table 1: Components and score standards version HEI-1990 (Kennedy et al., 1995)[21]

Components of the Healthy Eating Indexes	Score Points	Maximum (10 points)	Minimum (0 points)
Food groups			
Grains	0 to 10	6-11 servings	0 serving (no intake)
Vegetables	0 to 10	3-5 servings	0 serving
Fruits	0 to 10	2-4 servings	0 serving
Milk	0 to 10	2 - 3 servings	0 serving
Meat	0 to 10	2 - 3 servings	0 serving
Recommendations			
Total fat	0 to 10	≤30% of the tei[1]	≥45% of the tei
Saturated fat	0 to 10	less than 10% of the tei	≥15% of the tei
Cholesterol	0 to 10	less than 300 mg	≥of 450 mg
Sodium	0 to 10	less than 2.4 g	≥of 4.8 g
Diet variety	0 to 10	intake of 16 types of foods over three days	intake of 6 or fewer types of foods over three days

The subsequent HEI updates, namely HEI-2005 (Guenther et al., 2008) and HEI-2010 (Guenther et al., 2013), changed the food groups and nutrients but maintained the direct relationship between diet quality and score. The current subdivision is represented by food groups separated by compliance criteria and intake moderation.

Compliance parameters included total fruit (including juices); whole fruits (except juices); total vegetables (including all types of beans and peas not included in the total protein sources); green vegetables and beans (including all types of beans and peas not included in the total protein sources); whole grains; milk and dairy products; total protein sources; seafood and plant protein (including nuts, seeds, and soybean products); and fatty acids (the *ratio* between poly- and monounsaturated fatty acids to saturated fatty acids). Moderation parameters include refined grains, sodium, and empty calories (solid fats, alcohol, and added sugar) (Baiocchi de Carvalho, Dutra, Pizato, Gruezo, & Ito, 2014).

Table 2*:* Components and score standards version HEI-2005 (Guenter et al., 2008)[29]

Components of the Healthy	Score		
Eating Indexes	**Points**	**Maximum (5 - 10 points)**	**Minimum (0 points)**
Group adequacy			
Total fruits	5	≥0.8 cup *per*1,000 kcal	no intake
Whole fruits	5	≥0.4 cup *per* 1,000 kcal	no intake
Total vegetables	5	≥1.1 cups *per*1,000 kcal	no intake
Dark green and orangish	5	≥0.4 cup *per* 1,000 kcal	no intake
Vegetables and legumes			
Total grains	5	≥3.0 ounces[2] *per* 1,000 kcal	no intake
Whole grains	5	≥1.5 ounces *per* 1,000 kcal	no intake
Milk	10	≥1.3 cups *per* 1,000 kcal	no intake
Meat and beans	10	≥2.5 ounces *per* 1,000 kcal	no intake
Oils	10	≥12 g *per* 1,000 kcal	no intake
Group moderation			
Saturated fat	10	≤0.7% of the tei	≥15% of the tei
Sodium	10	≤0.7 g *per* 1,000 kcal	≥2.0 g per 1,000 kcal
SoFAAS calories	20	≤20% of the tei	≥50% of the tei

The scores are still based on the amount of energy coming from each group expressed as energy density (serving/1,000 kcal). For the compliance parameter, the scores are highest when intake equals or exceeds the RDA. In the moderation parameter, the maximum score indicates an intake equal to or below the RDA.

Generally, the HEI assesses diet quality and appropriateness of consumed food groups and nutrients. The results may indicate a need of nutritional interventions for specific groups or populations.

Table 3: Components and score standards version HEI-2010 (Guenter et al., 2013)[30]

Components of the Healthy	Score		
Eating Indexes	Points	Maximum (5 - 10 points)	Minimum (0 points)
Group adequacy			
Total fruits	5	≥0.8 cup *per* 1,000 kcal	no intake
Whole fruits	5	≥0.4 cup *per* 1,000 kcal	no intake
Total vegetables	5	≥1.1 cups *per* 1,000 kcal	no intake
Green vegetables and beans	5	≥0.2 cup *per* 1,000 kcal	no intake
Whole grains	10	≥1.5 ounces *per* 1,000 kcal	
Dairy products	10	≥1.3 cups *per* 1,000 kcal	no intake
Total protein sources	5	≥2.5 ounces *per* 1,000 kcal	no intake
Seafood and plant protein	5	≥0.8 ounces *per* 1,000 kcal	no intake
Fatty acids	10	(Pufas + Mufas)/saturated >2.5	no intake
Group moderation	10	≤1.8 ounces *per* 1,000 kcal	≥4.3 ounces per 1,000 kcal
Refined grains	10	≤1.1 g *per* 1,000 kcal	≥2.0 g *per* 1,000 kcal
Sodium	20	≤19% of the tei	≥50% of the tei
Empty calories			

Note: [1]Total energy intake; [2]Ounce: 1 ounce equals 28.35 g. SoFAAS: Solid Fats, Alcoholic beverages, and Added Sugar; PUFA: Polyunsaturated Fatty Acids; MUFA: Monounsaturated Fatty Acids.

The original Diet Quality Index (DQI)

The original Diet Quality Index (Patterson, Haines, & Popkin, 1994) was developed to assess the intake of eight food groups and the recommendations of the Committee on Diet and Health of the National Research Council Food and Nutrition Board (National Research Council, 1989) and of the United States government. Drewnowski *et al,* (1997) presented a simplified DQI version (DQI-a I), adapted for a dietary survey in France. In this adaptation, a maximum score of five points reflected the following attributes: (1) less than 30% energy from fats; (2) less than 10% energy from saturated fats; (3) cholesterol intake up to 300 mg per day; (4) more than 50% energy from carbohydrates; and (5) less than 10% energy from sucrose. These parameters were assessed in absolute terms (yes/no). In addition to these five DQI elements, the authors scored variety (dietary

variety) and diversity (number of food groups present in the diet). Thus, this index excluded the parameters protein, sodium, calcium, and fruit and non-starchy vegetable servings, present in the original version. These scores can be used to describe overall dietary characteristics and repeat or compare results across populations. Many dietary quality scores have significant associations with disease and mortality outcomes. The total score is easy to understand and use, and the summing process is simpler than in other statis tical methods for dietary pattern analysis.

According to de Carvalho, Dutra , Pizato, Gruezo, & Ito, (2014) in a second adaptation (DQI-a II), this same group of researchers modified one of the five DQI-a I score points, replacing the parameter 'less than 10% energy from sucrose' by the parameter 'sodium intake below 2,500 mg'. Again, dietary variety score was assessed separately. The authors used this instrument to compare the dietary patterns of American youth and older individuals by recording food intake during fourteen consecutive days. A significant association was not found between diet variety and DQI-a II, suggesting that the component 'variety' proposed by this instrument still presented limitations.

Later, changes in the American Food Guide Pyramid and Dietary Reference Intakes (DRI) were included in the Revised DQI (DQI-R), which introduced the measurement of food proportionality, moderation, and variety as dietary quality parameters (Haines, Siega-Riz, & Popkin, 1999). Proportionality regards the recommendation of consuming a higher number of servings of certain food groups and a fewer number of other food groups. The parameter moderation involves limiting the intake of food components that contribute to health risk, such as fat, salt, and sugar. Finally, variety includes inter- and intragroup variety, consequently, consumption of more food components. This instrument, with a maximum score of 100 points, is based on the food pyramid and the DRIs for calcium and iron.

Table 4: The DQI-R components: (Patterson, Haines, & Popkin, 1994)

Component	Score	Assessment Levels
Total fat ≤30% of the energy consumed	0-10 points	≤30% = 10
Total fat ≤30% of the energy consumed.	0-10 points	>30%; ≤40% = 5
Saturated fat ≤10% of the energy consumed	0-10 points	>40% = 0
		≤10% = 10
Saturated fat ≤10% of the energy consumed.	0-10 points	>10%; ≤13% = 5
Dietary cholesterol <300 mg/day	0-10 points	>13% = 0
		≤300 mg = 10
Dietary cholesterol <300 mg/day	0-10 points	>300 mg; ≤400 mg = 5
Two to four servings of fruits per day, % of recommended servings	0-10 points	>400 mg = 0
		≥100% = 10
Two to four servings of fruits per day, % of recommended servings	0-10 points	99% - 50% = 5
Three to five servings of non-starchy vegetables per day, % of recommended servings	0-10 points	<50% = 0
		≥100% = 10
Three to five servings of non-starchy vegetables per day, % of recommended servings	0-10 points	99% - 50% = 5
Six to eleven servings of grains per day, % of recommended servings	0-10 points	<50% = 0
		≥100% = 10
Six to eleven servings of grains per day, % of recommended servings	0-10 points	99% - 50% = 5
Calcium intake as % AI by age, % of recommended intake	0-10 points	<50% = 0
		≥100% = 10
Calcium intake as % AI by age, % of recommended intake	0-10 points	99% - 50% = 5
Iron intake as % RDA by age	0-10 points	<50% = 0
		≥100% = 10
Iron intake as % RDA by age	0-10 points	99% - 50% = 5
Diet diversity score	0-10 points	<50% = 10
		≥6
Diet diversity score	0-10 points	≥3; <6
Food moderation score	0-10 points	<3
		≥7
Food moderation score	0-10 points	≥4; <7
		<4

Mediterranean Diet Score

The term 'Mediterranean Diet' refers to the dietary pattern found in areas that produce olive oil. The first scientific evidence that Mediterranean populations had lower incidence of cardiovascular diseases appeared in the 1960s, which was attributed to a non-westernized diet (Baiocchi de Carvalho, Dutra, Pizato, Gruezo, & Ito, 2014).

According to Baiocchi de Carvalho, Dutra, Pizato, Gruezo, & Ito (2014), the Mediterranean Diet is characterized by high intake of olive oil (main source of lipids), non-starchy vegetables, legumes, whole grains, and fruits, including nuts; moderate intake of poultry and fish (depending on proximity to the coast); low intake of whole milk and dairy products and red meats; and low to moderate intake of wine as the main source of alcohol during the meals. Over time, other foods were incorporated because more information regarding the traditional Mediterranean Diet of reference became available, which included less typical foods such as eggs, animal fats, margarine, beverages with added sugar, cakes, pies, cookies, and sugar.

Generally, indices that estimate adherence to the Mediterranean Diet were constructed using deduction, that is, by combining specific components ordered by cut-off points, later added to compose a final score. The number of components (foods, food groups, or a combination of nutrients, foods, and food groups); classification categories for each component; assessment scales; statistical parameters (mean, median, or daily intake cut-off amounts); and the positive or negative contribution of each component to the total score varied greatly between indices, resulting in an important variation in internal consistency (Mila-Villarroel et al., 2011).

The Mediterranean-Style Dietary Pattern Score (MSDPS) was developed for the American population and is based on the recommended amounts of the 13 food groups of the Mediterranean Diet Pyramid (Willet, et al., 1995). Each food group receives a score from 0 to 10 according to compliance with the recommended intakes. The score decreases as the degree of exceedance of the recommendations increases. The maximum score was standardized to 100 points and weighted proportionally to the energy intake from Mediterranean diet foods, that is, these foods received higher weights.

Overall Nutritional Quality Index (ONQI)

Yale University researchers proposed a way for assessing the overall nutritional quality of food ranges called Overall Nutritional Quality Index (ONQI). The Overall Nutritional Quality Index scores foods from 1 to 100 based on their nutritional characteristics. Higher scores mean lower risk of NCD (Katz et al., 2009).

The algorithm, based on a literature review, selected nutrients based on scientific evidence of their health effects. Nutrients with beneficial health effects are placed in the numerator and those with detrimental health effects are placed in the denominator. Therefore, higher values reflected higher ONQI score. Table 5 shows the main nutrients included in the algorithm. Some nutritional factors, such as fat and protein quality, were placed in the numerator, while energy density and glycemic load were placed in the denominator.

Table 5*:* Nutrients and macronutrient factors of the Overall Nutritional Quality Index (ONQI) algorithm

Nutrients in the numerator (beneficial)	**Nutrients in the denominator (not beneficial)**	**Macronutrient factors**
Fiber	Saturated fat	Fat quality
Folate	Trans fat	Protein quality
Vitamin A, C, D, E, B6, and B12	Sodium	Energy density
Potassium	Sugar (total/added)	Glycemic load
Calcium	Cholesterol	
Zinc		
Omega-3 fatty acids		
Bioflavonoids		
Carotenoids		
Magnesium		
Iron		

In addition to the above-mentioned nutritional aspects, another key element of this algorithm was the use of a dietary parameter. The dietary contribution of each food was also assessed. In other words, the researchers determined how the nutrient content of a food contributed to its daily

requirement. Furthermore, this parameter also indicated how consuming a food could affect the recommended intake of its other nutrients. This parameter was named Trajectory Score (TS) (Katz et al., 2009).

Dietary Approaches to Stop Hypertension (DASH)

The DASH diet is a lifelong approach to healthy eating that is designed to help treat or prevent high blood pressure (hypertension). The DASH diet encourages you to reduce the sodium in your diet and eat a variety of foods rich in nutrients that help lower blood pressure, such as potassium, calcium, and magnesium.

The DASH trial, originally published in 1997, reviewed the impact of eating patterns on blood pressure management. Specifically, subjects were fed either a diet rich in fruits and vegetables or a combination diet that was both rich in fruits and vegetables and low-fat dairy foods and low in saturated fat. Table 6 and 7 provides an outline of the recommended number of servings from each food group, based on two different levels of caloric intake 1,600 and 2,000 calories per day.

Table 6*:* Recommended servings in the Dash Eating Pattern by calorie level

		Recommended Servings/Day in the Dash Eating Pattern	
Food Groups	***Serving Size***	***1.600 calories/day***	***2,000 calories/day***
*Grains**	1 slice bread 1 oz dry cereal ½ cup cooked rice, pasta, or cereal	6	6-8
Vegetables	1 cup raw, leafy vegetables ½ cup cut-up raw or cooked vegetables ½ cup vegetable oil	3-4	4-5
Fruits	1 medium fruit ¼ cup dried fruit ¼ cup fresh, frozen, or canned fruit ½ cup fruit juice	4	4-5
Fat-free or low-fat milk and milk	1 cup milk or yoghurt 1 ½ oz cheese	2-3	2-3

Food-based dietary indices have a number of advantages over those based on food and nutrient intakes. They retain the complexity of food intake and indirectly assess intakes of nutrient and no nutrient components in food (Kant, 1996). In addition, developing a food-based score may lend itself to further adaptation to short methods of dietary assessment that may be particularly relevant for use in monitoring and surveillance activities (Rafferty, Anderson, McGee, & Miller, 2002).

Table 7: Recommended servings in the Dash Eating Patter by calorie level

Food Groups	***Serving Size***	***1.600 calories/day***	***2,000 calories/day***
Lean meats, poultry, and fish	1- oz cooked meat, poultry, or fish 1 egg	3-6	≤6
Nuts, seeds, and legumes	1/3 cup or 1 ½ oz nuts 2 Tbsp peanut butter 2 Tbsp or ½ oz seeds ½ cup cooked legumes	3 per week	4-5 week
Fats and oils	1 tsp soft magarine 1 tsp vegetable oil 1 Tbsp mayonnaise 2 Tbsp salad dressing	2	2-3
Sweets and Added sugars	1 Tbsp sugar, jelly, or jam ½ cup sorbet or gelatin 1 cup lemonade	0	≤5 per week

Note. Whole grains are preferable for most servings because of their higher fiber and nutrient content

2.2.2. Empirical methods (data driven approaches)

Empirical methods also known as *a posteriori* method, use statistical approaches to provide information about existing dietary patterns within the population (Thorpe, Milte, Crawford, & McNaughton, 2016). A number of approaches are used to generate *a posteriori* diatary patterns as outlined below.

2.2.2.1. Exploratory Factor Analysis (EFA)

Exploratory factor analysis is a statistical approach that can be used to analyse interrelationships among a large number of variables and to explain these variables in terms of a smaller number of

common underlying dimensions. This involves finding a way of condensing the information contained in some of the original variables into a smaller set of implicit variables (called factors) with a minimum loss of information (Zaiontz, 2018).

Factor analysis is based on a correlation table. If there are *k* items in the study, then the correlation table has $k \times k$ entries of form r_{ij} where each r_{ij} is the correlation coefficient between item i and item j. The main diagonal consists of entries with value 1. Factor analysis is based on various concepts from Linear Algebra, in particular eigenvalues, eigenvectors, orthogonal matrices and the spectral theorem (Zaiontz, 2018).

Empirical analysis and understanding of Eigenvalues and Eigenvectors

The main aim of a factor analysis is to extract the factors, compute the scores for each factor and to interpret the factors. It is thus from this background that a factor analysis must be simple and interpretable (Suresh et al., 2014). A factor analysis involves two steps. First, based on the correlation matrix for all variables, the appropriateness of the factor analysis is evaluated. Second, it is necessary to decide which factor model should be used, the number of factors to be extracted and to assess how well the model fits the data. The criteria used to extract the factors can be maximizing variance or minimizing residual correlations (Suresh et al., 2014).

Suppose that S is a symmetric matrix of order p. If v is also a column vector of order p and λ is a scalar such that $Sv = \lambda v$, then λ is called the eigenvalue (also called characteristic root) of S and v is the corresponding eigenvector (also called characteristic vector) of S. If $v'v = 1$, then v is said to be the standardized eigenvector of S. The eigenvalues can be arranged as diagonal entries in a diagonal matrix D (a square matrix where all the diagonal elements are zero) with the corresponding eigenvectors arranged as columns in V. Matrix V is orthogonal meaning:

$V'V = I$. Thus, the eigen structure of S can be written in matrix form as:

$$SV = VD \text{ (also known as the eigenequation)}$$

Additionally, it is also the case that:

$$S = VDV'$$

Properties of Eigenvalues

1. If all the p eigenvalues of a symmetric matrix S are positive, then S is termed positive definite.

2. If any eigenvalue is zero, then S is singular.

3. The sum of diagonal elements of a symmetric matrix S is equal to the sum of its eigenvalues.

The most frequently used method for factor analysis is principal component factor extraction. The number of factors to be extracted is somewhat arbitrary but can be based on the following for principal components.

- Eigenvalue ($\lambda_j > 1$), i.e., factors with a variance less than 1 are no better than a single variable.
- Scree test criterion: this is done by plotting eigenvalues against the number of factors in their order of extraction. The number of extracted factors is determined by the point on the curve where the slope becomes horizontal. This point indicates the maximum number of factors to be extracted.

It is important to interpret the factors. This is done through a factor rotation procedure, which simplifies the factor structure giving more insight into each factor. The simplest case of rotation is the orthogonal procedure in which the axes are maintained at 90°. Varimax is one common

rotational method under the orthogonal procedure and maximizes the sum of the variances of the required loadings of the factor matrix S (Suresh et al., 2014).

Finally, the standardized variables $Z_1, Z_2, \ldots . Z_j$ can be formed as linear combinations of the factors $F_1, F_2, \ldots . F_k$. The magnitude of each of the coefficients in the factor loading matrix is a weight measuring the importance of the j^{th} j variable Z_j to the k^{th} factor Z_k. Scores on factors can be estimated once the factor loading matrix is available. These scores are standardized to have zero mean and unit standard deviation (Suresh et al., 2014).

2.2.2.2. Principal Component Analysis (PCA)

Principal Component analysis, or as commonly known PCA, is a dimensionality-reduction approach that is mostly used to reduce the dimensionality of large data sets. The reduction in datasets is done by transforming a large set of variables into smaller ones with minimal loss of information. Principal component analysis (PCA) and cluster analysis (CA) are two commonly applied empirical dietary pattern methods (Thorpe, Milte, Crawford, & McNaughton, 2016). PCA uses the correlation matrix of food intake variables to identify common patterns of food consumption within the data in order to account for the largest amount of variation in diet (Thorpe et al., 2016). Like in factor analysis, PCA are mostly used when researchers have a large number of potential variables to analyze and would like to summarize the information contained in those variables as efficiently as possible (Gleason et al., 2015).

PCA may be used for a variety of reasons (Gleason et al., 2015). It was noted that a researcher might use a data reduction technique like PCA or FA before collecting information from the full study sample in order to determine which questions to include in a survey instrument to best capture a particular construct of interest (Gleason et al., 2015). It is further noted that in a situation

like this, there may be many questions that capture some important aspects of the underlying concept. The researcher would collect data on all variables from a small subsample and conduct PCA or FA to identify the questions that best "hang together" and reflect the underlying construct. These inquiries would then be incorporated into a complete survey instrument that could be used to survey the entire target sample.

In other instances, the researcher might be working with information that has already been gathered on the entire sample of interest, but they want to exclude a sizable number of input variables to make any following analysis go more smoothly. More often than not, a limited set of summary measures might offer more helpful and understandable descriptive data on a certain underlying construct. A statistical issue known as multicollinearity, wherein high correlations among covariates make it difficult to identify their true relationship with the model's dependent variable, can be minimized if the constructs are intended to be used as covariates in a statistical technique like a regression model (Gleason et al., 2015).

The other method is the Varimax Orthogonal Rotation. The VARIMAX method of rotation is the most frequently used rotation method (Hair et al., 1998, as cited in (Suresh et al., 2014)). It minimizes the number of variables that have high loadings on a factor, so that the factors can be interpreted more easily. The relationship between the test points remains the same as before. However, the axes are altered to interpret the factors more easily (Suresh et al., 2014).

Assumptions of the Principal Component Regression

The assumptions are the same as those used in regular multiple regression:

- Linearity
- constant variance (no outliers)

- independence
- Since PC regression does not provide confidence limits, normality need not be assumed.

There are five steps mainly involved in conducting PCA and are discussed below:

1. Standardize the data sets.

The purpose of standardizing a range of continuous initial variables is so that each one of them contributes equally to the analysis (Jaadi, 2021). Importantly, the reason for conducting standardization first before PCA, is that PCA is sensitive regarding the variances of the initial variables. According to Karan (2021), the process of PCA identifies directions in which variances are greatest. That is, if all the variables are not standardized so that they have a mean of 0 and a standard deviation of 1, the variables with the largest scale will dominate the others when the components are calculated since the variance of a variable is measured on the same scale squared.

Standardizing can be done mathematically by subtracting the mean and dividing by the standard deviation for each value of each variable (Jaadi, 2021):

$$Z = \frac{value - mean}{standard\ deviation} \qquad (1)$$

2. Calculate the covariance matrix to identify correlations.

A covariance matrix is defined as a square matrix that displays the covariances between many different variables (Karan, 2021). The purpose of covariance matrix computation is to understand how the variables of the input data set are varying from the mean with respect to each other, that is, to Identify if there is any relationship between them (Jaadi, 2021).

According to Jaadi (2021), the covariance matrix, which has entries for all potential pairs of the initial variables, i is a $p \times p$ symmetric matrix (whereby p is the number of dimensions). For instance, for a 3-dimensional data set with three variables *x*, *y*, and *z*, the covariance matrix is a 3× 3 matrix of this nature:

$$\begin{bmatrix} Cov\,(x,x) & Cov\,(x,y) & Cov\,(x,z) \\ Cov\,(y,x) & Cov\,(y,y) & Cov\,(y,z) \\ Cov\,(z,x) & Cov\,(z,y) & Cov\,(z,z) \end{bmatrix} \quad (2)$$

Since the covariance of a variable with itself is its variance ($Cov(a,a) = Var(a)$), in the main diagonal (Top left to bottom right) we actually have the variances of each initial variable. Since the variables is commutative ($Cov(a,b) = Cov(b,a)$), the entries of the covariance matrix are symmetric with respect to the main diagonal which means that the upper and the lower triangular portions are equal (Jaadi, 2021).

The covariance matrix can be calculated as follows (Karan, 2021):

Covariance matrix for population (2 dimensions): $Cov(x,y) = \frac{\sum(X_i - \bar{X}) \times (Y_i - \bar{Y})}{N}$ *(3)*

Covariance matrix for sample (2 dimensions): $Cov(x,y) = \frac{\sum(X_i - \bar{X}) \times (Y_i - \bar{Y})}{N-1}$ *(4)*

3. Compute the eigenvectors and eigenvalues of the covariance matrix

The eigenvectors and eigenvalues are defined as non-zero vectors of a linear map, that when transformed, give rise to a scalar multiple of them (they do not change direction). The scalar is referred to as the eigenvector or eigenvalue and they help to identify the principal components (Karan, 2021).

The eigenvalues and eigenvectors associated with each eigenvalue of a matrix A

$$A\vec{v} = \lambda\vec{v} \quad (5)$$

is calculated by the following steps of Karan (2021):

(i) Calculate the roots of the characteristic polynomial of the matrix A

$$|A - \lambda/| = 0 \quad (6)$$

(ii) For each eigenvalue, determine all non-trivial solutions

$$|A - \lambda/|\vec{v} = 0 \quad (7)$$

Whereby:

A= square matrix

V= vector

λ = scalar value

4. Create a feature vector for Principal Component Analysis (PCA)

This step allows you to choose whether to keep all the components or discard those of lesser significance/ low eigenvalues, and form with the remaining ones a matrix of vectors that are called feature vector.

The feature vector is simply a matrix that has as columns the eigenvectors of the components that we decide to keep (Jaadi, 2021).

5. Recast the data along the principal component's axes

In the last step, we use the created feature vector to reorient the data from the original accessed to the once represented by the principal component. This is done by multiplying the transposed original data set by the transposed row feature vector and completing the process of Principal Component Analysis (Karan, 2021).

$$FinalDataset = FeatureVector^T * StandardizedOriginalDataSet^T \quad (8)$$

2.2.2.3. Correspondence Analysis (CA)

Principal component analysis (PCA) is expanded into correspondence analysis (CA), which is designed to handle nominal variables. The purpose of CA was initially to analyse contingency tables, which are data matrices with two non-negative entries and a sample of observations specified by two nominal variables. However, it was quickly expanded to analyse all data matrices with non-negative entries (Hervé & Michel, 2014). Conceptually comparable to PCA, CA analyses nominal or categorical data as opposed to continuous data. There are three significant distinctions between CA and PCA techniques. (1) PCA maximizes the amount of variance explained among measured variables, whereas CA maximizes the correspondence (measure of similarity of frequencies) between rows (represent measured variables) and columns (represent samples) of a table, (2) PCA assumes a linear relationship among variables, whereas CA anticipates a unimodal model, (3) In order to calculate the distances between samples in the complete ordination space of the CA, weighted Euclidean distance, a type of Euclidean distance, or chi-square distance are used. One disadvantage of CA is that it frequently results in an observable mathematical artifact known as the "arch" effect.

The basic idea of CA is that given K set of indicators $X^k, k = 1, \ldots . K$, with $X^k_{0,j}, X^k_{1,j}$ defining the binary responses corresponding to presence or absence of item X^k in household j. Put differently, $X^k_{0,j} = 1$ when $X^k_j = 0$ when the item is lacking at household j and $X^k_{1,j} = 1\ for\ X^k_j = 1$ for household j. Based on the a^{th} factorial axis is defined by:

$$F^a_j = \frac{1}{K}\sum_{k=1}^{K}\sum_{m_{k=0}}^{1} W^{a,k}_{m_k} X^k_{m_k}, j, \qquad (9)$$

where m_k indicates the value of X^k and the weights $W^{a,k}_{m_k}$ are the columns standards coordinates on the a^{th} factorial axis corresponding to $W^k_{m_k}, j$. Typically, a score is defined on the first factorial

axis, i.e., $a = 1$. A composite social vulnerability score F_j is obtained by applying a Dirac delta function, $\delta(k - a)$ (Ssempiira et al., 2018):

$$F_j = \frac{1}{K}\sum_{k=1}^{K}\sum_{m_k \epsilon \{0,1\}}^{1}\sum_{a=1}^{L}\delta(k-a)W_{m_k}^{a,k}X_{m_k}^{k}, j, j \qquad (10)$$

where L is the number of factorial axes used in the composite score. The Dirac delta function takes the value 1 when the weights related to $X_{m_k}^{a,k}, j$ are selected from the a^{th} factorial axis and 0 otherwise, that is, $\delta(k - a) = 1$ if $k - a$ and $\delta(k - a) = 0$ if $k \neq a$.

The interpretation of the score is eased by translating the weights so that the absence category, $(m_k = 0)$ of the X^K indicator received a zero weight and the presence one $(m_k = 1)$ received a strictly positive weight representing the gain in the readiness increase measured by the axis when a household i acquires the k^{th} item ((Ssempiira et al., (2018) as cited in Kazembe (2021)).

2.2.2.4. Cluster Analysis

Cluster analysis groups individuals with similar dietary patterns into mutually exclusive categories according to the mean of the food intake variables (Thorpe et al., 2016). The goal of cluster analysis is to group related observations in a dataset so that they are as similar to one another as possible and as dissimilar to one another as possible. Cluster analysis groups observations by similarities across rows as opposed to other data reduction approaches like factor analysis (FA) and principal components analysis (PCA), which group by similarities across variables (columns) of a dataset (Population Health Methods, 2018).

When it's crucial to distinctly distinguish various groups of sample participants within the population of interest, researchers may choose to employ cluster analysis. This can be the case if they want to use these groupings for further investigation. Therefore, cluster analysis techniques

are especially beneficial during the research's exploratory phase (Gleason, Boushey, Harris, & Zoellner, 2015). Once satisfactory clusters are formed, the clusters can be prepared for subsequent inferential statistics if a suitable a priori hypothesis has been established, such as by comparing mean outcomes among the newly defined groups of sample members (Gleason et al., 2015).

Several CA algorithms exist, with k-means being the most popular in nutrition research because it can handle a large number of input variables efficiently (Thorpe et al., 2016). K-means is one method of cluster analysis that groups observations by minimizing Euclidean distances between them. Euclidean distances are analogous to measuring the hypotenuse of a triangle, where the differences between two observations on two variables (x and y) are plugged into the Pythagorean equation to solve for the shortest distance between the two points (length of the hypotenuse) (Population Health Methods, 2018).

All variables must be continuous in order to use K-means clustering. Other approaches that do not need all variables to be continuous, including some hierarchical clustering algorithms, have different assumption. The number of clusters, k, must also be specified a priori when using K-means clustering. Although the data can be used empirically to do this, the choice should be made based on theory because poor decisions can result in incorrect clusters. (Population Health Methods, 2018).

In CA, distances are computed using simple Euclidean distance. If you want to use another distance or similarity measure, use the Hierarchical Cluster Analysis procedure. Scaling of variables is an important consideration. If your variables are measured on different scales (for example, one variable is expressed in dollars and another variable is expressed in years), your results may be misleading. In such cases, you should consider standardizing your variables before you perform

the *k*-means cluster analysis. The procedure assumes that you have selected the appropriate number of clusters and that you have included all relevant variables. If you have chosen an inappropriate number of clusters or omitted important variables, your results may be misleading (IBM knowledge center, 2018).

Data-driven approaches often identify eating patterns that do not reflect guidelines or knowledge about ideal eating patterns and therefore, lack of associations with health and disease may not be surprising.

2.2.2.5. The Least Absolute Shrinkage and Selection Operator (LASSO)

In order to enhance the prediction of clinically meaningful risk factors and the identification of predictive dietary patterns, Zhang et al., (2018) proposed a new dietary pattern analysis method using the advanced LASSO (Least Absolute Shrinkage and Selection Operator) model. LASSO is a regression-based method that penalizes the absolute value of the regression coefficients; in doing so, it regularizes the impact a coefficient may have in the overall regression (Zhang et al., 2018). The greater the penalization, the greater the shrinkage of a coefficient, with some coefficients shrinking to zero. Thus, the LASSO model is a type of automatic feature selection, a method that has found success in disciplines including computational chemistry, genomics, and neuroimaging. Therefore, it is thought to be very novel and inventive to evaluate the LASSO model systematically when assessing dietary data connected to health outcomes (Zhang et al., 2018).

2.2.3. Hybrid Methods

Further to the theoretical and empirical methods are hybrid methods, such as reduced rank regression and partial least squares regression that use a combination of theoretical knowledge and statistical approaches to determine dietary patterns (Thorpe et al, 2016).

Reduced rank regression (RRR) and partial least-squares regression (PLS) are proposed substitutes to PCA for deriving dietary patterns (Hoffmann et al., 2004). RRR and PCA both use dimension reduction techniques that produce uncorrelated summary variables that represent a bigger collection of original factors. RRR's objective is distinct from PCA's since it derives combinations of predictor factors that largely account for the data in a set of response variables. Key nutrients or biomarkers function as the response variables, and linear combinations of foods are derived which maximize the explained variance in these responses. The PLS method is a compromise between PCA and RRR; the goal of this approach is to explain variability in select nutrients or biomarkers, as well as foods (Colón-Ramos, Kabagambe , & Baylin, 2007).

2.2.3.1. Reduced Rank Regression (RRR)

RRR, otherwise known as the maximum redundancy analysis, is also one of the closely related data reduction methods. Reduced rank regression is based on statistics that are very similar to those used in PCA and FA. Each of these techniques uses a similar procedure to calculate the principal components, or factors, utilizing eigenvalues and factor loadings based on eigenvectors (Gleason et al., 2015). While the two methods discussed above derive the factors (principal components) by maximizing the explained variation of a set of predictor or input variables (such as food groups), reduced rank regression derives the factors by taking into account as much variation in a researcher-determined response variable or variables as possible (e.g., body mass index, nutrients, biomarkers) (Gleason et al., 2015). Reduced rank regression is particularly helpful when a

researcher wishes to quickly and effectively summarize a big collection of input variables in order to explain or predict a certain final outcome of interest. In that situation, based on prior study, the researcher could choose a collection of intermediate outcomes assumed to be related to the primary outcome of interest as response variables (Gleason et al., 2015).

A linear model is frequently suggested to relate response to composition in the study of the experimental features of mixes. The statistical method of linear regression analysis is then appropriate, and it is frequently used multiple times when there are a number of responses of interest. Now since the responses are frequently interconnected, it could be conceivable to utilize an empirical linear relationship to predict the approximate value of a specific response based on information about the others. In these situations, it is necessary to modify the process for calculating the regression coefficients of response on composition to account for the known existence of such correlations (whose linearity is implied by the mutual linear dependence of responses on composition). This leads to consideration of the multivariate regression model with a constraint imposed on the rank of the matrix of coefficients, sometimes termed reduced-rank regression. Such models have been studied, e.g. by Izenman (1975) and also by Burket (1964), who used a factor analysis model (Davies et al., 1982).

The RRR model is a multivariate regression model with a coefficient matrix with reduced rank. It is related to canonical correlations and involves calculating eigenvalues and eigenvectors (Johansen, 2008). The solution for the RRR analysis is related to the singular value decomposition of the full rank matrix. In RRR analysis, principal component analysis is first performed on Y followed by regressing X on the principal components. It is based on maximizing the covariance between the principal components and response variables.

Partial least squares depend on selecting components $t = Xw$ of the explanatory variables and $\mu = Yq$ of the responses that have maximum covariance, whereas principal component regression effectively ignores μ and selects t to have maximum variance, subject to orthogonality constraints. In contrast, reduced rank regression selects μ to account for as much variation in the predicted responses as possible, effectively ignoring the explanatory variables for the purposes of factor extraction. In reduced rank regression, the Y weights q_i are the eigenvectors of the covariance matrix $\grave{Y}_{LS}Y_{LS}$ of the responses predicted by ordinary least squares regression; the X-scores are the projections of the Y-scores Y_{qi} onto the X space (SAS Institute Inc, 2004).

RRR takes the first principal components of the ordinary regression matrix (Kiers & Smilde, 2007).

Coefficient matrix can be written as a product of two component matrices of lower dimension. It follows that the assumption of lower rank for the regression coefficient matrix leads to estimation results which take into account the interrelations among the multiple responses and use the entire set of explanatory variables in a systematic fashion (Reinsel, 2006). The model for reduced-rank regression may be written as:

$$Y_{N\times K} = D_{M\times K} + E^{*}_{N\times K}$$
$$rank\ (D) < s, \qquad (11)$$

where s is an integer to be specified. The interpretation of (11) is as follows $D_{M\times K}$ and $Y_{N\times K}$ are data matrices whose N rows contain measurements on M and K variables respectively for N individuals or experimental units. Assume that column means have been subtracted from each variable of X and Y. This corresponds to the situation that a constant term associated with each regression has been previously estimated (by maximum likelihood in the case of normality) and

allows us to consider the homogeneous model (1) without loss of generality. The problem is to estimate the unknown matrix of regression coefficients $D_{\mathrm{M\times K}}$ subject to the rank constraint $rank\ (D) < s < M$ which imposes the condition that the predictions shall be linearly dependent. Finally, E^* is the matrix of stochastic errors which are assumed to be uncorrelated row-wise, that is from unit to unit, but which may be correlated between variables measured on the same unit.

We shall assume a zero-mean K-variate multi-normal distribution for the rows of $E^*, e\sim N(0,\sum)$, $\sum$, is an unknown positive definite covariance matrix. It is natural to make explicit the reduced-rank nature of $D_{M\times K}$ by expression this matrix as the product of two matrices, $D_{M\times K} = Q_{M\times S}B_{s\times K}$, where $Q_{M\times S}$ is a matrix whose s columns are a set of linearly independent vectors representing a basis for the unknown subspace spanned by the columns of D, and $B_{s\times K}$ has K columns that define the appropriate linear combinations to represent the columns of D; i.e. the regression coefficients for each y-variable with respect to this basis. We shall choose Q to have the normalization $P_{N\times s} = X_{N\times M}Q_{M\times S}$; that is, the s columns of P are orthogonal linear combinations of the x-variables. This definition is consistent with canonical variate analysis and in fact it shall show that the columns of Q may be estimated as the s principal canonical linear combinations (Tso, 1981).

2.2.3.2. Partial Least-Squares Regression (PLSR)

PLSR enables working with small number of observation units and/or data set with multicollinearity and/or more than one response variable. PLSR involves information on both X and Y in the calculation of components and loadings by using singular value decomposition of $S = X\text{`}Y$ cross product matrix. PLSR is thus instrumental in in finding relations between 2 matrices (X and Y) (Sharifi, 2016).

According to Sharifi (2016), a PLSR model will try to find the multidimensional direction in the X space that explains the maximum multidimensional variance direction in the Y space. PLSR regression is particularly suited when the matrix of predictors has more variables than observations, and when there is multicollinearity among X values. PLSR is derived as following: Given a pair of input and output data matrices X and Y and assuming they are linearly related by

$$Y = XC + V \qquad (12)$$

where V and C are noise and coefficient matrices, respectively, the PLSR regression builds a linear model by decomposing matrices V and Y into bilinear terms, the general underlying model of multivariate PLSR is thus defined by:

$$X = TP^T + E \qquad (13)$$

$$Y = UQ^T + F \qquad (14)$$

whereby X is an $n \times m$ matrix of predictors, Y is an $n \times p$ matrix of responses; T and U are $n \times 1$ matrices that are, respectively, projections of X (the X score, component or factor matrix) and projections of Y (the Y scores); P and Q are, respectively, $m \times 1$ and $p \times 1$ orthogonal loading matrices; and matrices E and F are the error terms, assumed to be independent and identically distributed random normal variables. The decompositions of X and Y are made so as to maximise the covariance between T and U (Sharifi, 2016).

2.3. Application of Correlational Analysis in Dietary Patterns

The immediate determinant of nutritional status is dietary intake (calories, protein, fat, micronutrients, carbohydrates, and vitamins). Dietary intake must be sufficient in quantity and quality and nutrients must be consumed in appropriate combinations for the child to absorb them

(Suresh, Shailendra, & Prabuddha., 2014). The inadequate food intake of a child is the result of households not having enough resources (such as own food production, income, or in-kind transfers of food) for gaining access to food. Although sustained income growth can improve the nutritional status of a child through the household's access to various resources, other factors such as women's education and nutritional knowledge play an equally important role.

Namibia is a developing country undergoing rapid nutrition transition. The nutrition transition is underpinned by dietary changes in both rural and urban areas. The result of this poses a massive threat to public health with a higher impact on the poor. Furthermore, Namibia produces about 40 percent of the food it consumes and is highly dependent on imports (World Food Programme, 2019). This means that while food is available, price fluctuations can make it difficult to access food for 28 percent of Namibian families. This particularly affects the 80 percent of the population who depend on markets to fulfill their food needs. Smallholder farmers also have limited access to nutritious food due to recurrent droughts and floods, low productivity and limited access to land (World Food Programme, 2019). These limitations translate into poorly diversified diets with insufficient consumption of vitamins and minerals, which are at the root of persistent malnutrition.

While the essential role of nutrition in the promotion and enhancement of the overall quality and span of life is widely and well recognized, many Namibians suffer from nutritional disorders that are due to an inadequate food intake, both in terms of quality and quantity. These nutrition problems are related to diets, which are monotonous, deficient in food energy and contain few foods that are rich in vitamins and minerals (Ministry of Health and Social Services, 2013).

The World Bank's Global Monitoring Report (2012) on Food Prices and Nutrition provides an analysis of the current global situation in food insecurity and nutrition. According to Suresh,

Shailendra, & Prabuddha (2014), some of the consequences of higher food prices on undernutrition and the difficulties associated with meeting the Millennium Development Goals (MDGs) (now SDG's) noted in the World Bank's Global Monitoring Report (2012) are that:

- As food prices increase, the purchasing power of the poor decreases, the composition of their diet worsens, and their food consumption may decrease. These changes directly affect all targets of MDG 1 (SDG 1) on poverty, full and productive employment, and hunger.
- Malnutrition affects early childhood development and makes children more likely to drop out of school (MDG 2 (SDG 2)).
- An increase in food prices affects women and girls' consumption disproportionately (MDG 3 (SDG 5)).
- Undernutrition is linked directly to more than one-third of children's deaths each year (MDG 4 (SDG 3)).
- Pregnant women face heightened maternal mortality, through increased anemia, during a food price crisis (MDG 5 (SDG 3)).
- The adverse effects of a food crisis on the availability of health services and on health status bear on countries and individuals' abilities to combat the HIV/AIDS epidemic (MDG 6 (SDG 3))
- Undernutrition weakens the immune system and compounds the effect of diarrhea and waterborne diseases (MDG 7 (SDG 3)).
- Higher food prices have weakened intergovernmental coordination in food markets (MDG 8 (SDG 1)).

According to Suresh et al., correlation analysis can be useful for the following reasons:

1. Formulating nutrition targets (such as targets within a development plan).

2. Planning social development programs/projects for the vulnerable sections of the population (for example, government and non-government organizations use nutrition indicators in implementing, monitoring, and evaluating social developmental programs).

3. Using it as baseline and benchmark data (nutrition indicators often reflect the current nutrition situation with which future data can be compared at the start of an intervention project).

2.3.1. Understanding Correlation Analysis

A few concepts are associated with correlation analysis.

Suppose we have two random variables X and Y with means $\bar{X}$ and $\bar{Y}$ and standard deviations S_X and S_Y respectively. Then, the correlation coefficient can be computed as follows:

$$r = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\,(Y_i - \overline{Y})}{(n-1)S_X S_Y} \qquad (15)$$

Equation 15 can be interpreted as follows: suppose that an X value was above average and that the associated Y value was also above average. Then the product $(X_i - \bar{X})\,(Y_i - \bar{Y})$ would be positive as it's a product of two positive numbers. If the X value and the Y value were both below average, then the product would also be positive. Thus, a positive correlation indicates that large values of X are associated with large values of Y, while small values of X are associated with small values of Y. The correlation coefficient measures the strength of a linear relationship between any two variables and is always between −1 and +1. The closer the correlation is to + 1 or −1, the closer it is to a perfect relationship (Suresh et al., 2014).

The interpretation of r^2 (the square of r) can also be made in terms of variation in the dependent variable Y that is explained by the regression line. Suppose that the total deviation of an actual Y

from its mean Y can be expressed as the sum of two non-overlapping components, ($\hat{Y} - \bar{Y}$) and (Y $-\hat{Y}$). The first component represents that part of the total difference explained or accounted by the relationship of Y with X; the other component represents that part of the total difference remaining after accounting for the relationship of Y with X. Thus, we get:

$$\sum(Y - \bar{Y})^2 = \sum(\hat{Y} - \bar{Y})^2 + \sum(Y - \bar{Y})^2 \qquad (16)$$

Equation 16 can be interpreted as the sum of total variation to be equal to the sum of explained and unexplained variation. The ratio $\sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2$ is the proportion of total variation that remains unexplained by the regression equation. On the other hand, $1 - \sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2$ represents the proportion of the total variation in Y that can be explained by the regression equation. The above ideas can be summarized as follows:

$$r^2 = 1 - \frac{\sum(Y-\hat{Y})^2}{\sum(Y-\bar{Y})^2} = 1 - \frac{Unexplained\ variation}{Total\ Variation} = \frac{Explained\ Variatiom}{Total\ Variation} \qquad (17)$$

2.3.2. Path Analysis and Structural Equation Models

Structural Equation Modeling abbreviated as SEM, is a very general statistical modelling technique, which is widely used in the behavioral sciences (Hox, Moerbeek, & Van De Schoot, 2017). It can be viewed as a combination of factor analysis and regression or path analysis. The interest in SEM is often on theoretical constructs, which are represented by the latent factors. The relationship between the theoretical constructs are represented by regression or path coefficients between the factors. The structural equation model implies a structure for the covariances between the observed variables, which provides the alternative name covariance structure modelling. It

should be noted that the model can be extended to include means of observed variables or factors in the model, which makes covariance structure modeling a less accurate name (Hox et al., 2017).

Bardenheier et al., (2013) used structural equation modeling with factor analysis, which groups inter-correlated variables into a single factor or latent construct, and path analysis, which includes the direct and indirect effects of factors previously reported associated with prediabetes. Direct effects are depicted as an arrow emanating from an independent variable (exposure) leading and pointing to a dependent variable (outcome). An indirect effect is depicted as a mediating variable having an arrow pointing to it from an independent variable but also pointing to yet another dependent variable. A confounder is depicted as a variable with direct effects on both the exposure and the dependent variable. Correlations between the measurement errors of two variables are represented by two-headed curving arrows, in which case only the measurement error terms are correlated.

In general, latent variable models reduce measurement error by having multiple indicators per latent construct, the ability to test models with multiple dependent variables, and the benefit of testing multiple integrated models simultaneously rather than factors individually. In addition, structural-equation modeling examines the direct and indirect effects of mediators on dependent variables while allowing the examination of complex associations among multiple mediators (Bardenheier et al., 2013). Conversely, in a traditional regression model, mediators would not be included because they would block the pathway between the independent variable of interest and the dependent variable. Thus, in the structural-equation model, the independent factors and combined mediated relationships can be examined simultaneously, determining the impact of each of the dependent variables in the appropriate order. Thus, the SEM includes mediating effects

without sacrificing indirect effects of interest. For each relationship in the SEM model, only data missing for either the independent or dependent variable would be missing from that equation (Bardenheier et al., 2013).

Furthermore, Castro et al., (2016) used the exploratory structural equation modelling (ESEM) using the robust maximum likelihood (MLR) parameter estimation and the oblique Geomin rotation applied to food group variables in order to empirically derive dietary patterns. ESEM is a multivariate statistical technique that can be interpreted as a combination of EFA and SEM (Castro et al., 2016). It relies on the covariance structure of the observed variables and is indicated when the researcher has a weak hypothesis about how multiple-observed variables load on the factors, a common situation in dietary pattern analyses. In addition, Castro et al., (2016) used ESEM since it allows for investigators to test the significance of factor loadings, which contributes to reduce the subjectivity during modelling and interpreting dietary patterns.

2.4. Measurement Errors in Dietary Assessment

Measurement error is a major issue in dietary surveys. Research has indicated consistent errors in self-reported dietary intake, using the available dietary assessment methods (Devlin, Mcnulty, Nugent, & Gibney, 2012). Dietary intake is commonly over- or under-reported leading to implausible energy intake in population group, where the latter may be considered the most disadvantageous to research studies. Under-reporting of dietary intake can happen in three ways, where subjects can (1) deny ever eating the food at all; (2) fail to report the correct portion size consumed or (3) fail to report how many times the food is actually consumed (Devlin et al., 2012).

Measurement error can be in the form of random error and/or systematic error. The former reduces precision and the latter results in incorrect estimates. Hence, it is important to quantify and correct

for these effects. In general, the errors that affect the validity of a dietary assessment method are systematic and those associated with reproducibility are random (Gibson, Charrondiere, & Bell, 2017).

Random errors may occur across all respondents and all days, causing associations to be underestimated and even failure to detect associations in the first place. This type of error can be minimized by increasing the number of measurements. *Systematic errors* may be respondent-, food- or interviewer-specific and can result in underestimated or overestimated associations; these type or errors cannot be minimized by increasing the number of measurements.

Major measurement errors are due to nonresponse bias, respondent biases, interviewer biases, respondent memory lapses, incorrect estimation of portion size, supplement usage, coding errors, mistakes in the handling of mixed dishes, etc. (Gibson et al., 2017). Table 8 describes possible sources of error that should be considered in different dietary assessment methods. Depending on the population group being studied, it is important to employ appropriate strategies to optimize the information being retrieved and reported to the investigator, and to minimize errors.

Considerable efforts have been made in developing statistical techniques to deal with these errors and to enhance the performance of various methods. Linear regression calibration, energy adjustment and analysis of variance can be used to correct for random and systematic errors during the data analyses stage (Bennett, et al., 2017).

Table 8*:* Measurement errors in dietary diversity

Source of error	Estimated food records*	Weighed food records	24-hour recall *∞	Dietary history	FFQ	Brief dietary instruments	Duplicate meal method
Food composition table	+	+	+	+	+	-	-

Food coding	+	+	+	+	+	+	+
Incorrect weighing of food	-	+	-	-	-	-	+
Reporting error	+	+	+	+	+	+	-
Diet variations with time and season	+	+	+	+	+	+	+
Wrong frequency	-	-	-	+	+	+	-
Modified eating pattern	±	±	+	+	+	+	+
Respondent memory lapses	-	-	+	+	+	+	-
Portion size estimation	-	-	+	+	+**	-	-
Respondent bias	±	±***	+	+	+	+	-
Interview bias	±	±	±	+	±	-	-

Note. Adapted from FAO (2002).

*Image-assisted dietary assessment methods inherit the sources of errors of conventional dietary assessment methods
+ Random errors
- Systematic Errors
∞ Dietary diversity score inherits the sources of errors of the 24-hour recall, when data was collected by a 24-hR recall questionnaire
± means that there is a possibility for the method to be affected by the source of error
**Only in cases of quantitative FFQ
***Occurs in cases where a field worker or nutritionist is weighing the food (happens in low resource countries).

2.5. Dietary Diversity

Approaches to measuring food richness of nutrients in diets can be done by considering the diversity of food groups consumed. Dietary diversity score is a measure of quality of food consumed. Measures of dietary diversity tend to be of two types: those based on whether an individual food is consumed or not and those that are based on whether any food from a particular food group is consumed. When comparing dietary diversity based on food groups and individual foods, regression analysis shows that dietary diversity based on food groups is a stronger determinant of nutritional adequacy (Ruel, 2003).

The Household Dietary Diversity Score (HDDS) adopts the food group approach and asks how many of 12 different groups were consumed in the household over a specific recall period (usually 24 hours) (Hoddinott and Yohannes, 2002; Swindale and Ohri-Vachaspati, 2005; Swindale and Bilinsky, 2006b). The classification recommended for Africa by the by FAO is shown in Table 9, which groups food types primarily based on their nutrient content. The number of food groups consumed in a household provides a measure of the quality of the diet by reflecting the dietary diversity.

The respondent is asked a yes/no 24-hour recall question for each of the twelve food groups with regard to food consumed by household members in the home or prepared in the home for consumption by household members outside the home e.g. at lunchtime when at work. The sum of the 'yes' responses provide a score out of 12 for each household. The individual household scores are then averaged to construct a HDDS for a given population. The closer the score is to twelve the greater the dietary diversity.

Table 9: Food types in groups to construct HDDS

No	Food Groups
1.	Any (local food) bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice, wheat, or (any other local grain)
2.	Potatoes, yams, cassava, or any foods made from roots and tubers
3.	Vegetables
4.	Fruits
5.	Beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or other organ meats
6.	Eggs
7.	Fresh or dried fish or shellfish
8.	Food made from beans, peas, lentils or nuts
9.	Cheese, yoghurt, milk or other milk products
10.	Foods made with oil, fat or butter
11.	Sugar/honey
12.	Any foods such as condiments/tea/coffee

2.5.1. Modelling Approach for Household Dietary Diversity Score (HDDS)

A Poisson model is used in this study to investigate the influence of different explanatory variables on household dietary diversity. The value of the dependent variable HDDS varies from 0 to 12, which is a count variable. The value of HDDS is assumed to have a Poisson distribution with expectation μ, for independent variables X_i, the Poisson regression model for expected counts can be specified as an exponential function (Zhong , et al., 2018). For the dependent variable HDDS, the Poisson regression model is as follows:

$$\mu_i = E(HDDS_i|X_i) = exp(\beta_0 + \beta_i X_i) \quad (18)$$

where HDDS is the HDDS of household i, X_i refers to the vector of independent variables, and β_o and β_i are the constant and the coefficient vector for independent variables, respectively. The alternative log–linear model can be written as:

$$ln(\mu_i) = exp(\beta_0 + \beta_i X_i) \quad (19)$$

2.6. Data Overview

This study used a pool of data from two surveys: the Namibia Household and Income Expenditure Survey (NHIES) (2015/16) and the AFSUN- AFSUN-HCP Household Food Security Baseline Survey (2016). The sample design for the NHIES 2015/16 survey was a stratified two-stage cluster sample, where the first stage units were geographical areas designated as the Primary Sampling Units (PSUs) and the second stage units were the households. The up-to-date list of households in the selected PSU were prepared during the listing stage of fieldwork, and 12 households were systematically selected in each PSUs (NSA, 2018). The AFSUN- AFSUN-HCP is a validated questionnaire which collects a wide range of demographic, economic and food consumption and sourcing data at the household level. Households surveyed in the ten constituencies of Windhoek

were identified using a two-stage sampling design. Primary sampling units (PSUs) were first randomly selected from a master frame developed and demarcated for the 2011 Population and Housing Census. Within the 10 constituencies, a total of 35 PSUs were selected covering the whole of Windhoek, and 25 households were systematically selected in each PSU. The sampled PSUs and households were located on maps, which were used to select households for in-person interviews. Household heads (or their spouses/ partners) were recruited to complete the survey.

Food security was assessed using three standardized, cross-cultural measures of household food access: the Household Food Insecurity Access Prevalence (HFIAP) indicator, the Household Dietary Diversity Score (HDDS) and Months of Inadequate Household Food Provisioning (MIHFP). The HFIAP is a self-reporting measure of the frequency of occurrence of nine separate food-related conditions in the household in the 30 days prior to the survey. The HFIAP allocates each household to one of four categories (food secure, mildly food secure, moderately food insecure and severely food insecure). For purposes of this dissertation, the four categories were binned into two: food secure and food insecure (combining the mild, moderate and severe food insecure categories). The HDDS is a dietary intake metric which captures information on the types of foods consumed by the household in the 24 hours prior to the survey. The value of the HDDS variable ranges from 0 to 12 (based on the allocation of all food items into 12 separate food groups). An individual or household was considered to have consumed a particular food group if they had consumed at least one food item from that group in the previous 24 hours. For this dissertation, the HDDS was binned into lower HDDS (0–5) and higher HDDS (6-12). Months of Inadequate Household Food Provisioning (MIHFP) assess households on the months that they did not have enough food. The value of MIHFP ranges from 0-12, based on the number of months.

For this dissertation, the MIHFP was binned into seasonal (0-5 months) and persistence (6-12 months).

In order to assess the consumption of unhealthy diets, the data was aggregated into convenience and non-convenience foods. Convenience and non-convenience foods have been categorized based on the source purchased and further measured on the number of times an individual in the household made use of a source, weekly or monthly basis and estimates were made separately for each. Convenience food sources include fast foods/take-away, restaurants, spaza/tuckshop, Street seller/trader/hawker and begging from the streets while non-convenience food sources comprise supermarkets, small-shops, Open markets, and food grown by households in rural areas.

Data on the socio-economic status of all households were extracted for the analysis and included the following variables: Household size, Residence, Sex of Head, Age of Head, Main source of income, Highest level of educational attainment of household head, work status, marital status, Access to Radios, Access to Televisions and Access to Internet. Furthermore, Data on a range of non-communicable diseases (including diabetes, heart problems (CVD), and hypertension/stroke) were extracted for analysis.

Various analytical approaches have been used to determine the relationship between dietary patterns and non-communicable diseases as well as to derive dietary patterns from food consumption data. The study used Poisson models for count data at different levels. Specifically, the study applied bivariate count models on convenience and non-convenience food preferences/consumptions, copula joint modelling of food insecurity, HDDS and MIHFP while observing partial observability and sample selection. The dissertation further applied multiple-indicator, multiple cause modelling to examine the relationship between foods consumed and non-

communicable diseases through factor analysis, principal component analysis and structural equation modelling.

Several designs are necessary in interpreting the data from the study. First, the cross-sectional nature of the survey limits the generalizability of our results. For consistent capture of food intakes, it has been proposed that longitudinal types of study design are preferable to capture the stability of food intake. That said, the four-week and 24-h recall questions used in this study have been extensively used and shown consistent results to those reported in cohort or longitudinal studies. Secondly, the Poisson regression model provides a basis for the analysis of count data. The Poisson regression model is often of limited use because empirical count data sets typically exhibit over-dispersion and/ or an excess number of zeros. In order to analyze bivariate count data, the plain Poisson regression model was extended. Thirdly, PCA approach has been criticized as exploratory in nature and a posteriori in its approach. This study extended the analysis to factor analysis as a secondary step to enable identification of dietary patterns. Fourth, this study was based on disease outcomes reported at the individual level, while other factors such as food security, dietary diversity, lived poverty and housing type were measured at household level. Our analysis did not adjust for such hierarchical structure in the data, but the consistency of our findings suggests that the approach used was robust.

2.6.1. Research Ethics

Ethical clearance was obtained from the University of Namibia, Research and Ethics Committee (UREC) and permission to conduct the research from University of Namibia Centre for Postgraduate Studies (UCPS). The respondents were explicitly asked for their verbal informed consent to voluntarily participate in the study and recorded in the questionnaire. Prior to starting

each interview, the study objectives were explained to the respondents. It was also clarified to them that the data collected would be kept strictly confidential, analyzed anonymously and used for research purposes only.

2.7. Conclusion

In summary, this chapter reviewed methods of dietary pattern analysis. Three types of dietary pattern methods were reviewed, namely the a priori, a posterior (empirical methods) and hybrid methods. The chapter further reviewed correlation analysis in dietary pattern, using path analysis and structural equation models. To strengthen understanding of dietary patterns, the topic of dietary diversity was reviewed to understand different approaches to measuring dietary diversity and other explanatory factors. The methods reviewed forms part of analysis and were explored to explain behaviors in dietary patterns, dietary diversity and non-communicable diseases in subsequent chapters.

CHAPTER 3: COUNT MODELS APPLICATION ON DIETARY DIVERSITY IN NAMIBIA

Laina Mbongo[1]*, Lawrence Kazembe, Lillian Pazvakawambwa
Department of Computing, Mathematics and Statistical Sciences,
University of Namibia, Windhoek, Namibia

Abstract

Consuming a wide variety of foods ensures an adequate intake of essential nutrients and leads to better diet quality and optimal health outcomes. High-quality diets further help unlock the development potential of individuals, boost economic productivity and reduce demands on expenditure in areas such as health and social protection. The objective of this study is to examine factors associated with household dietary diversity scores (HDDS). Count data arise frequently in analyzing HDDS, but regularly violate the equi-dispersion constraint imposed by the Poisson distribution, the most popular distribution for analyzing these data. Here, a range of count regression models were employed to permit for possible over-dispersion and zero-inflation in the response variable, computed as the number of types of food consumed over the 24-hour recall period using household survey data from the NHIES (2015/16) dataset. Model selection was based on AIC and BIC with model having small value of AIC considered the best. Poisson Inverse Gaussian regression model had the best AIC and BIC and thus deemed best fit in this study. Findings suggest a moderate diverse diet with HDDS mean of 6.5, with less consumption of food made from beans/lentils; eggs; fruits/vegetables, and more consumption in starch food. Some of the risk factors for HDDS included educational level, sex of head of household and main source of income. Poisson Inverse Gaussian regression was found to fit the data best due to its low AIC.

Keywords: dietary diversity, count models, food security, Namibia

[1]* Corresponding Author Email: inambongo@gmail.com

3.1. Introduction

The issues of chronic food insecurity, poverty, and malnutrition continue to be fundamental human welfare challenges in developing and developed countries (Suresh et al., 2014). Subsequently, ending all forms of malnutrition and providing access to safe, sufficient and nutritious food for all people year-round by 2030 is one of the targets of the United Nation's Sustainable Development Goals (SDGs), specifically SDG2 (aiming to end hunger, achieve food security, improve nutrition, and promote sustainable agriculture). Food security and nutrition is intangible for the majority of the rural population and a significant proportion of those living in urban areas in Namibia. In 2013, the Emergency Food Security Assessment (EFSA) found 16% of the Namibian population to be severely food insecure while 22% of the population was moderately food insecure; further, the 2015 Global Hunger Index ranked Namibia 87th out of 120 countries assessed, with an index score of 31.8 indicating a serious food problem (National Planning Commission of Namibia, 2016).

Food insecurity means that people do not have "physical, social and economic access to food of sufficient quantity and quality in terms of variety, diversity, nutrient content and safety to meet their dietary needs and food preferences for an active and healthy life, coupled with a sanitary environment, adequate health, education and care" at all times (FAO, 2010). Food security is not only concerned with ensuring adequate supply of dietary energy but includes a diet which is sufficient to meet all nutritional needs, thus incorporating sufficient intake of vitamins and minerals (Kennedy., 2009). Consuming a wide variety of foods among and within groups is a recommended strategy to help ensure adequate intake of micronutrients (FAO/WHO, 2002). When the first indicators of food security were developed, the focus was mainly on national food supply, emphasizing the concept that if adequate quantities of food were available, food security needs would be met (Kennedy., 2009). However, the supply of large quantities of foods only implies that

food security needs will be met but the truth is individuals are instead just consuming monotonous diets. Over the years, the definition of food security evolved and considered dietary diversity as an outcome measure of food security.

There are several food security and dietary diversity measurement approaches (Vellema, Desiere, & D'Haese, 2015). For example, anthropometric measures are used to monitor the growth of children under 5 years, recalls of food consumed in the past 24 hours or over a longer reference period are recorded to measure the intake of macronutrients and micronutrients and data on food expenditure are used to define food poverty lines, whereas experience –based responses such as the Household food insecurity access scale (HFIAS) elicit perceived consequences of not having enough food. These indices add quantitative elements to qualitative aspects.

Food security is analyzed using different measures to capture its spectrum of access, availability, stability and utilization. The most commonly measure used to measure food accessibility and utilization, given its relative simplicity is the Household dietary diversity score (HDDS). It was developed as a quick-to-use survey-based indicator to measure the impact on household food access of programs with improvements in food security as their core objective (Swindale & Bilinsky, 2006). The Household Dietary Diversity Score (HDDS) is meant to reflect, in a snapshot form, the economic ability of a household to access a variety of foods whereas the Individual dietary diversity scores (IDDS) aim to reflect food utilization and nutrient adequacy (FAO, 2010).

Although all these indices and measurements aim to show food security levels and overall quality of diet, they often focus on specific food features, depending on the contexts and objectives of their usage. Counts of foods or food groups have been the most frequently used method in dietary diversity analysis. The Poisson regression model provides a basis for the analysis of count data.

The classical Poisson regression model for count data is often of limited use because empirical count data sets typically exhibit over-dispersion and/ or an excess number of zeros (Achim, Christian , & Simon, 2008). The issue of excess number of zeros can be addressed by extending the plain Poisson regression model in various directions: e.g., using sandwich covariance's or estimating an additional dispersion parameter (Quasi-Poisson Model) (Achim, Christian , & Simon, 2008). Negative binomial regression, having a second parameter, is now a standard way of modeling over dispersed Poisson data. A number of other models based on the Poisson and negative binomial models have been designed to appropriately compensate for over dispersion or under dispersion (Hilbe, 2014).

To the best of our knowledge, the present study is the first comparison study extending the application of various count models from the basic Poisson regression model and hence this dissertation contributes to the literature by aiming on the issue of dietary diversity and the use of different models to analyze count data. Knowledge and application of different count models in dietary diversity measures may solve the problem of dispersion and heterogeneity and identify relationships among dietary diversity, healthful dietary patterns, and health outcomes.

3.2. Materials and Methods

3.2.1. Study Area and Sampling Design

The study used cross-sectional survey data from the Namibian Household and Income Expenditure (NHIES) of 2015/2016. The sample design for the survey was a stratified two-stage cluster sample, where the first stage units were geographical areas designated as the Primary Sampling Units (PSUs) and the second stage units were the households. The up-to-date list of households in the

selected PSU were prepared during the listing stage of fieldwork, and 12 households were systematically selected in each PSUs (NSA, 2018).

The food groups in the NHIES 2015/2016 were re-grouped and re-arranged in order to make up the 12 food groups. The dietary diversity score was then created on the new groups. Statistical R software Version 3.6 was used to fit the count models.

3.2.2. Household Dietary Diversity Score (HDDS)

The Dietary Diversity Score adopts the food group approach and asks how many of 12 different groups were consumed in the household over a specific recall period (usually 24 hours) (Hoddinott and 40 Yohannes, 2002; Swindale and Ohri-Vachaspati, 2005; Swindale and Bilinsky, 2006b). A cross tabulation was done between the dietary diversity and the demographic variables. The classification recommended for Africa by the FAO is shown in Table 10, which food group types primarily based on their nutrient content. The number of food groups consumed in a household provides a measure of the quality of the diet by reflecting the dietary diversity.

Table 10: Food Types in groups to construct HDDS

No	Food groups
1.	Any local food including: bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice
2.	Potatoes, yams, cassava, or any foods made from roots and tubers
3.	Vegetables
4.	Fruits
5.	Beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or organ meats
6.	Eggs
7.	Fresh or dried fish or shellfish
8.	Food made from beans, peas, lentils or nuts
9.	Cheese, yoghurt, milk or other milk products
10.	Foods made with oil, fat or butter
11.	Sugar/honey
12.	Any foods such as condiments/tea/coffee

3.2.3. Overview of the Models

3.2.3.1. Foundation for Count Data Models

The underpinning factor for the development of count data models is the Poisson distribution. Most of the count data models belong to Generalized Linear Models. For the dependent variable HDDS, the Poisson regression model is as follows:

$$\mu_i = E(HDDS_i|X_i) = exp(\beta_0 + \beta_i X_i) \quad (20)$$

where HDDS is the HDDS of household i, X_i refers to the vector of independent variables, and β_o and β_i are the constant and the coefficient vector for independent variables, respectively. The alternative the generalized linear model with link log can be written as:

$$log(\mu_i) = x_i'\beta \quad (21)$$

3.2.3.2. Poisson Model

The Poisson distribution is usually used as a standard model for count data and was derived as a limiting case of the binomial distribution by Poisson. GLM's have two primary features; firstly, for some dependent variables μ, the probability distribution of *y* given μ is a member of the exponential family (This distribution is known as the Poisson distribution). Secondly, there is a "link function" which is the transformation (g) that linearizes the expected value of *y*. The Poisson distribution models the probability of *Y* events (in this case, number of food groups consumed (dietary diversity score) by a household) with the formula:

$$Pr(Y = y|\mu) = \frac{e^{-\mu}\mu^y}{y!} \qquad y = 0,1,2 \ldots . 12 \quad (22)$$

$$\mu = E(y|x_k) = Var(y|x_k) = exp(\beta_0 + \beta^T x) \quad (23)$$

The Poisson distribution has a single parameter to be estimated, μ, or the mean, which is also sometimes referred to as the location parameter. The unique feature of the Poisson distribution is that the mean and variance are the same. The higher the value of the mean of the distribution, the greater the variance or variability in the data.

3.2.3.3. Quasi Poisson Model

Using the mean regression function and the variance function from the GLM while leaving the dispersion parameter Φ unrestricted is one strategy for dealing with over-dispersion (Achim, Christian, & Simon, 2008). Therefore, Φ is not assumed to be fixed at 1 but is estimated from the data. Similar coefficients are obtained with this method as with the Standard Poisson model, but over-dispersion is taken into account in the inference (Achim et al., 2008). Subsequently, both the Quasi-Poisson and Sandwich-adjusted Poisson models adopt the estimating function view of the Poisson model and do not correspond to models with fully specified likelihoods.

3.2.3.4. Negative Binomial Model

The Negative binomial distribution is used for modeling over-dispersed data and is a standard generalization of the Poisson distribution. The negative binomial is a two-parameter model – with mean (μ) and dispersion (α) parameters. Let y_i $(i = 1,2, \dots, n)$ be a non-negative integer valued random variable representing the i_{th} outcome and y_i be the associated outcome of interest. The unconditional negative binomial distribution of y_i is expressed as:

$$p(y) = \frac{\Gamma(\alpha+y_i)}{\Gamma(\alpha)y_i!}\left(\frac{\beta}{1+\beta}\right)^{yi}\left(\frac{1}{1+\beta}\right)^{\alpha}, y_i = 0,1,2 \ldots \ldots \quad (24)$$

The above distribution has a mean

$$E(y_i) = \alpha\beta \quad (25)$$

And variance:

$$var(y_i) = \alpha\beta + \alpha\beta^2 \quad (26)$$

3.2.3.5. Poisson Inverse Gaussian Regression (PIG Model)

PIG regression is used to model count data that have a high initial peak and that may be skewed to the far right as well as data that are highly Poisson over dispersed (Hilbe, 2014). The PIG probability distribution, as a variety of Sichel distribution, can be given as:

$$f(y;\mu,\alpha) = \sqrt{\frac{\phi}{2\pi y^3}}exp\left(\frac{-\phi(y-\mu)^2}{2\mu^2 y}\right) \quad (27)$$

whereby $\{y, \mu, \phi\} > 0$.

The negative binomial model and the Poisson inverse Gaussian (PIG) model are both mixture models. The inverse Gaussian model combines Poisson and inverse Gaussian distributions, whereas the negative binomial model combines Poisson and gamma distributions.

In contrast to the gamma distribution that is inherent to the negative binomial model, the PIG model assumes that over-dispersion in a Poisson model is best characterized, or shaped, according

to the inverse Gaussian distribution. The inverse Gaussian is a mixture of Poisson and inverse Gaussian distributions, with an inverse Gaussian variance of μ^3/Φ .

3.2.3.6. Generalized Poisson (GP) Model

Similar to the negative binomial and PIG models, the generalized Poisson has a second parameter, also referred to as the dispersion or scale parameter. Also, like the previously introduced models, the generalized Poisson reduces to Poisson when the dispersion is zero. The nice feature of the generalized Poisson, however, is that the dispersion parameter can have negative values, which indicate an adjustment for Poisson under dispersion.

The generalized Poisson probability function is based on (Consul, 1989) and (Famoye, 1993), and this parameterization on (Harris, Yang, & Hardin, 2012). Suppose Y_i is a count response variable (number of food groups consumed) that follows a generalized Poisson distribution. The probability density function of Y, $i = 1, 2, \ldots . . n$ is given by:

$$f(y_i) = Pr(Y_i = y_i) = \left(\frac{\lambda_i}{1+\alpha\lambda_i}\right)^{yi} \frac{(1+\alpha y_i)^{yi-1}}{y_i!} exp\left[\frac{-\lambda_i(1+\alpha y_i)}{1+\alpha\lambda_i}\right], y_i = 0,1,2 \ldots \qquad (28)$$

The mean and variance of Y are given by:

$$E(Y_i|x_i) = \lambda_i, var(Y_i|x_i) = \lambda_i(1+\alpha\lambda_i)^2 \qquad (29)$$

when α is called the dispersion parameter.

3.2.3.7. Hurdle Models

A hurdle model is primarily used to handle count response variables that have more or fewer zero counts than permitted by the distributional assumptions of the count model being applied to the data. The fundamental concept behind a hurdle model is to divide it into two distinct processes: a binary process generating positive counts (1) versus zero counts (0) (process 1) and another that produces only positive counts (process 2)). A binary model is frequently used to represent the binary process, while a zero-truncated count model is frequently used to represent the positive count process. Values below the hurdle are given the value of 0 and above the hurdle point they are given the value of 1. Starting with the binomial process, suppose that is the probability value when the value for the response variable is zero and that 1 is the probability value when the response variable is a positive integer. The probability mass function is given by:

$$Pr(Y = y) = \begin{cases} \pi, & y = 0 \\ 1 - \pi, & y = 1,2, \dots \end{cases} \qquad (30)$$

The zero-truncated Poisson has the probability mass function:

$$pr(Y = y|Y \neq 0) = \begin{cases} \frac{\lambda^y}{(e^y - 1)!}, & y = 1,2, \dots . \\ 0, & Otherwise \end{cases} \qquad (31)$$

Thus, the unconditional probability mass function for Y is

$$Pr(Y = y| \neq 0) = \begin{cases} \pi, & y = 0 \\ (1 - \pi)\frac{\lambda^y}{(e^\lambda - 1)y!}, & y = 1,2, \dots . \end{cases} \qquad (32)$$

3.2.3.8. Zero-Inflated Models

There are some situations where a major source of over-dispersion is a mass of zero counts, and the resulting over-dispersion cannot be modeled accurately with negative binomial estimation. In this regard, one can use zero-inflated (Poisson or negative binomial) estimation methods.

The first component is the binary, which is usually modeled as a logit or probit. Unlike the hurdle model, though, in which the binary component has 1's consisting of all counts greater than 0 and 0's consisting of the 0 counts in the data, the zero-inflated models have a binary component with a value of 1 for all 0's in the data and 0 for all other counts greater than 0. The count component simply models all of the counts from zero to greater than zero. In this sense, the binary model is directly modeling the 0's in the data by revaluing them to 1's. The positive count values are the combined reference level for the binary model. In zero-inflated model, counts are estimated as

$$Pr(Y = 0) = Pr(Bin = 0) + (1 - Pr(Bin = 0)) * Pr(Count = 0) \qquad (33)$$

$$Pr(Y > 0) = (1 - Pr(Bin = 0)) + PDFcount \qquad (34)$$

Since we will use the logic model for the binary component of zero-inflated models, the logic equation for the probability of 0's as $1/(1 + exp(xb))$:

$$Pr(Bin = 0) = \frac{1}{1+exp(x\beta)} \qquad (35)$$

For a zero-inflated Poisson Model with a logit binary component, we have the equation

$$(y = 0) = log(\frac{1}{1+exp(-x\beta_b)} + \frac{exp(-exp(x\beta))}{1+exp(x\beta_b)}) \qquad (36)$$

$$(y > 0) = log\left(\frac{1}{1+exp(-x\beta_b)}\right) - exp(x\beta) + y(x\beta) - log\Gamma\,(y + 1) \qquad (37)$$

where the b in equation 36 indicates the $x\beta$ is a binary model component and without the subscript that $x\beta$ is from the component, in this case Poisson.

3.2.3.9. Conway-Maxwell Poisson Regression Model

The Conway-Maxwell-Poisson (CMP) distribution is a generalization of the Poisson distribution that enables you to model both under dispersed and over dispersed data. The Conway-Maxwell-Poisson distribution is defined as:

$$P(Y_i = y_i|x_i, z_i) = \frac{1}{Z(\lambda_i, v_i)} \frac{\lambda_i^{yi}}{(y_i!)^{vi}}, y_i = 0,1,2 \ldots\ldots \quad (38)$$

where the normalization factor is

$$Z(\lambda_i, v_i) = \sum_{n=0}^{\infty} \frac{\lambda_i^n}{(n!)^{vi}} \quad (39)$$

and

$$\lambda_i = exp(\left(\acute{X}_i\beta\right) \quad (40)$$

$$v_i = -exp((\acute{g}_i\delta) \quad (41)$$

The β vector is a $(k+1) * 1$ parameter vector. (The intercept is β_0, and the coefficients for the k regressors are $\beta_{1,\ldots\ldots}\beta_k$). The δ vector is an $(m+1) * 1$ parameter vector. (The intercept is represented by δ_0, and the coefficients for the m regressors are $\delta_1 \ldots . \delta_k$). The covariates are represented by x_i $\mathrm{x_i}$ and x_i vectors.

One of the restrictive properties of the Poisson model is that the conditional mean and variance must be equal:

$$E\ (y_i|x_i) = V(y_i|x_i) = \lambda_i = exp(\acute{X}_i\beta) \qquad (42)$$

The CMP distribution overcomes this restriction by defining an additional parameter, *v*, which governs the rate of decay of successive ratios of probabilities such that:

$$\frac{P(Y_i=y_i-1)}{P(Y_i=y_i)} = \frac{(y_i)^{vi}}{\lambda_i} \qquad (43)$$

The introduction of the additional parameter v, allows for flexibility in modeling the tail behavior of the distribution. If $v = 1$, the ratio is equal to the rate of decay of the Poisson distribution. If $v > 1$ the rate of decay decreases, enabling you to model processes that have longer tails than the Poisson distribution (over dispersed data). If $v < 1$, the rate of decay increases in a nonlinear fashion, thus shortening the tail of the distribution (under-dispersed data).

3.2.4. Comparison of the Models of Goodness-of-Fit

The goodness of fit of a statistical model describes how well the model fits into a set of observations. The two commonly used goodness-of-fit statistics for model selection are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) calculated as:

$$AIC = -2L + 2q \qquad (44)$$

$$BIC + -2L + qln(N) \qquad (45)$$

Whereby L is the maximum likelihood, q is the number of parameters estimated and N is number of observations. When comparing models, lower values of either the AIC or BIC indicate a better fit. The AIC were mainly used to conclude because they have an advantage that they can be used to descriptively compare all models regardless of whether one is nested or not within another.

3.2.5. Statistical Analysis

Descriptive statistics were generated to summarize the levels of Household Dietary Diversity and the types of food consumed by household members. Different count models were then applied to determine the best fitted model.

3.3. Results

3.3.1. Descriptive Results

An increase in the average number of different food types or groups consumed reflects an improvement in the household's diet. Figure 1 shows that local foods/grains, meats and foods made with oil, fat or butter and sugar/honey were the most consumed food groups. The households consumed less of food made from beans, pears, lentils or nuts, eggs, fruits and cheese, yoghurt, milk or other milk products.

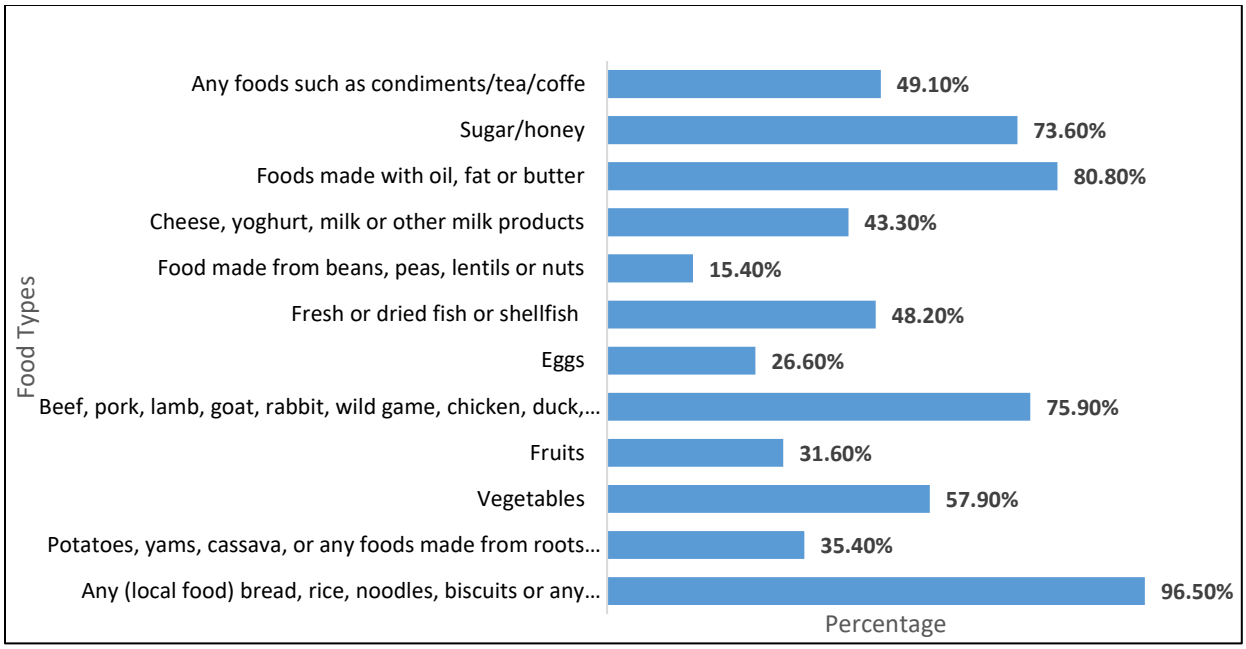


Figure 1: Percent distribution of Food Types consumed by households.

3.3.2. Household Dietary Diversity

Dietary diversity was divided into two categories, Low diversity and High diversity. A cross tabulation was done between the dietary diversity and the demographic variables. Table 11 shows that at least a household consisting of 1-3 members has about 43% of low dietary diversity. About 75.1% of rural residences had low diversity and households that were headed by males, about 53.4 % had low diversity. Of the age groups, at least 25.9 percent of high diversity and 20.6 in the low diversity were headed by a person in the age group of 30-39. Table 11 further shows that 60.1% of individuals of whom main source of income comes from salaries and wages had a high dietary diversity. It was further shown that individuals with just Primary education had about 35 % low diversity and those with secondary education had about 51.8% high dietary diversity. Accessibility to radio, television and internet was also assessed. The table in addition shows that majority did not have access to radio, television and internet. Socio- demographic variables such as place of

residence, number of members in a house, main source of income, educational level, and access to infrastructures such as radios, televisions and internet were significant at 5% level of significance.

Table 11: Frequency distribution: Dietary Diversity and Socio-Demographic Characteristics

Variable		Low Diversity	High Diversity	Chi-Square P-Value
Number of Household Members	1-3 members	1810(43.2%)	2791 (47.4%)	<0.001
	4-6 members	1446(34.5%)	2080(35.4%)	
	7-9 members	634(15.1%)	696(11.8%)	
	>10 member	298(7.1%)	316(5.4%)	
Residence	Rural	3148(75.1%)	2375(40.3%)	<0.001
	Urban	1042(24.9%)	3513(59.7%)	
Sex of Head	Female	1956(46.6%)	2660(45.1%)	0.140
	Male	2240(53.4%)	3234(54.9%)	
Age of Head	10-19 Years	66(1.6%)	55(0.9%)	<0.001
	20-29 Years	537(12.8%)	973(16.5%)	
	30-39 Years	864(20.6%)	1528(25.9%)	
	40-49 Years	849(20.2%)	1331(22.6%)	
	50-59 Years	697(16.6%)	946(16.1%)	
	60+	1183(28.2%)	1061(18.0%)	
Main source of income	Salaries and Wages	1453(34.6%)	3541(60.1%)	<0.001
	Pension	840(20.0%)	576(9.8%)	
	Subsistence Farming	740(17.6%)	453(7.7%)	
	Business Income	284(6.8%)	574(9.7%)	
	Remittances/Grants	512(12.2%)	485(8.2%)	
	Drought/ in-kind Receipts	247(5.9%)	55(0.9%)	
	Commercial Farming	4(0.1%)	46(0.8%)	
	Others	116(2.8%)	164(2.8%)	
Highest level of educational attainment of household head	No Formal Education	1298(21.4%)	607(10.7%)	<0.001
	Primary	1463(35.4%)	1262(22.2%)	
	Secondary	1234(29.9%)	2936(51.8%)	
	Tertiary	113(2.7%)	833(14.7%)	
	Not Stated	20(0.5%)	35(0.6%)	
Access to Radios	Yes	887(40.4%)	1507(48.0%)	<0.001
	No	1309(59.6%)	1631(52.0%)	
Access to Televisions	Yes	509(14.2%)	675(25.8%)	<0.001
	No	3076(85.8%)	1943(74.2%)	
Access to Internet	Yes	183(4.5%)	557(11.8%)	<0.001
	No	3883(95.5%)	4158(88.2%)	

Household dietary diversity indicates how the diet of households varies, consisting of different food types. Figure 2 shows the distribution of household dietary diversity. The diversity of the diet of households was found to be moderate (mean = 6.35).

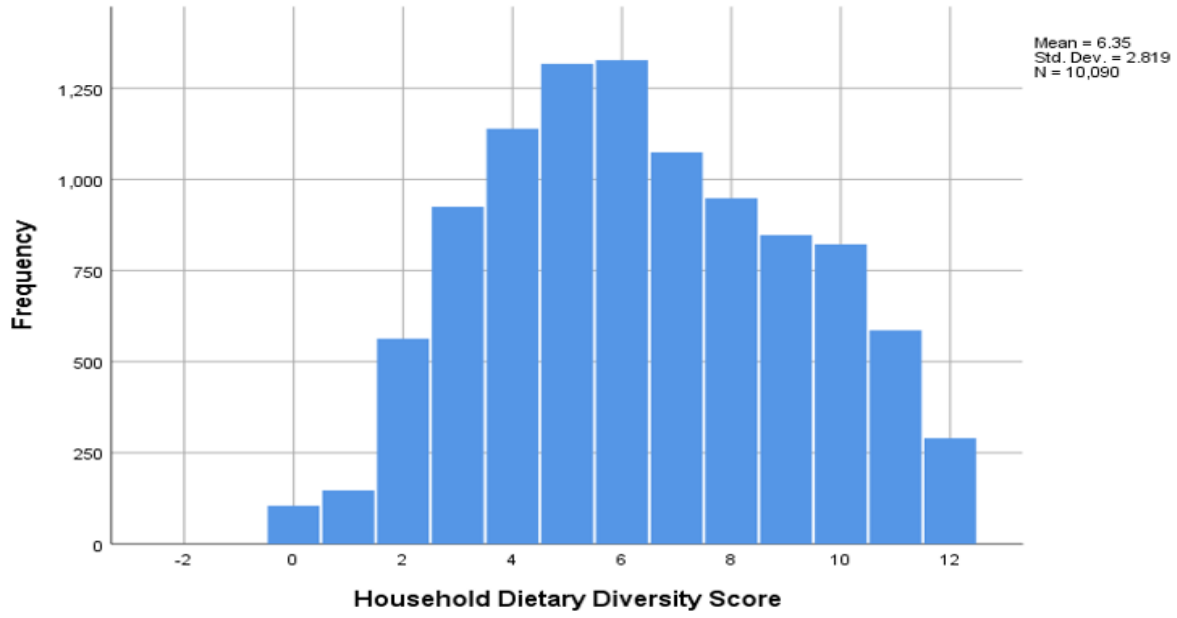


Figure 2: Histogram of Household Dietary Diversity Score (NHIES 2015/16)

3.3.3. Comparative Fit of the Different Models Used

Various count models were fitted, and the best model was selected. The models AIC and BIC are presented in Table 12. Poisson Inverse Gaussian Regression had both the lowest AIC and BIC and was selected to be the best fit.

Table 12: Summary of fitted count regression models (GLM) for NHIES (2015/16)

Model	**2 x Log Likelihood**	**Akaike Information Criterion (AIC)**	**Bayesian Information Criterion (BIC)**	**Δ (AIC – BIC)**
Poisson Regression Model	-13053.21	13103	13253.67	150.67
Negative Binomial Regression	-13053.23	13105	13261.72	156.72
Poisson Inverse Gaussian Regression	**-12805.2**	**12857**	**13013.69**	**157**
Poisson Logit Hurdle	-12958	13058	N/A	N/A
Conway-Maxwell Poisson	-13039.80	13092	13248	156

3.3.3.1. Poisson Inverse Gaussian Regression

The Poisson Inverse Gaussian Regression was found to have the lowest AIC (12857) and BIC (13013.69) compared to other models. Table 13 and 14 of the Poisson Inverse Gaussian Regression shows that the variables Residence, Sex of head of household, Main source of income, Pension, Subsistence Farming, Remittances/grants, Drought/in-kind receipts and others), Level of Education (Primary, Secondary, Tertiary and not Stated) and Access to Televisions had an effect on dietary diversity.

Table 13: Summary of fitted Poisson Inverse Gaussian model (PIG) for NHIES (2015/16)

Intercept	Estimate (B)	Standard Error	t- Value	P- Value
(Intercept)	1.805	0.083	21.713	<0.001 ***
Number of Household Members: 4-6 Members	-0.010	0.021	-0.482	0.630
Number of Household Members: 7-9 Members	-0.035	0.031	-1.110	0.267
Number of Household Members: >10 Members	0.044	0.041	1.076	0.282
Number of Household Members: 1-3 Members	*Reference*			
Residence: Rural	-0.152	0.020	-7.792	<0.001***
Residence: Urban	*Reference*			
Sex of Head of Household: Male	-0.083	0.018	-4.580	<0.001***
Sex of Head of Household: Female	Reference			
Age of Head of Household: 20-29 Years	0.071	0.074	0.968	0.333
Age of Head of Household: 30-39 Years	0.037	0.073	0.499	0.617
Age of Head of Household: 40-49 Years	0.018	0.074	0.244	0.808
Age of Head of Household: 50-59 Years	0.065	0.076	0.864	0.387
Age of Head of Household: 60+ Years	0.132	0.082	1.614	0.107
Age of Head of Household: 10-19	*Reference*			
Main source of income: Pension	-0.182	0.047	-3.847	<0.001***
Main source of income: Subsistence Farming	-0.143	0.032	-4.455	<0.001***
Main source of income: Business income	-0.010	0.031	-0.350	0.727
Main source of income: Remittances/grants	-0.141	0.031	-4.465	<0.001***
Main source of income: Drought/in-kind receipts	-0.442	0.043	-10.173	<0.001***
Main source of income: Commercial Farming	0.203	0.266	0.762	0.446
Main source of income: Others	-0.169	0.046	-3.703	<0.001***

Table 14: Summary of fitted Poisson Inverse Gaussian model (PIG) for NHIES (2015/16) continued…..

Intercept	Estimate (B)	Standard Error	t- Value	P- Value
Main source of income: Salaries and wages	*Reference*			
Education: Primary	0.091	0.025	3.651	<0.001***
Education: Secondary	0.241	0.026	9.339	<0.001***
Education: Tertiary	0.452	0.048	9.383	<0.001***
Education: Not Stated	0.249	0.105	2.372	0.018*
Education: No formal education	*Reference*			
Access to Radios: Yes	-0.037	0.020	-1.894	0.028.
Access to radios: No	*Reference*			
Access to Televisions: Yes	-0.117	0.026	-4.437	<0.001***
Access to Televisions: No	*Reference*			
Access to Internet aerial/dish: Yes	-0.052	0.038	-1.377	0.169
Access to Internet aerial/dish: No	*Reference*			
Signif. Codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. No. of observations in the fit: 3037, D.F: 26**				
Residual Degrees of Freedom:3011 at cycle: 20. Degrees of Freedom for the fit: 26. Global Deviance: 12864.51				
AIC: 12857. BIC: 13013.69				

3.4. Discussion

Count models have been applied in many studies dealing with count data. Comparison of different models is a very important aspect in identifying the best possible fit model. The objective of this chapter was to apply count models in dietary diversity in Namibia and it was comparison in nature whereby the model with the lowest AIC was regarded as the best model. Furthermore, this study aimed at establishing the risk factors of dietary diversity. The dietary diversity score was established, and different socio-economic variables have also been assessed against the diversity of a household.

Eight (8) regression models were fitted for the analysis of count data. These models have been described in the context of modeling Household Dietary Diversity Score. The results indicate that Poisson Inverse Gaussian Regression performs better than Poisson, Negative binomials, Hurdle Regressions and Zero-Inflated models. The Poisson Inverse Gaussian model had the least AIC (12857). AIC is usually interpreted in the lower as better fit. The study further revealed that the hurdle Poisson and Zero inflated Poisson are better compared to Poisson as depicted by their lower AICs of 13058 and 13087 respectively. The PIG model assumes that over-dispersion in a Poisson model is best described, or shaped, according to the inverse Gaussian distribution rather than the gamma distribution that is inherent to the negative binomial model (Hilbe, 2014).

This study confirms the notion that households consumed less of food made from beans, peas, lentils or nuts, eggs, fruits and cheese, yoghurt, milk or other milk products (Figure 1). This can be due to inability to purchase animal products which are often sold in the local market and rarely consumed at home (IFPRI, 2015). The study concurs with findings from Pazvakawambwa & Nickanor (2018) that dietary diversity in Namibia is low and primarily starch based which limits

the consumption of essential micronutrients leading to a lack of sufficient nutrient- dense foods to achieve micronutrient adequacy. Male-headed households had significantly low dietary diversity scores compared to those in the female-headed households. Additionally, sex and educational level of the household head contribute to improved dietary diversity.

A previous study in Tanzania reported that households who were provided with nutritional education improved the quality of their household diets (Pillai, Kinabo, & Krawinkel, 2016). Households headed by women significantly had higher household dietary diversity scores. This could be because in Sub Saharan Africa (SSA), women- controlled income often has greater benefits for the nutrition, health and well-being of all household member, especially children, than men-controlled income (FAO, 1997). Dietary problems may be primarily quantitative in the most underprivileged areas, such as rural areas during seasonal food shortages or urban areas under acute poverty. As a result, the dietary deficiency then appears to be chiefly energy related (Savy, Martin-Prevel, Sawadogo, Kameli, & Delpeuch, 2005).

In Germany, Thiele & Weiss, (2003) noted that household size, age, sex composition, employment status and level of education were the major determinants of food diversity. The same trend was found in this study. Education is likely to have an impact on the household's nutritional knowledge and skills to conceptualize and use nutritional promotional messages, which consequently contribute to better dietary diversity (Rajendran, et al., 2017).

3.5. Conclusions

Food and nutrition insecurity remains a complex problem in developing countries such as Namibia. This study concluded that the Inverse Gaussian Poisson was the best fit to analyze Household Dietary Diversity. Our study revealed that the diet among households was moderate (mean=6.3

out of 12), and that the intake of foods from beans, pears, lentils or nuts, eggs, fruits was low among the households. Male-headed households and low-income households had significantly low dietary diversity scores compared to those in the female-headed households.

3.6. Acknowledgements

The authors had support from the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

CHAPTER 4: A BIVARIATE COUNT MODELLING APPROACH IN ANALYZING CONVENIENCE AND NON-CONVENIENCE CONSUMPTION OF FOOD PREFERENCE IN WINDHOEK, NAMIBIA

Laina Mbongo[2]*, Lawrence Kazembe, Lillian Pazvakawambwa
Department of Computing, Mathematics and Statistical Sciences,
University of Namibia, Windhoek, Namibia

Abstract

Globalization coupled with urbanization has placed a significant effect on the food systems of developing countries in the world, leading to a myriad lifestyle change that has become one of the most important influences in dietary patterns. The nutritional transition has affected the dietary pattern and nutrient intake greatly and has led to a rise in the purchases and consumption of processed and convenience foods, which are prepared foods designed for simplicity of consumption but are associated with rising rates of diet-related non-communicable diseases in Low- and middle-income countries (LMICs). This chapter jointly analyzed paired consumption of both convenience and non-convenience food that are exhibiting correlations, using a household food security survey, conducted in Windhoek. This is illustrated by applying both the untruncated and the right-truncated bivariate Poisson models to examine factors that influence convenience and non-convenience food consumption patterns both on a weekly and monthly basis. The results found that overall, the untruncated (conditional/marginal) model fitted the data better. Whereas the consumption of food on a monthly basis was more on the non-convenience foods, the purchases of convenience was frequent on a weekly basis and from multiple food sources. The choice of food purchase both at weekly and monthly preference was influenced by sex, marital status, education level and work status.

Key words: Convenience and Non-Convenience foods, Bivariate Poisson Regression, (Un)Truncated, urbanization.

[2]* Corresponding Author Email: inambongo@gmail.com

4.1. Introduction

Globalization has had a significant effect on the food systems of developing countries around the world. As a complex and multifaceted phenomenon, globalization is considered by some as a form of capitalist expansion which entails the integration of local and national economies into a global, unregulated market economy (Shalmali, 2007). Forces manifested by globalization, such as market and trade liberalization, capital flow, and urbanization have changed the nature of food systems by increasing the diversity and affordability of food, but also by changing its quality and nutritional value (Black, 2016).

In developing countries undergoing rapid urbanization combined with globalization, the process includes changes in the sociocultural environment such as mass media marketing and the widespread availability of less traditional foods, which play an important role in influencing tastes and preferences. Consumer's food choices or preferences have been attributed to factors such as growing foreign direct investments that contributes to the rise of fast-food restaurants and western-style supermarkets by offering greater variety, quality, convenience and competitive prices in high-value added foods. In urban areas, men and women are driven into the workforce in order to maintain their lifestyles. Working hours and commuting times are often long and, with growing numbers of family members entering the workforce, there is less time available to prepare food and hence there is a greater desire and necessity to consume meals outside the home.

Traditional meals and mealtimes are replaced by spontaneous often unplanned food purchases on street corners or in small kiosks that provide family members with at least one and often several meals per day. Street foods are becoming increasingly important as both a cheap and quick meal option and as an income-generating strategy. Secondary factors such as marketing, advertising, the

appeal of new products, new retail outlets including supermarkets and multinational fast-food chains contribute to dietary adaptation and convergence. Aside from the driving force of time constraints, part of the rapid adoption of new foods in the diet stems from successful advertising (Lang, 2003).

Convenience foods are prepared food designed for simplicity of consumption. These foods products are prepared food products that can be sold as ready-to-eat dishes; as room-temperature, precooked and frozen products and hot products. Convenience food can include products such as candy, soft drinks, fast food; nuts, fruits, processed meats and cheeses and canned products such as soups and pasta dishes (Jackson et al., 2018). Consumption is often associated with rising rates of diet-related non-communicable diseases in Low- and Middle-Income Countries (LMICs) (Khan, 2013).

Studies of Dietary patterns have become popular in nutritional epidemiology (Smith, Emmett, Newby, & Northstone, 2011). Traditional analysis examined diseases in relation to a single or a few nutrients or foods. However, people do not eat isolated nutrients. Instead, they eat meals consisting of a variety of foods with complex combinations of nutrients. A typical household would consume either convenience or consume non-convenience or consume both types of food. The high degree of inter-correlation among nutrients as well as among foods makes it difficult to attribute effects to single dietary components. For these reasons, a more prudent analysis that simultaneously estimates factors that are associated with the preference of non-convenience and convenience food groups is required.

Event counts such as the number of convenience and non-convenience foods consumed are likely to be jointly dependent (Karlis & Ntzoufras , 2005). To understand fully the drivers of preference

of food choices (Convenience vs. Non-convenience), we will employ multidimensional measures that jointly estimate the risk factors and are able to accommodate heterogeneity attributes. Different count data may possess different characteristics and therefore cannot be used with particular count data models. Poisson regression model provides a basis for the analysis of count data. Due to the over-dispersion and/or excess number of zeros that are frequently present in empirical count data sets, the Poisson regression model is frequently only of limited use. In order to analyze bivariate count data, the plain Poisson regression model needs to be extended.

Bivariate Poisson models are appropriate for modeling paired count data exhibiting correlation and require joint estimation (Karlis & Ntzoufras, 2005). The application of bivariate count models often assumes a bivariate Poisson distribution which assumes the conditional mean of each count variable equals the conditional variance. One more shortcoming of commonly used bivariate count models is that they can only accommodate non-negative correlation between the paired counts.

The bivariate Poisson is the most widely used model for bivariate counts. It was proposed by (Holgate, 1964) and presented by (Johnson & Kotz, 1969). Leiter & Hamdan (1973) proposed bivariate probability models applicable to traffic accidents and fatalities. Several approaches have been discussed by various authors with the development of bivariate Poisson distribution with various assumptions. Amongst others, the most comprehensive one has been proposed by (Kocherladota & Kocherlakota , 1992). The bivariate Poisson form is further shown using a trivariate reduction method (Jung, 1993) allowing for correlation between the variables. Karlis & Ntzoufras (2005) implemented the maximum likelihood estimation for bivariate Poisson models and their diagonal inflated variations. Furthermore, (Islam & Chowdhurry, 2015), (Chowdrhurry & Islam, 2016), (Islam & Chowdhurry, 2017) and (Chowdhury & Islam, 2019) developed

untruncated, zero-truncated and right truncated bivariate Poisson model for covariates dependence based on the extended generalized linear model. Despite vast application of bivariate poisson regression models in count data, there is no literature on application of bivariate poisson in the area of food consumption. This study thus extends bivariate count modelling approach to analyze convenience and non-convenience consumption of food preference.

4.2. Materials and Methods

4.2.1. The AFSUN-HCP Data

This study used the AFSUN-HCP Household Food Security Baseline Survey (2016) which collected a wide range of demographic, economic and food consumption, and sourcing data at the household level. Households surveyed in the ten constituencies of Windhoek were identified using a two-stage sampling design. Primary sampling units (PSUs) were first randomly selected from a master frame developed and demarcated for the 2011 Population and Housing Census. Within the 10 constituencies, a total of 35 PSUs were selected covering the whole of Windhoek, and 25 households were systematically selected in each PSU. The sampled PSUs and households were located on maps, which were used to select households for in-person interviews. Household heads (or their spouses/ partners) were recruited to complete the survey.

4.2.2. Outcome Variables

We consider two possibly dependent and correlated response variables namely Y_1, which is the total number of households consuming convenience food (CONVENIENCE) and Y_2, which is the total number of times each household consumes non-convenience foods (NON-CONVENIENCE).

4.2.3. Explanatory Variables

The regressor variables in this study are: Age of head of household (1- <19, 2- 20-29, 3- 30-39, 4- 40-49, 5- >50), Sex of head of household (1-Male, 2-Female), Marital Status Sex of head of household (1- Unmarried, 2- Married, 3- Living together/cohabitating, 4- Widowed), Educational level Sex of head of household (1- None, 2- Primary education, 3- Secondary education, 4- Tertiary education) and Work Status Sex of head of household (1- Self-employed, 2- Formal employed, 3- Unemployed).

Convenience and non-convenience foods have been categorized based on the source purchased and further measured on the number of times a household made use of a source, weekly or monthly basis and estimates were made separately for each. Convenience food sources include fast foods/take-away, restaurants, spaza/tuckshop, Street seller/trader/hawker and begging from the streets while non-convenience food sources comprise supermarkets, small-shops, Open markets, and food grown by households in rural areas.

4.2.4. Review of Bivariate Count Models Data

4.2.4.1. Bivariate Poisson Regression

The Bivariate Poisson model is the most used model among the bivariate count models. A well-established approach, according to (Chou & Steenh, 2011), is to generate the bivariate Poisson distribution by convolutions of Poisson random variables (Kocherladota & Kocherlakota , 1992). Let Y_1 represent the convenience food and Y_2 be non-convenience food consumed by household members over a week and month time period:

$$y_{1i} = y_{1i}^{*} + \mu_i \qquad (46)$$

$$y_{2i} = y_{2i}^{*} + \mu_i \quad (47)$$

Where y_{1i}^{*} ~Poisson (λ_{1i}) and y_{2i}^{*} ~Poisson (λ_{2i}) are independently distributed. The joint probability density function of the bivariate Poisson can be defined as follows:

$$f(y_{1i}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{exp(-\lambda_{ji})\lambda_{j_i}^{y_{ji}}}{y_{ji!}}\right] exp(-\lambda_3) \sum_{s=0}^{m} \binom{y_{1i}}{s} \binom{y_{2i}}{s} s! \left(\frac{\lambda_3}{\lambda_{1i}\lambda_{2i}}\right)^{s} \quad (48$$

Where $m = min(y_{1i}, y_{2i})$ and $\lambda_{ji} = exp(x_{ji}\beta)$. The Poisson distribution is known to be restrictive due to its equi-dispersion property, viz., with the mean and variance both equal to μ_j.

$$E(Y_{it}) = Var(Y_{it}) = \mu_{it}, \quad t = 1,2 \quad (49)$$

The model allows only for non-negative correlation between the counts and restricts the mean to be equal to the variance for each of the respective marginal distributions (Chou & Steenh, 2011). The marginal distributions of the model are still Poisson, and the correlation between the two count variables (conditioned on the covariates) is individual specific, being a function of the λ_{ji} and λ_3.

$$corr(y_1, y_2) = \lambda_3/\sqrt{(\lambda_1)(\lambda_2)} \quad (50)$$

The maximum likelihood estimator of the correlation between y_1 y_2 shown by (Leiter & Hamdan, 1973) is:

$$\hat{corr}(y_1, y_2) = \left(\frac{\bar{y}_2}{(\bar{y}_1 + \bar{y}_2)}\right)^{\frac{1}{2}} \quad (51)$$

4.2.4.2. Bivariate Truncated Poisson Model

The bivariate truncated models are mostly used if the observations (y_{1i}, x_{1i}) or (y_{2i}, x_{2i}) or both in some ranges are totally lost and the joint distribution of observed counts is restricted. When the count data are only observed over a portion of the response variable's range, this is referred to as truncation (Cameron & Trivedi , 1999). A series may be truncated from below (left truncated) or truncated from above (right truncated) or un-truncated (Gurmu & Elder, 2008). The truncated models can be grouped as follows:

4.2.4.3. Un- Truncated Bivariate Poisson

The untruncated model in this study is defined as, the number of occurrences of convenience food Y_1 over a week or month follows Poisson distribution with parameter λ_1 and the occurrence of non-convenience food, Y_2, is also Poisson with parameter, $\lambda_1 y_1$. The joint pdf of Y_1 and Y_2 is:

$$g(y_1, y_2) = \frac{e^{-\lambda_1} {\lambda_1}^{y_1} e^{-\lambda y_1} (\lambda_2 y_1)^{y_2}}{y_1! y_2!},\ y_1 = 0,1, \ldots.: \lambda_1, \lambda_2 > 0 \qquad (52)$$

4.2.4.4. Zero- Truncated Bivariate Poisson

The joint distribution of the Zero Truncated Bivariate Poisson model can be obtained from the marginal and conditional distributions (Chowdhury & Islam, 2019):

$$g(y_1, y_2) = \ g_2(y_2|y_1).\, g_1(y_1) = \frac{(\lambda_2 y_1)^{y_2}}{y_2!(e^{\lambda_2 y_1} - 1)} * \frac{{\lambda_1}^{y_1}}{y_1!(e^{\lambda_1} - 1)} = \frac{(\lambda_2 y_1)^{y_2} {\lambda_1}^{y_1}}{y_1!\, y_2!(e^{\lambda_1} - 1)(e^{\lambda_2 y_1} - 1)} \qquad (53)$$

where the link functions are:

$$ln\, \lambda_1 = X'\beta_1 \; and \; ln\, \lambda_2 = X'\beta_2 \qquad (54)$$

The log-likelihood function is:

$$ln\,L = \sum_{i=1}^{n}\left[y_{1i}(x_i'\beta_1) - ln(y_{1i}!) - ln\left(e^{e^{x_i\beta_1}} - 1\right) + y_{2i}(x_i'\beta_2) + y_{2i}\,ln(y_{1i}) - ln(y_{2i}!) - ln\left(e^{y1i^{e^{x_i}'^{\beta_2}}} - 1\right)\right] \quad (55)$$

4.2.4.5. Right – Truncated bivariate Poisson.

Right truncation results from loss of observations greater than some specified values (Cameron & Trivedi , 1999). The joint distribution of the right truncated bivariate Poisson distribution for number of occurrences of convenience food, Y_1, in a week or month interval and number of occurrences of non-convenience food, Y_2, can be represented by:

$$g(y_1, y_2) = g(y_2|y_1).g(y_1) = c_1c_2e^{-\lambda_1}{\lambda_1}^{y_1}e^{-\lambda_2y_1}(\lambda_2y_1)^{y_2}/(y_1!\,y_2!) \quad (56)$$

the bivariate exponential form for the joint distribution of Y_1 and Y_2 can be shown as:

$$g(y_1, y_2) = e^{\{y_1\,ln\,\lambda_1 + y_2\,ln\,\lambda_2 - \lambda_1 - \lambda_2y_1 + y_2\,ln\,y_1 - ln\,y_1! - ln\,y_2! + ln\,c_1 + ln\,c_2\}} \quad (57)$$

The loglikelihood function is:

$$ln\,L = \sum_{i=1}^{n}\{y_{1i}\,ln\,\lambda_1 + y_{2i}\,ln\,\lambda_2 - \lambda_1 - \lambda_2y_{1i} + y_{2i}\,ln\,y_{1i}! - ln\,y_{21}! + ln\,c_1 + ln\,c_{2y_1}\} \quad (58)$$

$$= \sum_{i=1}^{n}\{y_{1i}x\grave{}_{1i}\beta_1 + y_{2i}x\grave{}_{2i}\beta_2 - e^{x\grave{}_{1i}\beta_1} - e^{x\grave{}_{1i}\beta_2}y_{1i} + y_{2i}\,ln\,y_{1i} - ln\,y_{1i}! - ln\,y_{2i}! + ln\,c_1 + ln\,c_{2y_1}\} \quad (59)$$

4.2.5. Other Bivariate Regression Models

4.2.5.1. Bivariate Negative Binomial Regression Model

Lakshminarayana et al., (1999) defined a bivariate Poisson distribution as a product of Poisson marginals with a multiplicative factor and correlation coefficient can be positive, zero, or negative depending on the value of λ, the multiplicative factor parameter. Famoye (2010) adopted adopted a similar approach as Lakshminarayana et al., (1999) and defined the bivariate negative binomial distribution as a product of negative binomial marginals. The probability function of the bivariate negative binomial distribution is given by:

$$P(y_1, y_2) = \binom{m_1^{-1}+y_1-1}{y_1} \theta_1^{y_1}(1-\theta_1)^{m_1^{-1}} \binom{m_2^{-1}+y_2-1}{y_2} \theta_2^{y_2}(1-\theta_2)^{m_2^{-1}} \times [1 + \lambda(e^{-y_1} - c_1)(e^{-y_2} - c_2)], \quad (60)$$

Where $c_t = E(e^{-Y_t}) = [\frac{1-\theta_t}{1-\theta_t e^{-1}}]$ $(t = 1, 2)$ and $y_1, y_2 = 0,1,2 \ldots$. Furthermore, the marginal distributions of $Y_t (t = 1, 2)$ is defined as a negative binomial with the following mean and variance:

$$\mu_t = \frac{m_t^{-1}\theta_t}{1-\theta_t} \quad (61)$$

$$\sigma_t^2 = m_t^{-1}\theta_t/(1-\theta_t)^2 \quad (62)$$

Additionally, the correlation coefficient can either be positive, zero, or negative depending on the value of the multiplicative factor parameter λ, is defined by:

$$\rho = \lambda c_1 c_2 A_1 A_2/(\sigma_1 \sigma_2) \quad (63)$$

4.2.5.2. Bivariate Generalized Poisson Regression Model (BGPR)

Famoye (2010) defined a BGPR as a product of univariate generalized Poisson marginals which allows negative, zero, or positive correlation.

The probability distribution of the BGPR is given by (Famoye, 2010b):

$$P(y_1, y_2) = \prod_{t=1}^{2}\left[\frac{\theta_t^{y_t}(1+\alpha_t y_t)^{y_t-1}}{y_t!} exp[-\theta_t(1+\alpha_t y_t)]\right][1+\lambda(e^{-y_1}-c_1)\,(e^{-y_2}-c_2)] \quad (64)$$

Whereby $c_t = E(e^{-Y_t}) = \exp[\theta_t(s_t - 1)$. Additionally, the marginal distributions of $Y_t(t = 1, 2)$ is defined as a negative binomial with the following mean and variance:

$$\mu_t = \theta_t/(1-\alpha_t\theta_t) \quad (65)$$

$$\sigma_t^2 = \theta_t/(1-\alpha_t\theta_t)^3 \quad (66)$$

The correlation coefficient can be written as $\rho = \sigma_{12}/(\sigma_1\sigma_2) = \lambda(c_{11} - c_1\mu_1)(c_{22} - c_2\mu_2)/(\sigma_1\sigma_2)$. The correlation coefficient can either be positive, zero or negative depending on the value of the multiplicative factor, λ (Famoye, 2010b).

4.2.5.3. Bivariate Poisson Inverse Gaussian (BPIG)

Suppose Y_1, Convenience foods, and Y_2, non-Convenience foods, are two random variables that are Poisson distributed and independent from each other and has mean $\nu\mu_1$ and $\nu\mu_2$ with variance $\mathrm{V}ar(Y_1) = \mu_1 + \mu_1^2\tau$ and $Var(Y_2) = \mu_2 + \mu_2^2\tau$. Variable V is defined as a random variable that has an Inverse Gaussian distribution with the following probability density function (Mardalena et al., 2021):

$$g(v) = (2\pi\tau v^3)^{-0.5}e^{-(v-1)^2/2\tau v}, \quad v > 0 \quad (67)$$

Furthermore, the BPIG distribution based on the inverse Gaussian mixture distribution is defined by the following joint distribution:

$$f(y_1, y_2;\ \mu_1, \mu_2, \tau) = \left(\frac{2z}{\pi}\right)^{\frac{1}{2}} \frac{\mu_1^{y_1}\mu_2^{y_2} e^{\frac{1}{\tau}} K_{s(z)}}{(z\tau)^{y_1+y_2}\ y_1! y_2!} \quad (68)$$

With $= y_1 + y_2 - \frac{1}{2}$, $z = \sqrt{\frac{1}{\tau^2} + \frac{2(\mu_1+\mu_2)}{\tau}}$, and $K_{y_1+y_2-\frac{1}{2}}\left(\frac{1}{\tau}\sqrt{1+2\tau(\mu_1+\mu_2)}\right)$.

The Bivariate Poisson Inverse Gaussian Regression is defined as a regression model with two correlate variables (Mardalena et al., 2021). Suppose y_{ij} is the jth response variable for the ith observation and given a random sample $(Y_{i1}, Y_{i2}) \sim BPIG\ (\mu_{ij}, \tau)$ where $i = 1,2, \ldots, n$ and $j = 1,2$, then the BPIGR model can be defined as follows:

$$ln\left[\frac{E(Y_{ij})}{q_i}\right] = X_i^T \beta_j \quad (69)$$

Whereby $E(Y_{ij}) = \mu_j = q_i e^{X_i^T \beta_j}$, q_i is the exposure variable, $X_i^T = [1 \quad x_{1i} \quad x_{2i} \quad \ldots \ x_{pi}]_{1\times(p+1)}$ is the kth predictor variable vector $(k = 1,2, \ldots, p)$ for the ith observation $(i = 1,2, \ldots . n)$ and the jth response variable $j = 1,2,\ \beta_j = [\beta_{jo} \quad \beta_{j1} \quad \beta_{j2} \ldots \ldots \ldots \beta_{jp}]$ is a regression coefficient vector with $(k+1) \times 1$ dimension for the jth response variable.

4.2.5.4. Bivariate Poisson-Laguerre Polynomial Model

If $g(v_{1i}, v_{2i})$ is approximated by Laguerre polynomial of order one, we obtain the bivariate Poisson- Laguerre polynomial density given by:

$$f(y_{1i,}, y_{2i}|x_i) = \left[\prod_{j=1}^{2} \frac{(\theta_{ji})^{y_{ji}}}{y_{ji!}}\right] M^{(y_1,y_2)}(-\theta_{1i}, -\theta_{2i}) \quad (70)$$

$$Where\ M^{(y_i,y_2)}(-\theta_{1i}, -\theta_{2i}) = \left[\prod_{j=1}^{2} \frac{\Gamma(y_{ji}+\alpha_j)}{\Gamma(\alpha_j)}\ \lambda_j^{\alpha j}\left(\lambda_j + \theta_{ji}\right)^{-(a_j+y_{ji})}\right] \Psi_i \quad (71)$$

$$With\ \lambda_j = \frac{1}{1+\rho_{11}^2}\left[\alpha_j + \rho_{11}^2(\alpha_j + 2)\right] \quad (72)$$

$$And\ \Psi_i = \frac{1}{1+\rho_{11}^2}[1 + 2\rho_{11}^2\sqrt{\alpha_1\alpha_2}(1-n_{1i})(1-n_{2i}) + \rho_{11}^2\alpha_1\alpha_2(1-2n_{1i}+n_{1i}\xi_{1i})(1-2n_{2i}+ n_{2i}\xi_{2i})] \quad (73)$$

$$n_{ji} = \frac{y_{ji}+\alpha_j}{\alpha_j}\left(1+\frac{\theta_{ji}}{\lambda_j}\right)^{-1}\ and\ \xi_{ji} = \frac{y_{ji}+1+\alpha_j}{\alpha_j}\left(1+\frac{\theta_{ji}}{\lambda_j}\right)^{-1} \quad (74)$$

Unlike the bivariate Poisson-lognormal distribution, the Poisson-Laguerre polynomial model has a closed form and can be easily implemented within the likelihood framework (Chou & Steenh, 2011).

4.2.5.5. Bivariate Hurdle and Zero-Inflated Model

When the observed data shows a high frequency of the zero-zero condition ($Y_1 = 0, Y_2 = 0$), zero-modified count models are applied. There are two approaches to treating this issue. First, the bivariate hurdle model, which consists of two parts: a binary outcome model (logit or probit) in the first part and a bivariate truncated count model in the second (Mullahy,1986). The interpretation that positive observations result from passing the zero-zero hurdle or threshold is made possible by this partition. The bivariate hurdle model is appealing because it reflects a two-part decision-making process (Chou & Steenh, 2011). The probability density function of the bivariate hurdle model is given by:

$$h(y_{1i}, y_{2i}|x_i) = \begin{cases} \pi_I & y_{1i} = 0, y_{2i} = 0 \\ (1-\pi_i)\frac{f(y_{1v}y_{2i}|x_i)}{1-f(y_{1i}=0,y_{2i}=0)} & y_{1i} > 0,\ \ y_{2i} > 0 \end{cases} \quad (75)$$

where $\pi_i = \Pr(y_{1i} = 0, y_{2i} = 0)$ is defined as the cumulative density function (CDF) of the logit or probit regression selection model and $\frac{f(y_{1v}y_{2i}|x_i)}{1-f(y_{1i}=0,y_{2i}=0)}$ is the probability density function of a bivariate truncated count regression model.

Secondly, another approach to model excess zeros in the count data is the bivariate zero-inflated count models. Bivariate zero-inflated model assumes that the zero counts come from two sources not one source as in the bivariate hurdle model (Chou & Steenh, 2011). The zero-inflated model is used when a count data set shows a large proportion of zeros. A bivariate zero-inflated model can be constructed by increasing the probability of the event $(Y_1 = 0, Y_2 = 0)$ and decreasing the other joint probabilities.

A logit or probit model is used to determine the probability of counts being the zero-zero state. The bivariate zero-inflated probability density function is given by:

$$h(y_{1i}, y_{2i}|x_i) = \begin{cases} \pi_i + (1-\pi_i)f(y_{1i} = 0, y_{2i} = 0) & y_{1i} = 0, y_{2i} = 0 \\ (1-\pi_i)f(y_{1i}, y_{2i}|x_i) & y_{1i} > 0,\ \ y_{2i} > 0 \end{cases} \qquad (76)$$

where $\pi_i = \Pr(y_{1i} = 0, y_{2i} = 0)$ is the cumulative density function (CDF) of the logit or probit regression, and $f(y_{1i}, y_{2i}|x_i)$ is the density function.

4.2.5.6. Bivariate Censored Model

A sequence is said to be censored from below (left censored) or censored from above (right censored). When high counts are not observed, censored samples may result, or they may be required by the survey's design. Thus, right censoring is the most common form in the analysis of bivariate count models (Chou & Steenh, 2011). Given the bivariate counts are right censored at $r = (r_1, r_2)$ so that $y_{ji} = 1, 2, \dots, r_j \; for \; j = 1,2..$ Letting $f(y_{1i}, y_{2i};\ \varphi)$ denote the complete

bivariate density (Gurmu and Elder, 2000), the log-likelihood function for the right-censored bivariate count model is:

$$LL(y_1 y_2|\varphi) = \sum_{i=1}^{n} d_i[ln\, f(y_{1i}, y_{2i};\ \varphi)] + [1 - d_i]\, ln[\, 1 - \sum_{l=0}^{r_1-1}\sum_{m=0}^{r_2-1} f(y_{1i} = l, y_{2i} = m;\ \varphi) \qquad (77)$$

where $d_i = 1$, if y falls in the uncensored region, and $d_i = 0$ otherwise.

4.2.5.7. **Diagonal Inflated Bivariate Poisson Model**

One drawback of the bivariate Poisson model is that since its marginal distributions are Poisson distributions, which demand that the mean and variance be equal, they cannot handle overly or inadequately scattered data (Yunteng, Yao-Jan, Jonathan, & Yinhai, 2011). Karlis & Ntzoufras (2005) proposed the diagonal inflated bivariate Poisson model to fix this problem. This model uses a more general form developed on the basis of zero-inflated models and the probabilities of the diagonal elements are inflated in the probability table. The diagonal inflated bivariate Poisson model can be defined based on the bivariate regression model as follows (Yunteng et al., 2011).

$$f_{IBP}(x, y) = \begin{cases} (1 - p_m) f_{BP}(x, y|\lambda_1, \lambda_2, \lambda_3) & x \neq y \\ (1 - p_m) f_{BP}(x, y|\lambda_1, \lambda_2, \lambda_3) + p_m f_D(x|\theta, J), & x = y \end{cases} \qquad (78)$$

where $p_m$is the mixing proportion, $p_m f_D(x|\theta, J)$ is the probability mass function of a discrete distribution $D(x; |\theta)$. $D(x; |\theta)$ can be a Poisson, geometric, or a simple discrete distribution. The marginal distributions of a diagonal inflated bivariate Poisson regression model are mixtures of distributions with one Poisson component.

4.2.6. Comparison of the Models of Goodness-of-Fit

The goodness of fit of a statistical model describes how well it fits into a set of observations. The two commonly used goodness-of-fit statistics for model selection are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) calculated as:

$$AIC = -2L + 2q \quad (79)$$

$$BIC + -2L + qln(N) \quad (80)$$

When comparing models as to fit, lower values of either the AIC or BIC indicate a better fit. The AIC were mainly used to conclude because they have an advantage that they can be used to descriptively compare all models regardless of whether one is nested or not within another.

4.2.7. Statistical Analysis

Descriptive statistics were generated to summarize the levels of convenience and non- convenience food consumption by household members. In this study, we fitted a bivariate Poisson regression model using both the joint and conditional arguments. The Bivariate Poisson model is recommended for modelling paired count data exhibiting correlation. Estimations and tests for over and under-dispersion for both the right truncated and the Untruncated Bivariate Poisson regression to relate convenience and non-convenience food consumption with bio-demographic and socio-economic variables were performed. R package for bivariate Poisson GLM with covariates “bpglm” was used to fit the models (Chowdhury & Islam, 2019).

4.3. Results

4.3.1. Frequency Distribution of Consumption of Convenience and Non-convenience Food

Table 15 shows frequencies of household's consumption of convenience and non-convenience foods. Households purchased convenience foods on a weekly basis were more from street sellers/traders/hawkers (46.1%) and Spaza/Tuck-shops (33.7%), while 13.3 percent were obtained food from fast foods/Takeaways and 6.2 percent from restaurants. However, monthly purchases were fewer and increased for non-convenient food purchases with 80.1 percent of the food purchased on a monthly basis at Supermarkets.

Table 15: Convenience and Non-Convenience Food Sources

Source	**Frequency**	
	Weekly	**Monthly**
Convenience Food Sources		
Fast foods/Take away	60 (13.3%)	64 (56.6%)
Restaurants	28 (6.2%)	18 (15.9%)
Spaza/Tuckshop	152 (33.7%)	12 (10.6%)
Street seller/trader/hawker	208 (46.1%)	17 (15.0%)
Begging	3 (0.7%)	2 (1.8%)
Non-Convenience Food Sources		
Supermarket	174 (30.9%)	544 (80.1%)
Small Shops	115 (20.4%)	36 (5.3%)
Open Markets	272 (48.3%)	83 (12.2%)
Food grown by households in rural areas	2 (0.4%)	16 (2.4%)

Table 16 shows the cumulative total number of purchases from each source constituting convenience and non-convenience food. At least 90.7% of individuals did not shop from any source in a week and 7.5 % utilized only one source within the convenience food sources. Non-Convenience food sources mostly provide food to be cooked at home.

Table 16: Convenience food sources

Convenience Food Sources	**Weekly**		**Monthly**	
	Count	***Frequency (%)***	***Count***	***Frequency (%)***
	0 = fast food/take away	3557 (90.7%)	0	3507 (89.4%)
	1 = Restaurants	295 (7.5%)	1	297 (7.6%)
	2 = Spaza/Tuck-shops	58 (1.5%)	2	94 (2.4%)
	3 = Street seller/trader/hawker	8 (0.2%)	3	17 (0.4%)
	4 = Begging	4 (0.1%)	4	7 (0.2%)
	Mean	0.115	**Mean**	0.395
	Std. dev	0.312	**Std. dev**	0.714

Furthermore, Table 17 shows that at least 7.3 percent visited at least one food source and 3.2 percent visited at least 2 sources on a weekly basis. Additionally, Table 17 shows that 8.9 percent visited at least one source monthly and 7.5 percent 2 sources. Overall, on the non-convenience food sources, households preferred to shop monthly than weekly.

Table 17: Non-convenience food sources

Non-convenience Food Sources	**Weekly**		**Monthly**	
	Count	***Frequency (%)***	***Count***	***Frequency (%)***
	0 = Supermarket	3500 (89.2%)	0	3182 (81.1%)
	1 = Small shops	288 (7.3%)	1	350 (8.9%)
	2 = Open markets	127 (3.2%)	2	296 (7.5%)
	3 = Food grown by households in rural areas	7 (0.2%)	3	94 (4.4%)
	Mean	0.14	**Mean**	0.31
	Std. dev	0.445	**Std. dev**	0.714

4.3.2. Bivariate Distribution of Outcome Variables

Table 18 shows the relationship between Convenience food consumed weekly and non-convenience food consumed weekly as well as monthly consumption in a contingency format. The p-values indicate significant associations. About half of the households purchase non-convenience food at least from one (1) or two (2) non-convenience shops on a weekly basis. The same pattern is noted with the monthly purchasing whereby above half of the purchases were made from one

(1) non-convenience source. There is also an observable upward trend on purchasing, food from Convenience food sources. Households tend to purchase convenience food from multiple sources (more than 2) on a weekly and monthly basis (Table 18).

Table 18: Crosstabulation of Convenience and Non-convenience Food Sources

Convenience ~Weekly	**Non-Convenience ~ Weekly**					***P-Value***
	0=Super-market	***1=Small shops***	***2= Open markets***	***3= Food grown by households in rural areas***	***Total***	
0 = Fast foods/Take away	3386	146	27	0	3559	<0.001
1 = Restaurants	90	119	84	1	294	
2 = Spaza/Tuck-shops	21	22	12	2	57	
3= Street seller/trader/hawker	3	1	2	2	8	
4 = Begging	0	1	2	1	4	
Total	3500	289	127	6	3922	
Convenience~ Monthly	**Non-Convenience ~ Monthly**					***P-Value***
	0=Super-market	***1=Small shops***	***2= Open markets***	***3= Food grown by households in rural areas***	***Total***	
0 = Fast foods/Take away	3175	217	112	7	3511	<0.001
1 = Restaurants	6	100	123	65	294	
2 = Spaza/ Tuck-shops	1	30	52	10	93	
3= Street seller/trader/hawker	0	3	6	8	17	
4= Begging	0	0	3	4	7	
Total	3182	350	296	94	3922	

4.3.3. Application of Bivariate Poisson Models on the Convenience and Non-Convenience Food Sources

4.3.3.1. Comparative Fit of Bivariate Poisson Models

Various Bivariate Poisson regression models were jointly fitted on Convenience and Non-Convenience data. AIC and BIC values are presented in Table 19 were used to select the best model. Firstly, the Bivariate Poisson model was fit with constant only for both the untruncated

and the right truncated models. Secondly, the Bivariate Poisson model was fit with covariates for both models. The AIC of 3646.976 Untruncated Bivariate Poisson (Full model) on a weekly basis was the least among all the fitted models and thus the full model fits the data better.

Table 19: Summary of the Fitted Bivariate Poisson Regression Models

Frequency	Model	2x Log Likelihood	Akaike Information Criterion (AIC)	Bayesian Information Criterion (BIC)
Weekly	Untruncated Bivariate Poisson (Constant only)	-1960.897	3925.795	3938.343
	Untruncated Bivariate Poisson (Full model)	-1787.488	3646.976	3870.246
Monthly	Untruncated Bivariate Poisson (Constant only)	-2222.053	4448.106	4460.655
	Untruncated Bivariate Poisson (Full model)	-2011.454	4094.909	4318.178

4.3.3.2. Weekly utilization of food sources: Untruncated Bivariate Poisson Regression

Firstly, the Bivariate Poisson model was fit with constant only as depicted in Table 20. The output shows the detail model statistics (AIC, BIC, etc.,) and parameter estimates (coefficients, t-value, p-value, adjusted S.E, and adjusted p-values). The AIC of 3925.8 in the reduced model is greater than the AIC of 3908. 5 in the Full model, thus the full model fits the data better.

Table 20: Fit for Bivariate Poisson Model (marginal/conditional): Constant only (reduced model)

Variable name	Coeff.	S.E	t.value	p.value	Adj.S.E	Adj. p.value
Y1: Constant	-2.170	0.047	-45.921	<0.001	0.054	<0.001
Y2: Constant	0.225	0.042	5.328	<0.001	0.0494	<0.001

Note. Loglik. = −1960.897, AIC = 3925.795, AICC = 3925.8, BIC = 3938.343, Deviance = 3087.749, P-1=1.35, P-2=1.37

The results of the fit of bivariate Poisson model are shown in Table 21 and 22 for both unadjusted and adjusted for over- or under-dispersion. It further provided the detail models statistics (e.g.,

AIC, BIC, etc) and parameter estimates showing the coefficients, standard error, t-value, p-value adjusted standard error and adjusted p-values. Here we model two possibly correlated dependent variables: (1) Convenience foods (2) non-Convenience foods. The Bivariate Poisson regression models shows that the variables education (secondary), age (all categories) and Income (20,000-49,999) are important determinants of both convenience and non-convenience food groups. The positive coefficients of education (none, primary), work (self-employed), sex (male), marital status (living together) shows a higher association on the convenience food sources.

Table 21: Fit of Bivariate Poisson Model (marginal/conditional) for both unadjusted and adjusted, for over- or under-dispersion (Full model)

Variable Names	**Coefficients (Coeff)**	**Standard Error (s.e)**	**t.value**	**p.value**	**Adj.s.e**	**Adj.p.value**
Convenience: Constant	-2.218	0.617	-3.596	0.000	0.719	0.002
Education: None	0.064	0.239	0.269	0.788	0.278	0.818
Education: Primary	0.014	0.219	0.066	0.948	0.255	0.955
Education: Secondary	-0.355	0.207	-1.714	0.087	0.241	0.141
Education: Tertiary (reference)						
Work: Self-employed	0.113	0.220	0.514	0.607	0.257	0.659
Work: Formal employed	-0.071	0.148	-0.478	0.633	0.172	0.682
***Work: Unemployed* (*reference*)**						
Sex: Male	0.108	0.100	1.078	0.281	0.117	0.355
Sex: female (reference)						
Marital: Unmarried	-0.102	0.604	-0.169	0.866	0.703	0.885
Marital: Married	-0.134	0.609	-0.219	0.826	0.710	0.851
Marital: Living together	0.674	0.605	1.114	0.265	0.705	0.339
***Marital: Widowed* (*reference*)**						
Age: <19	0.134	0.223	0.601	0.548	0.260	0.606
Age: 20-29	0.336	0.221	1.522	0.128	0.258	0.192
Age: 30-39	0.067	0.232	0.290	0.772	0.270	0.804
Age: 40-49	-0.042	0.238	-0.176	0.860	0.278	0.880
>50 (reference)						
Income: 2500-4999	-0.007	0.272	-0.025	0.980	0.317	0.983
Income: 5000-9999	-0.421	0.395	-1.066	0.286	0.461	0.360
Income: 10000-19999	-0.250	0.525	-0.477	0.634	0.611	0.683
Income: 20000-49999	-0.771	1.016	-0.759	0.448	1.184	0.515
Income: 0-2499 (reference)						

Table 22: Fit for Bivariate Poisson Model (marginal/conditional) for both unadjusted and adjusted, for over- or under-dispersion (Full model)…… Cont.

Variable Names	Coeff.	s.e	t.value	p.value	Adj.s.e	Adj.p.value
Non-convenience: Constant	-1.387	1.022	-1.358	0.175	1.205	0.250
Education: None	0.276	0.235	1.173	0.241	0.277	0.320
Education_ Primary	0.215	0.219	0.981	0.327	0.258	0.406
Education_ Secondary	0.422	0.211	2.002	0.045	0.249	0.090
Education_ Tertiary (reference)						
Work: Self-employed	0.227	0.195	1.165	0.244	0.230	0.324
Work: Formal employed	-0.087	0.137	-0.636	0.525	0.161	0.590
***Work: Unemployed* (*reference*)**						
Sex: Male	0.162	0.096	1.691	0.091	0.113	0.152
Sex: female (reference)						
Marital: Unmarried	0.408	1.037	0.393	0.694	1.224	0.739
Marital: Married	0.843	1.027	0.821	0.412	1.212	0.487
Marital: Living together	0.343	1.037	0.330	0.741	1.223	0.779
***Marital: Widowed* (*reference*)**						
Age: <19	0.854	0.280	3.051	0.002	0.330	0.010
Age: 20-29	0.781	0.272	2.869	0.004	0.321	0.015
Age: 30-39	0.926	0.263	3.523	0.000	0.310	0.003
Age: 40-49	0.662	0.258	2.561	0.010	0.305	0.030
>50 (reference)						
Income: 2500-4999	0.005	0.239	0.019	0.985	0.282	0.987
Income: 5000-9999	0.214	0.330	0.647	0.517	0.389	0.583
Income: 10000-19999	0.637	0.372	1.709	0.087	0.439	0.147
Income: 20000-49999	1.568	0.658	2.383	0.017	0.776	0.043
Income: 0-2499 (reference)						

Note. Loglik. = −1787.488, AIC = 3646.976, AICC = 3648.446, BIC = 3870.246, Deviance = 2805.341, P-1=1.36, P-2=1.39

4.4. Discussion

Several models exist for different Count data types. It is critical to know the properties and assumptions of different models. Bivariate Poisson model are appropriate for modelling paired count data exhibiting correlation (Karlis & Ntzoufras, 2005) . In this study, we used the dataset of

Windhoek AFSUN 2016 Household dataset to relate the Convenience and Non-Convenience Food Consumption both on a weekly and monthly base.

Convenience food often implies a lack of effort or concern, whether by choice or necessity. Highly processed food production and consumption are steadily increasing in both high-income and lower-income countries (Pan American Health Organization (PAHO), 2015). Parallel to this, the prevalence of obesity and other diet-related chronic non-communicable diseases (NCDs), such as type II diabetes, hypertension, and some common cancers, is increasing worldwide (Lim, et al., 2012). This study found that the households consume convenient food more often on a weekly basis and tend to utilize multiple convenient sources. On the other hand, the study revealed that households did not visit non-convenient food sources as often on a weekly basis and rather purchased their convenient foods monthly. According to Martinez Steel, Popkin, Swinburn, & Monteiro, (2017), the poor nutritional quality of ultra-processed foods coupled with their high availability, low cost, and aggressive marketing, which result in excessive consumption, can lead to obesity and other chronic diet related NCDs.

This chapter employed both the Untruncated and the Right-Truncated Bivariate models. The Bivariate Poisson models were further expanded to allow for covariates, for both the Univariate Poisson Regression with constants only and the Untruncated Poisson Regression full model. The parameter estimates confirmed that the variables age, marital status and educational level had an effect on the convenience food consumption. Hwang & Choe (2016) explained that households headed by younger, more educated, and time constrained managers were more likely to buy prepared meals. Employment creates time constraints from both the time spent working and the time spent commuting. These time constraints shift consumer demand from grocery store foods to

restaurant meals. The shift to full-service restaurants is most notable when all adults in the household are employed (Rahkovsky, Jo, & Carlson, 2018). The model statistics, particularly the AICs were used to compare and select the best fitted model. The Untruncated models specifically the full model proved to fit the data best compared to the right truncated model.

4.5. Conclusions

There has been tremendous growth and demand of the convenience food industry recently. Traditional meals and meals prepared at home are replaced by, often, unplanned food purchases from street corners, take-aways or restaurants. Convenience foods are described to be cheap and easy to prepare but the health benefits are questionable. The aim of this study was to apply bivariate count modelling approaches in analysing convenience and non-convenience consumption of food preference in Windhoek households. This study used the frequency of purchasing food from Convenience food Sources and Non-Convenience food sources variables from the AFSUN Windhoek dataset, 2016. In order to model frequency of occurrence/ count data that are correlated and needs to be jointly estimated, this study employed bivariate Poisson regression models, both un-truncated and right truncated. Although the consumption of food on a monthly basis was more on the non-convenience foods, the purchases of Convenience was frequent on a weekly basis and in multiple food sources. Convenience foods are mostly highly processed and of poor nutritional quality and can lead to a higher prevalence of NCDs. The untruncated models fit the data best. In conclusion, the model proved that the variables age, marital status, educational level of head of household and work status influenced the choices of food a household makes.

4.6. Acknowledgements

The authors had support from the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government. The data used in this study is from the AFSUN-HCP Household Food Security Baseline Survey (2016) for Windhoek and the usage of this dataset highly assisted the authors to achieve the objectives of this study.

CHAPTER 5: COPULA JOINT MODELLING OF FOOD INSECURITY INDICATORS WITH APPLICATION TO FOOD INSECURITY PREVALENCE (FIP), HOUSEHOLD DIETARY DIVERSITY SCORE (HDDS) AND MONTHS OF INADEQUATE HOUSEHOLD FOOD PROVISIONING (MIHFP)

Laina Mbongo[3]*, Lillian Pazvakawambwa, Lawrence Kazembe
Department of Computing, Mathematics and Statistical Sciences,
University of Namibia, Windhoek, Namibia

Abstract

Food insecurity is expressed using various indicators to measure availability, access, utilization and stability. Some of the indicators used are household food insecurity prevalence (HFIP), household dietary diversity score (HDDS) and months of inadequate household food provisioning (MIHFP). These measures are often assumed to be independent, since they capture different spectrums of food insecurity. However, these are correlated to each other, and their dependence has rarely been analyzed. This study used generalized joint regression models through copulas to estimate the relationship between food security outcomes/indicators and exposure variables. Both Bernoulli and Poisson marginals were assumed to quantify both binary and count response variables. We further explored partial observability and sample selection in the outcomes. A national cross-sectional survey, NHIES, of 2015/2016 was used in this analysis. The results indicated that both the Frank copula and bivariate normal copula fitted the data better of establishing the relationship between HFIP and HDDS (AIC=2287.296), and between HFIP and MIHFP (AIC=2072.708) respectively. The partial observability and sample selection analysis to account for measurement errors indicated that there was no statistically significant relationship

[3]* Corresponding Author Email: inambongo@gmail.com

between the food insecurity indicators and the exposure variables. The chapter thus concluded that copula approaches provide an advantage of analyzing jointly two outcomes in order to test for significant relationships between high-level hierarchical effects (e.g., random effects). Specifically, the bivariate normal and the frank copula were found to fit the data best. One unique feature of the Gaussian Copula is that it does not allow for a different dependence structure between the outcomes while the frank copula does not have tail dependence and it can model both positive and negative dependencies as the normal copula.

Keywords: Copulas, Sample Selection, Partial observability, Household food insecurity prevalence (HFIP), Household dietary diversity score (HDDS), Months of inadequate household food provisioning (MIHFP)

5.1. Introduction

Food security (FS), according to FAO (2002) exists when all people, at all times, have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy lifestyle. Food security thus encompasses four dimensions namely: (1) food availability which addresses the "supply side" of food security and is determined by the level of food production, stock levels and net trade; (2) food accessibility (economic and physical), an adequate supply of food at national or international level does not itself guarantee household level food security; (3) utilization, which is commonly understood as the way the body makes the most of various nutrients in the food; (4) stability: Even if food intake is adequate today, one is still considered to be food insecure if there is inadequate access to food on a periodic basis, risking a deterioration of your nutritional status (FAO, 2010).

A variety of food security measures have been proposed to capture the four components above. These aim to capture the extent of food insecurity at individual and household level. Foremost is the household food insecurity prevalence (HFIP), a categorical measure that classifies each household into either food secure, mildly food insecure, moderately food insecure or severely food insecure. Households are categorized as increasingly food insecure as they respond affirmatively to more severe conditions and/or experience those conditions more frequently (Coates, Swindale, & Bilinsky, 2007). Measures of household dietary diversity (HDD) tend to be of two types: those based on whether an individual food is consumed or not and those that are based on whether any food from a particular group is consumed. According to Coates, Swindale, & Bilinsky (2007), the resource available to the household and the management and availability of these resources throughout the year defines food access, hence the need to estimate the proportion of households

with an inadequate food supply in a month. This is considered as months of inadequate household food provisioning (MIHFP).

Although the definition of food security is clear, measurements of the different dimensions of food security are rare. Modelling of food insecurity, dietary diversity and months of inadequate food provisioning has often been applied independently at individual and household level. The main question of interest is, what are the chances that households or individuals that are food insecure are the same households that lack diversity in their diets and further experienced inadequate food provisioning throughout the year. The analysis of interdependence among two or more FS outcomes will help us to see the overall picture among outcomes and their correlations. Joint analysis has several advantages including avoiding multiple tests, increased power, better control of Type I error rates and efficiency handling of missing data (Leon & Wu, 2011).

According to Nieman (2015), the proper implementation of strategic probits and logits, however, is often made impossible by the outcome- rather than actor-specific structure of available data. While there are data on the aggregated outcomes of an interaction, there is no record of each player's actions at each of the interaction stages. During the analysis of observational data, it is often difficult to have data available for each actor at each information set of the game but instead the data is only available for the outcome of an interaction, with little to no data on the individual actions that led to the observed outcome. This translates that observational data such as food insecurity and dietary diversity are only partially observed. Traditional logistic and probit models often ignore the underlying partial observability problem, that might potentially lead to incorrect inferences.

The importance of dealing with these challenges motivated this study to employ alternative strategies that provide great flexibility in joint modeling of multimodal data. When there is an association between the two outcomes, a joint model will provide interesting and improved results than modelling the responses separately. The joint models significantly improve median log-loss and absolute residuals of cross-validation predictions (Broatch & Karl, 2017). Additionally, the joint models provide the ability to test for significant relationships between high-level hierarchical effects (e.g., random team effects) since significant predictions for outcomes at individual level may not be important at the group level.

Survey data are sometimes affected by systematic non-participation (Marra & Radice, 2017). This can occur through various ways including directly declining to participate in the study. If individuals are selected into (or out of) the sample based on a combination of observed and unobserved characteristics then models that ignore such a mechanism will most likely yield estimates which are not representative of the population of interest. The bias arising from ignoring such systematic non-participation is known as non-random sample selection bias. Another bias arises through partial observability. Partial observability typically occurs when two decisions are made to jointly determine an outcome. By jointly determining the outcome, one might not be able to observe the specific responses of the two decisions but can only observe the joint outcome. The unobserved specific responses often lead to partial observability biasness. The bivariate Probit with partial observability acknowledges the biasness by assuming that the model which determines the observed outcome is a bivariate Probit in which only one of the four outcomes is observed (Marra & Radice, 2017).

The Copula approach is defined as a useful method for deriving joint distributions. The approach relates an arbitrary joint distribution to its corresponding univariate marginal distribution via copula (Skalar (1959) as cited in Kazembe (2016). Copulas have been applied in many applications of statistics such as in insurance, econometrics, medicine, marketing, spatial, time series and even sports (Perrone and Muller, 2016). Copula is a multivariate dependence structure for joint distribution of random variables that are parted from the marginal distribution of individual random variables (Zimmer & Trivedi, 2006). Copulas first link the marginal distribution together to form the joint distribution and then define the nonparametric measures of dependence of pairs of random variables.

In this chapter, we explored joint modelling of HFIAP, HDDS and MIHFP as joint of binary and count variables using copulas. To address shortcomings in traditional logistic and Probit models, we further conducted a bivariate Probit model with partial observability and sample selection to estimate HFIAP and HDDS, as well as HFIAP with MIHFP jointly.

5.2. Materials and Methods

5.2.1. Data

The study used cross-sectional survey data of the Namibian Household and Income Expenditure (NHIES) of 2015/2016. In order to be comparable with standards recommended for Africa by FAO, food groups in the NHIES 2015/2016 were re-grouped and re-arranged in order to make up the 12 food groups for the analysis of HFIP, HDDS and MIHFP. Statistical package R Version 3.6 was used to compute joint modelling of copulas. Three outcome variables were used in this chapter, namely Household food insecurity prevalence, household dietary diversity score, and months on inadequate household food provisioning.

5.2.2. Joint Modeling (JM)

Joint modelling has been defined according to the type of data used. This chapter adopted the definition of Marra & Radice (2017). Let us assume that there are two binary random variables (Y_{i1}, Y_{i2}) , , for $i = 1, \ldots, n$, where n represents the sample size. The probability of event $(Y_{i1} = 1, Y_{i2} = 1)$ can be defined as:

$$p_{11i} = P(Y_{i1} = 1, Y_{i2} = 1) = C(P(Y_{i1} = 1), P(Y_{i2} = 1); \theta_i), \qquad (81)$$

Where $P(Y_{ij} = 1) = 1 - F_j(-\eta_{ji})\ for\ j = 1,2, F_j\ (\cdot)$ is the cumulative distribution function (cdf) of a standardized univariate distribution (in this case Gaussian, logistic or Gumbel), $\eta_{ji} \in n_{ij} \epsilon \mathbb{R}$ is an additive predictor, C is a two-place copula function and θ_i is an association parameter measuring the dependence between the two random variables.

The marginal c.d.f.s in this model are conditioned on covariates through η_{1i} and η_{2i}, but for notational convenience they are suppressed when expressing them. The dependence parameter is provided as a function of an additive predictor because, for instance, the strength and direction of the relationship between the two marginals may differ between sets of observations. That is, $\theta_i = m(\eta_{ci})$, where m is a one-to-one transformation which ensures that θ_i lies in its range.

5.2.3. Parameter Estimation

The model specification allows for a high degree of flexibility in modeling covariate effects. If an unpenalized approach is employed to estimate the model's parameters, then over-fitting is the likely consequence. To prevent this, Marra & Radice (2017) maximized $\ell_p(\delta) = \ell(\delta) - \frac{1}{2}\delta^T S \delta$, where ℓ_p is the penalized model's log-likelihood, $\delta^T = (\beta_1^T, \beta_2^T, \beta_c^T)$ and $S = diag(D_1, D_2, D_c)$.

The smoothing parameter vectors are collected in the overall vector $= (\lambda, \lambda_2^T, \lambda_c^T)$. Practically, it is advised that estimation of δ and λ should be obtained by using a stable and efficient trust region algorithm that is based on first and second order analytical derivative information, with integrated automatic multiple smoothing parameter selection (Marra & Radice, 2017).

5.2.4. Bivariate Binary Model with Non-random Sample Selection

According to Marra and Radice (2017), non-random sample selection occurs when individuals select themselves into (or out of) the sample based on a combination of observed and unobserved characteristics. Marra and Radice (2017) further noted that models that fail to take into account such a systematic selection could produce results that are unrepresentative of the population of interest. By adopting a two-equation structural latent variable framework where one equation defines the selection process (Y_{i1}) and the other describes the outcome Y_{i2}, a bivariate binary selection model may be used to address this problem and correct for non-random sample selection. (Y_{i1}) indicates whether an individual is selected into the sample whereas (Y_{i2}) is the outcome which is observed only if the individual is selected. In the same vein, to the endogenous model, the errors of the two equations are expected to follow a bivariate distribution with association parameter θ_i. In this case, the first additive looks like (Marra & Radice, 2017):

$$n_2 = \beta_{20} 1_{ns} + Z_{21}\beta_{21} + \cdots + Z_{2k2}\beta_{2k2} = Z_2\beta_{2,} \quad (82)$$

$$n_c = \beta_{c0} 1_{ns} + Z_{c1}\beta_{c1} + \cdots + Z_{ckc}\beta_{ckc} = Z_c\beta_{c,} \quad (83)$$

where 1_{ns} is an n_s-dimensional vector made up of ones corresponding to the selected observations, and Z_2 and Z_c have n_s rows. The log-likelihood function of the sample is:

$$\ell = \sum_{i=1}^{n} \{` I_{11i}\, log(p_{11i}) + I_{10i}\, log(P_{10i}) + (1 - y_{i1})\, log(p_{oi})\} \quad (84)$$

, where $p_{oi} = F1(-\eta_{1i})$.

5.2.5. Bivariate Probit Model with Partial Observability

The definition of partial observability in this section is derived from Marra & Radice (2017). The model tackles a problem in which an observed binary outcome reflects the joint realization of two unobserved binary outcomes. Therefore, the joint event $(Y_{i1} = 1, Y_{i2} = 1)$ has probability p_{11i} whereas all the other events have probability $1 - p_{11i}$.

The second predictor is defined as:

$$n_2 = \beta_{20} 1_n + Z_{21}\beta_{21} + \cdots + Z_{2k2}\beta_{2k2} \quad (85)$$

The log-likelihood function can be written as:

$$\ell = \sum_{i=1}^{n} \{ I_{11i}\, log(p_{11i}) + (1 - I_{11i})\, log(1 - p_{11i}) \} \quad (86)$$

Quantities of interest include estimates for p_{11i} and the impacts the covariates have on these probabilities. Note that this model is defined using Gaussian margins and a Gaussian copula.

5.2.6. The Copula Theory

The copula theory is used to determine the joint distribution of two variables and three variables in order to find the interdependence structure among the food security metrics. The copula theory was introduced by Sklar in 1959. It provides the opportunity to combine several single-variable distributions in various families of one, two, or multivariable distributions considering the interdependence of the variables (Mesbahzadeh, et al. 2019). According to Mesbahzadeh et al., (2019), one of the most important advantages of using copulas functions is that the structure of dependency between variables can be defined even if marginal distributions are different, meaning that in order to define a joint distribution function having equal marginal functions for each variable is not necessary.

5.2.7. Copula Functions

Copula functions include a variety of families such as Elliptical (t copula, Normal), Archimedean (Gumbel, Clayton, Frank, Ali-Mikhail-Haq), Extreme value (Husler-Reiss, Galambos, Tawn, and t-EV,Gumbel) and other families, namely, Plackett and Farlie-Gumbel-Morgensterm . Families of Archimedean and Elliptical are mostly considered (Mesbahzadeh, et al. 2019). In this chapter, we used the commonly used bivariate copulas. Table 23 shows a brief description of some copula functions:

Table 23: Copula families (Trivedi and Zimmer, 2005)

	Copula type	**Joint CDF**	$\boldsymbol{\theta}$	**Kendall $\boldsymbol{\tau}$**
Archimedean family	Frank	$C(\mu, v;\ \theta) = 1 /\theta \ln[1 + \frac{(e^{-\theta\mu} - 1)(e^{-\theta v} - 1)}{e^{-\theta}} - 1]$	$R/\{0\}$	$1 - \left(\frac{4}{\theta}\right)(1 - D_1(\theta))$ $D_k(x) = k/x^k \int_0^x t^k /(\exp(t) - 1)dt$
	Rotated Joe	$1 - [1 - \prod_{i=1}^{m}(1 - (1 - \mu_i)^{\theta})]^{1/\theta}$	$(-\infty, -1)$	$-1 - 4\int x\, log(x)(1 - x)^{-\frac{2(1+\theta)}{\theta}}.dx$
	Rotated Gumbel	$C(\mu, v, \theta) = \mu + v - 1 + c(1 - \mu, 1 - v)$	$(-\infty, -1)$	$-1 - (1/\theta)$
	Rotated Clayton	$C(\mu, v, \theta) = \mu + v - 1 + c(1 - \mu, 1 - v)$	$(-\infty, 0)$	$(\theta/(2 - \theta)$
Elliptical Family	Gaussian	$C(\mu, v) = \int_0^{\mu} \Phi(\Phi^{-1}(v) - pxy\Phi^{-1}(t) /\sqrt{1 - p^2xy)}dt$	$(-1, +1)$	$(2/\pi)\ arcsin(\theta)$

5.2.8. Estimation of Parameters of Copula Function

Both parametric and nonparametric methods are used to estimate the parameters of copula function. In the parametric method, the relationship between generator function of each copula and Kendall coefficient Equation (87) is used (Mesbahzadeh, et al. 2019).

$$\tau = \frac{(c-d)}{\binom{n}{2}} \qquad (87)$$

In this equation, *c* and *d* are the number of pairs of concordant and discordant variables and *n* is number of observations. Two pairs of variables (X_i, Y_i) and (X_j, Y_j) are concurring if $X_j > X_i$ and $Y_j > Y_i$ or $X_i > Y_j$ and $Y_i > Y_j$. Alternatively, if $(X_i - X_j)\ (Y_i - Y_j) > 0$, variables are concordant, and if $(X_i - X_i)\ (Y_i - Y_j) < 0$, variables are discordant. In the parametric method, using the maximum log-likelihood function Equation (88), parameter of θ is estimated (Mesbahzadeh, et al. 2019).

$$L(\theta) = \sum_{k=1}^{n} log[c_\theta\{F_1(x_{1k}), \dots, F_p(X_{pk})\}] \qquad (88)$$

, where c_θ is the copula density function; F is the marginal distribution function; and $x_{1k}, x_{2k}, \dots, x_{pk}\ k = 1, \dots, n$ are the dependent random variables.

Log-likelihood function estimates parameter of θ using density copula function. If dependent random variables are as $x_{1k}, x_{2k} \dots, x_{pk}\ (k = 1, \dots, n)$ with copula function of $F_\theta\left(x_{1k,\dots,x_{pk}}\right) = C_\theta(F_1(X_{1k}), \dots, F_p(X_{pk}))$.

5.2.9. Goodness-of-Fit test for Copula Function

For selecting the best copula function, value of joint empirical probability of the variables were calculated through empirical copula in Equation (89) and then is compared with the values resulted from other copula functions (Archimedean and Elliptical families) (Mesbahzadeh, et al. 2019).

$$C_n(\mu, v) = \frac{1}{n}\sum_{t=1}^{n} 1(\mu_t < \mu, V_t < v) \quad (89)$$

Whereby μ and v are the empirical probabilities of the two variables.

To compare empirical copula with each copula functions, Normalized Root Mean Square Error (NRMSE) and Nash–Sutcliffe coefficient were selected equations (90) and (91)).

$$NRMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n} \frac{(P_{ei}-P_i)^2}{(P_{ei},max-P_{ei},min}} \quad (90)$$

$$NSE = 1 - \frac{\sum_{n=1}^{N}(P_{ei}-P_i)^2}{(P_{ei}-\bar{P}_i)^2} \quad (91)$$

Where P_{ei} is the value of empirical copula and P_i is the value of the copula theory.

Additionally, two criteria, namely, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Equation (92) and Equation 93) and are used. Furthermore, in equation 92 and equation 93, k is the model parameter, n is the number of observations and L is the value of the maximum log-likelihood function.

$$AIC = 2k - 2\,ln(L) \quad (92)$$

$$BIC = 2n\,log\,L + k\,log(n) \quad (93)$$

5.3. Results

5.3.1. Food Security, Dietary Diversity, and Months of Inadequate Food Provisioning

The following analysis shows the relationship between food security and socio-household characteristics (Table 24).

The variable sex (male, P=0.022), marital status (married (living with spouse), P =0.18), Education (primary and secondary, P<0.001), work status (working full time, P= 0.008, not working, P=0.020), access to water (no piped water, P= 0.037) showed a statistically significant relationship with food insecurity prevalence. Additionally, variables such as marital status (single, P=0.033), education (No education, P= 0.020, secondary, P=0.024), work status (working full time, P = 0.002 and not working-looking, P=0.008), Tenure status (owner/family, P= 0.046), water (no piped water, P= 0.012) and access to a flushing toilet (no toilet, P=0.012) had a statistically significant relationship with HDDS. In terms of MIHFP, variables such as marital status (not married but living with partner, P=0.004 and going steady (in a relationship), P=0.004, education (no education, P=<0.001 and secondary, P=0.001) and household structure (male centered, P= 0.031 and nuclear, P=0.011) were significant at 5% (Table 24).

Table 24: Association between HFIP, HDDS MIHFP and socio-household characteristics

	HFIP		**HDDS**		**MIHFP**	
	Pearson's Chi-Square					
Sex:	**Value**	**P-Value**	**Value**	**P-Value**	**Value**	**P-Value**
Male	5.239	**0.022**	0.066	0.797	1.345	0.246
Female	*Reference*					
Marital Status:						
Married (living with spouse)	5.552	**0.018**	3.732	0.053	1.959	0.162
Married (not living with spouse)	0.091	0.763	0.390	0.532	0.986	0.321
Not married (living with partner)	2.378	0.123	0.864	0.353	8.406	**0.004**
Going steady (in a relationship)	0.559	0.445	2.913	0.088	8.291	**0.004**
Single (not in a relationship)	0.330	0.566	4.536	**0.033**	0.791	0.374
Divorced separated	0.018	0.894	0.040	0.842	0.691	0.406
Widower/Widow	*Reference*					
Education:						
None	2.005	0.157	5.441	**0.020**	15.489	**<0.001**
Primary	35.588	**<0.001**	3.099	0.078	0.155	0.694
Secondary	33.010	**<0.001**	5.105	**0.024**	7.255	**0.007**
Tertiary	*Reference*					
Work status:						
Working full-time	6.958	**0.008**	9.454	**0.002**	3.723	0.054
Working part-time/casual work	0.249	0.618	1.000	0.317	0.054	0.815
Not working - looking	5.445	**0.020**	7.101	**0.008**	3.117	0.077
Not working - not looking	*Reference*					
Tenure Status:						
Owner/Family	0.428	0.513	3.141	**0.046**	0.830	0.362
Tenant/Lodger	0.047	0.828	2.348	0.125	0.613	0.434
Tied accommodation	*Reference*					
Household Structure:						
Female centered	0.200	0.655	0.207	0.649	0.056	0.814
Male centered	0.327	0.567	0.231	0.631	4.632	**0.031**
Nuclear	0.685	0.408	0.280	0.596	0.317	**0.011**
Extended	1.593	0.672	0.327	0.877	-1.942	0.573
Under 18 headed household	*Reference*					
No piped water - private	4.372	**0.037**	6.318	**0.012**	0.205	0.651
Piped Water - Private	*Reference*					
No electricity	1.927	0.165	0.037	0.847	1.323	0.250
Electricity available	*Reference*					
No toilet	0.250	0.617	6.332	**0.012**	0.762	0.383
Toilet available	*Reference*					

5.3.2. Logistic and Poisson Regression Models: HFIP, HDDS and MIHFP

Table 25 and Table 26 provides a summary of the logistic regression and Poisson regression models. Predictable variables such as education and accessibility to water influenced the food security level of a household. The household Dietary Diversity is affected by the educational level of the head of Household as well as accessibility to amenities such as electricity and toilet facilities. Months on Inadequate food Provisioning (MIHFP) is another indicator to measure the food security of a household. MIHFP was influenced by various factors including marital status (specifically by those that are not married but living with partners and those that are going steady in a relationship), Educational level, tenure status and the household structure. All these predictors were significant at 5% level (p-value<0.05).

Table 25: Modelling of FIP, HDDS and MIHFP

	Logistic Regression Model (HFIP)		**Poisson Regression Model (HDDS)**		**Poisson Regression Model (MIHFP)**	
Coefficients	**Std. Err**	**P-Value**	**Std. Err**	**P-Value**	**Std. Err**	**P-Value**
(Intercept)	623.74	0.978	0.617	0.004	1.942	0.011
Sex:					0.030	**0.002**
Male	0.342	0.087	0.064	0.812		
Female	*Reference*					
Marital Status:					-0.054	0.809
Married (living with spouse)	1.195	0.152	0.215	0.173		
Married (not living with spouse)	1.318	0.771	0.238	0.188	-0.490	0.090
Not married (living with partner)	1.200	0.348	0.216	0.353	0.462	**0.030**
Going steady (in a relationship)	1.194	0.579	0.214	0.269	-0.545	**0.013**
Single (not in a relationship)	1.180	0.411	0.211	0.517	0.012	0.952
Divorced / separated	1.755	0.677	0.349	0.924	0.248	0.466
Widower/Widow	*Reference*					
Education:						
None	0.905	**0.001**	0.141	**0.001**	0.798	**0.006**
Primary	0.805	**0.001**	0.110	**0.002**	0.388	0.172
Secondary	0.815	0.072	0.113	0.045	-0.002	0.928
Tertiary	*Reference*					

Table 26: Modelling of FIP, HDDS and MIHFP ….cont.

	Logistic Regression Model (HFIP)		**Poisson Regression Model (HDDS)**		**Poisson Regression Model (MIHFP)**	
Employment						
Working full-time	0.496	0.271	0.092	0.269	-0.223	0.090
Working part-time/casual work	0.513	0.210	0.110	0.804	-0.154	0.237
Not working - looking	0.506	0.964	0.095	0.353	-0.006	0.624
Not working - not looking	*Reference*					
Tenure Status:						
Owner/Family	623.733	0.979	0.422	0.813	1.340	**0.049**
Tenant/Lodger	623.734	0.977	0.431	0.945	1.244	0.074
Tied accommodation	*Reference*					
Household Structure:						
Female centered	1.588	0.438	0.325	0.519	-1.942	**0.011**
Male centered	1.585	0.634	0.325	0.697	-1.942	**0.011**
Nuclear	1.595	0.777	0.326	0.680	-1.942	**0.011**
Extended	1.593	0.672	0.327	0.877	-1.942	**0.011**
Under 18 headed household	*Reference*					
No piped water - private	0.426	**0.035**	0.083	0.305	-0.159	0.157
Piped Water - private	*Reference*					
No electricity	0.620	0.119	0.141	**0.076**	-0.351	0.089
Electricity available	*Reference*					
No toilet	1.317	0.933	0.207	**0.019**	-0.141	0.807
Toilet available	*Reference*					
AIC	**468.81**		**1781.6**		**1752.019**	
BIC	**559.1909**		**1871.9515**		**1838.909**	

5.3.3. Joint Modelling of Household Food Insecurity Prevalence (HFIP) and Household Dietary Diversity Score (HDDS)

Generalized Joint Regression model was conducted, and copula estimates were performed using binary-binary margins (probit) to estimate several copulas with endogenous treatment, where the bivariate distributions are chosen so that the dependence is allowed. This is mainly because the models based on the Gaussian and Frank Copulas suggest that the dependence between the outcomes is positive, thus implying copulas which allow for negative association when the data do not support this will be misleading (Marra & Radice, 2017). The AICs were used to determine

the best fitted model. According to Table 27, all the models are more or less equally good as their AICs did not differ much, however the Frank copula had the least AIC.

Table 27: AICs for copula models: FIP and HDDS

Family	**Df**	**AIC**
Bivariate Normal	65	2288.355
Frank	65	**2287.296**
Rotational Clayton	65	2288.158
Gumbel	65	2288.349

Table 28 shows that all the predictor variable estimates obtained for the Frank copula were not significant at 5%, thus indicating no existence of any positive association between the unstructured terms of the model equations.

Table 28: Estimates for Frank copula model (Margins: Bernoulli, Bernoulli)

Coefficients	**Estimate**		**Std. err**		**P-Value**	
	HFIP	**HDDS**	**HFIP**	**HDDS**	**HFIP**	**HDDS**
(Intercept)	7.587	2.408	7.144	7144.461	1.000	0.999
Sex:						
Male	-3.671	-4.211	5.435	5434.891	1.000	1.000
Female						
Marital Status:						
Married (Living with spouse)	0.7376	2.495	3.043	5434.891	1.000	1.000
Married not (living with spouse)	0.322	2.732	3.043	3042.879	1.000	1.000
Not married (living with partner)	0.381	1.561	3.043	3042.879	1.000	1.000
Going steady (in a relationship)	0.312	1.932	3.043	3042.879	1.000	1.000
Single (not in a relationship)	0.120	9.466	3.043	3042.879	1.000	1.000
Divorced / separated	3.043	-1.206	3.043	3042.879	1.000	1.000
Widower/Widow						
Education:						
None	-0.328	-1.724	3.972	3971.539	1.000	1.000
Primary	-1.399	-1.466	3.972	3971.539	1.000	1.000
Secondary	-1.202	-1.350	3.972	3971.539	1.000	1.000
Tertiary	*Reference*					
Working full – time	0.421	-2.964	3.972	3971.539	1.000	1.000
Working part-time/Casual	0.211	-4.213	3.972	3971.539	1.000	1.000
Not working - looking	0.260	-4.865	3.972	3971.539	1.000	1.000
Not working- not looking	*Reference*					
Tenure Status:						
Owner/Family	-6.031	8.254	4.537	4543.257	1.000	0.999
Tenant/Lodger	-6.225	-4.681	4.537	4543.257	1.000	0.999
Tied accommodation	*Reference*					
Household Structure:						
Female centered	0.301	7.240	3.575	3574.794	1.000	1.000
Male centered	0.027	-9.714	3.575	3574.794	0.999	1.000
Nuclear	-0.137	-2.192	3.575	3574.794	0.999	1.000
Extended	-0.005	-8.334	3.575	3574.794	1.000	1.000
Under 18 headed household's	*Reference*					
No piped water- private	-0.249	-2.351	5.435	5434.892	1.000	1.000
Piped Water – Private	*Reference*					
No Electricity	-0.694	1.628	5.435	5434.892	1.000	1.000
Electricity available	*Reference*					
No Toilet	-0.050	1.919	5.435	5434.892	1.000	1.000
Toilet Available	*Reference*					
AIC: 2287.296; BIC; 2542.719, Theta= 0.419(-0.342, 1.05), Tau = 0.0464 (-0.0379, 0.116)						

5.3.4. Joint modelling of Household Food Insecurity Prevalence and Months of Inadequate Household Food Provision (MIHFP)

The joint modelling of food insecurity prevalence and months of inadequate food provisioning using different copula models shows that the Bivariate Normal copula is the preferred model (lowest AIC).

Table 29: AICs for copula models: HFIP and MIHFP (margins = Bernoulli, Poisson)

Family	**Df**	**AIC**
Bivariate Normal	65	**2072.708**
Frank	65	2074.352
Gumbel	65	2108.451

The estimates for the Bivariate Normal copula independent variables proved to have no positive association at 0.05 significant level (app *P*-values>0.005, and Theta (-0.32(-0.417, -0.219).

5.3.5. Sample Selection and Partial Observability: Food Insecurity Prevalence and Dietary Diversity Score

Sample selection and Partial observability were conducted to observe specific household responses. Table 30 shows that the determinants variables were not significant at 5%, suggesting that there is no statistically significant relationship between HFIP, HDDS and the independent variables. Sex of head of household was found to have a statistically significant relationship with household food insecurity prevalence (P-value<0.05) (Table 31).

Table 30: Sample selection: Food Insecurity Prevalence (HFIP) and Household Dietary Diversity Score (HDDS) (margins= Bernoulli, Bernoulli)

	Estimate		**Std. err**		**P Values**	
Coefficients	**HFIP**	**HDDS**	**FIP**	**HDDS**	**FIP**	**HDDSx**
(Intercept)	-7.808	-1.473	7082.429	8192.000	0.999	1.000
Sex:	-0.158	-14.423	0.211	8192.000	0.454	0.999
Male	*Reference*					
Female						
Marital Status:						
Married (Living with spouse)	-0.734	-14.723	3148.404	8192.000	1.000	0.999
Married not (living with spouse)	-0.790	-14.632	3148.404	8192.000	1.000	0.999
Not married (living with partner)	-0.693	-14.641	3148.404	8192.000	1.000	0.999
Going steady (in a relationship)	-1.170	34.629	3148.404	8192.000	1.000	0.999
Single (not in a relationship,	-6.934	-14.638	3148.404	8192.000	1.000	0.999
Divorced / separated	-1.333	-14.692	3148.404	8192.000	1.000	0.997
Widower/Widow)	*Reference*					
Education:						
None	-6.675	-14.574	2524.724	8192.000	0.998	0.999
Primary	-6.295	-14.541	2524.724	8192.000	0.998	0.999
Secondary	-5.995	-14.206	2524.724	8192.000	0.998	0.999
Tertiary	*Reference*					
Working full – time	-6.227	-14.289	2524.724	8192.000	0.998	0.999
Working part-time/Casual	-6.409	-14.405	2524.724	8192.000	0.998	0.999
Not working - looking	-5.520	-14.330	2524.724	8192.000	0.998	0.999
Not working- not looking	*Reference*					
Tenure Status:						
Owner/Family	-0.645	-14.553	5205.962	8192.000	1.000	0.999
Tenant/Lodger	-5.632	40.998	7740.02	8192.000	0.999	0.999
Tied accommodation	*Reference*					
Female centered	6.173	8.795	3918.062	8192.000	0.999	0.999
Male centered	5.990	8.860	3918.062	8192.000	0.999	0.999
Nuclear	5.945	8.793	3918.062	8192.000	0.999	0.999
Extended	-0.777	-15.056	6819.300	8192.000	1.000	0.999
Under 18 headed household's	*Reference*					
No piped water- private	-42.093	-18.209	5414.556	8192.000	0.994	0.999
Piped Water – Private	*Reference*					
No Electricity	68.147	8.714	5414.556	8192.000	0.990	0.999
Electricity available	*Reference*					
No Toilet	-3.959	69.296	6817.927	8192.000	1.000	0.999
Toilet Available	*Reference*					
AIC: 473.073; BIC; 27.626, Theta= 100(87, 100), Tau = 0.961 (0.955, 0.961)						

Table 31: Partial Observability: HFIP and HDDS (margins= Bernoulli, Bernoulli)

	Estimate		**Std. err**		**P-Values**	
Coefficients	**HFIP**	**HDDS**	**HFIP**	**HDDS**	**HFIP**	**HDDS**
(Intercept)	0.044	87.082	66.736	59.9021	0.999	0.146
Sex:						
Male	0.484	-24.148	0.210	20.396	**0.021**	0.236
Female	*Reference*					
Marital Status:						
Married (Living with spouse)	-0.899	-17.342	32.724	20.338	0.978	0.395
Married not (living with spouse)	-0.976	-18.269	32.724	20.397	0.976	0.369
Not married (living with partner)	-0.714	-27.820	32.724	20.344	0.982	0.171
Going steady (in a relationship)	-1.340	-27.326	32.724	20.333	0.967	0.179
Single (not in a relationship)	-5.858	-28.334	32.724	8192.000	0.962	0.997
Divorced / separated	-1.431	-19.456	32.724	20.346	0.965	0.339
Widower/Widow	*Reference*					
Education:						
None	-5.588	-11.029	24.434	20.304	0.819	0.587
Primary	-5.063	-12.938	24.434	20.304	0.836	0.523
Secondary	-4.557	-11.750	24.434	20.304	0.852	0.563
Tertiary	*Reference*					
Working full – time	-5.181	-6.048	24.455	20.343	0.832	0.766
Working part-time/Casual	-5.320	-4.986	24.454	20.342	0.828	0.806
Not working - looking	-4.725	-7.736	24.454	20.206	0.847	0.703
Not working- not looking	*Reference*					
Tenure Status:						
Owner/Family	-0.605	-2.319	55.287	24.218	0.991	0.924
Tenant/Lodger	-4.201	-4.795	81.594	8192.000	0.959	0.999
Tied accommodation	*Reference*					
Household Structure:						
Female centered	5.030	-23.047	53.353	58.850	0.925	0.695
Male centered	4.676	-20.748	53.354	58.881	0.930	0.724
Nuclear	4.971	-20.148	53.354	58.845	0.926	0.732
Extended	-0.498	-18.254	91.977	939.496	0.994	0.984
Under 18 headed household's	*Reference*					
No piped water- private	0.019	0.600	53.585	51.438	0.999	0.990
Piped Water – Private						
	Reference					
No Electricity	15.060	0.927	43.674	62.046	0.730	0.988
Electricity available	*Reference*					
No Toilet	-2.810	4.488	102.710	940.904	0.978	0.996
Toilet Available	*Reference*					
AIC: 3093.055; BIC: 3347.609, Theta= 0.272 (0.258, 0.289), Tau = 0.175 (0.166, 0.187)						

5.3.6. Sample Selection and Partial Observability: Food Insecurity Prevalence and Months of Inadequate Food Provision

Table 32 and Table 33 shows results from sample selection and partial observability. Apart from Sex, all other determinants variables were not significant at 5%, suggesting that there is no statistically significant relationship.

Table 32: Sample Selection: FIP and MIHFP (margins= Bernoulli, Poisson)

	Estimate		**Std. err**		**P Values**	
Coefficients	**HFIP**	**MIHFP**	**HFIP**	**MIHFP**	**HFIP**	**MIHFP**
(Intercept)	-5.728	9.692	6.639	7.213	0.993	0.999
Male	-1.014	-2.485	2.738	9.693	0.711	**0.010**
Female	*Reference*					
Married (Living with spouse)	-1.690	-1.157	1.940	3.580	0.999	0.999
Married not (living with spouse)	-1.735	-3.383	1.940	3.580	0.999	0.999
Not married (living with partner)	-1.552	4.724	1.940	3.580	0.999	0.999
Going steady (in a relationship)	-2.151	3.294	1.940	3.580	0.999	0.999
Single (not in a relationship)	-6.446	-1.786	1.940	8.192	0.999	1.000
Divorced / separated	-2.211	6.423	1.940		0.997	0.997
Widower/Widow	*Reference*					
Education:						
None	-1.970	2.266	1.793	3.979	0.999	0.999
Primary	-1.704	5.795	1.793	3.979	0.999	0.999
Secondary	-7.878	7.974	1.793	3.979	0.999	0.999
Tertiary	*Reference*					
Working full – time	-8.081	3.578	1.721	3.979	0.999	0.999
Working part-time/Casual	-1.029	-3.057	1.721	3.979	0.999	0.999
Not working - looking	-3.480	5.031	1.721	3.979	0.999	0.999
Not working- not looking	*Reference*					
Owner/Family	7.217	3.053	2.648	5.457	0.978	0.999
Tenant/Lodger	6.849	1.441	2.565	8.192	0.978	0.999
Tied accommodation	*Reference*					
Female centered	6.864	2.179	2.565	3.979	0.978	0.999
Male centered	6.845	5.120	2.565	3.979	0.978	0.999
Nuclear	6.817	-2.316	2.565	3.979	0.978	0.999
Extended	6.277	3.609	2.664	8.192	0.981	0.999
Under 18 headed household's	*Reference*					
No piped water- private	1.725	6.291	3.792	5.457	1.000	0.999
Piped Water – Private						
No Electricity	-7.443	2.641	3.932	8.192	0.984	0.999
Electricity available	*Reference*					
No Toilet	-6.591	1.611	2.121	5.457	0.999	0.999
Toilet Available	*Reference*					
AIC: 678.045, BIC: 985.021, Theta= 0.535 (0.382, 0.65), Tau = 0.359 (0.25, 0.45)						

Table 33: Partial Observability: FIP and MIHFP (margins= Bernoulli, Poisson)

	Estimate		**Std. err**		**P-Values**	
Coefficients	**FIP**	**MIHFP**	**FIP**	**MIHFP**	**FIP**	**MIHFP**
(Intercept)	0.056	87.082	77.736	49.9021	0.999	0.146
Sex:						
Male	0.556	-24.148	0.223	20.396	0.432	0.236
Female	*Reference*					
Marital Status:						
Married (Living with spouse)	-0.899	-17.342	32.724	20.338	0.978	0.395
Married not (living with spouse)	-0.976	-18.269	32.724	20.397	0.976	0.369
Not married (living with partner)	-0.714	-27.820	32.724	20.344	0.982	0.171
Going steady (in a relationship)	-1.340	-27.326	32.724	20.333	0.967	0.179
Single (not in a relationship)	-5.858	-28.334	32.724	8192.00	0.962	0.997
Divorced / separated	-1.431	-19.456	32.724	20.346	0.965	0.339
Widower/Widow	*Reference*					
Education:						
None	-5.588	-11.029	24.434	20.304	0.819	0.587
Primary	-5.063	-12.938	24.434	20.304	0.836	0.523
Secondary	-4.557	-11.750	24.434	20.304	0.852	0.563
Tertiary	*Reference*					
Working full – time	-5.181	-6.048	24.455	20.343	0.832	0.766
Working part-time/Casual	-5.320	-4.986	24.454	20.342	0.828	0.806
Not working - looking	-4.725	-7.736	24.454	20.206	0.847	0.703
Not working- not looking	*Reference*					
Tenure Status:						
Owner/Family	-0.605	-2.319	55.287	24.218	0.991	0.924
Tenant/Lodger	-4.201	-4.795	81.594	8192.00	0.959	0.999
Tied accommodation				0		
	Reference					
Female centered	5.030	-23.047	53.353	58.850	0.925	0.695
Male centered	4.676	-20.748	53.354	58.881	0.930	0.724
Nuclear	4.971	-20.148	53.354	58.845	0.926	0.732
Extended	-0.498	-18.254	91.977	939.496	0.994	0.984
Under 18 headed household's	*Reference*					
No piped water- private	0.019	0.600	53.585	51.438	0.999	0.990
Piped Water – Private	*Reference*					
No Electricity	15.060	0.927	43.674	62.046	0.730	0.988
Electricity available	*Reference*					
No Toilet	-2.810	4.488	102.710	940.904	0.978	0.996
Toilet Available	*Reference*					
AIC: 4031.044, BIC: 4712.055, Theta= -0.32(-0.417, -0.219), Tau = -0.207 (-0.274, -0.14)						

5.4. Discussion

Various food security measurements exist to measure the extent of food insecurity both at individual and household level. This chapter particularly applied bivariate joint regression models using copulas (Bivariate Normal, Frank, Rotational Clayton, Gumbel) to model food insecurity prevalence, Household Dietary Diversity and Months of Inadequate Food provisioning. The Bivariate Poisson models are appropriate for modeling paired count data exhibiting correlation and require joint estimation (Karlis & Ntzoufras, 2005). Sample selection and partial observability are errors that arise during the collection of data. For example, the implementation of strategic models is often made impossible by the outcome-rather than actor-specific structure of available data: while there are data on the aggregated outcomes of an interaction, there will be no record of each player's actions at each of the interaction stage.

Food insecurity is a major problem in the country. About 63% of the population are food insecure. This means, this proportion of the country does not have physical and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy lifestyle at all times. Food security puts an emphasis on all the 4 dimensions to be met: Food availability, Food accessibility, Food utilization and Food stability (FAO, The State of Food Security in the world, 2002). Dietary diversity is very critical in measuring food security. This means that most households consume a monotonous diet that lacks variety of diets. Additionally, food accessibility is defined by the availability of resources to the households throughout the month. According to Nickanor (2014), most households did not have enough resources for food in the months of January. January Precedes the month of December, that is mostly referred to as the Festival Month. Most households have utilized their savings and bonuses on these social gatherings. Apart from that, during the month of January, households further have to capitalize on

other mandatory expenditures such as school uniforms and school fees and rural households investing in ploughing/ farming activities, as it is a rainy season. This leaves most households with little to spend on foods (Nickanor, 2014).

The results from the logistic and Poisson logistic regression models indicated that educational level of the head of household and accessibility to water influenced the food security level of a household. Other factors that influenced food security included marital status, tenure status and the household structures (female centered, male centered, Nuclear, Extended, under-18 headed households). Education improves food security more directly in two ways; firstly, by improving skills and income generating potentials, secondly, through greater employability opportunities and increased incomes from better employment (Ajieroh, 2009). The household structure also affects food insecurity of that house. Larger households tend to have a negative impact on individual caloric availability. The size of a household has a potential to directly affects is food insecurity level through its influence on consumption pattern (Nickanor, 2014).

This study utilized copula functions to jointly estimate the variables. A joint model provides improved results on modelling associations between two outcomes, rather than modelling them separately. It significantly improves the log-loss and absolute residuals of cross-validation predictions (Broatch & Karl, 2017). Frank copula fit the data better to estimate the relationship between food insecurity prevalence and household dietary diversity score. Bivariate Normal copula was the best to model an association between food insecurity prevalence and months of inadequate food provisioning. Sample selection and partial observability were conducted to determine the relationship between Food Insecurity Prevalence and Dietary diversity as well as between food insecurity prevalence and months of Inadequate household food provisioning. The

models found that, apart from sex, all other social-demographic variables were not significant at 5% indicating a non-relationship between the exposure and outcome variables.

5.5. Conclusions

Generalized Joint Regression Models are used to model jointly binary outcomes. The aim of this study was to jointly model food insecurity indicators with application to FIP, HDDS and MIHFP. Copula approaches relate an arbitrary joint distribution to its corresponding univariate marginal distributions. Copulas were applied in this study to investigate the relationship between household food insecurity prevalence (HFIP) (1. Food Secure 2. Food Insecure); household dietary diversity score (1. Low diversity 2. High diversity) and Months of Inadequate Household Food Provisioning (MIHFP) (Seasonal, persistent). Food insecurity in Namibia is high and less varied with a monotonous diet. Households were further found to be more food insecure during the month of January. Measurement errors were accounted for by the modelling of sample selection and partial observability.

The Copula approach is defined as a useful method for deriving joint distributions. The approach relates an arbitrary joint distribution to its corresponding univariate marginal distributions via Copula. Specifically, five (5) Copula families namely the Bivariate normal, Frank, Survival, Clayton, Gumbel and the Survival Gumble were used in this analysis. The Frank Copula was identified to fit the data between FIP and HDDS better while the Bivariate normal better fitted the data between FIP and MIHFP. Sample selection and Partial Observability were conducted to observe specific responses between the three indicators. The socio-demographic variables were all not significant at 5% indicating a non-relationship between the exposure and outcome variables.

5.6. Acknowledgements

The authors had support from the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

CHAPTER 6: MULTIPLE-INDICATOR, MULTIPLE CAUSE MODELLING TO EXAMINE THE RELATIONSHIP BETWEEN FOODS CONSUMED AND NON-COMMUNICABLE DISEASES

Laina Mbongo[4]*, Lawrence Kazembe, Lillian Pazvakawambwa
Department of Computing, Mathematics and Statistical Sciences,
University of Namibia, Windhoek, Namibia

Abstract

Non-Communicable diseases are commonly associated with the dietary patterns of an individual. quantifying the disease's burden over a household's or individual's health has been a topic of great interest to researchers as well as policymakers. Various measurement approaches of NCD's that account for different types of biasness is required to correctly identify explanatory variables. This chapter used Namibia Household and Income Expenditure (NHIES) survey of 2015/16 variables to examine relationships between NCDs and the type of foods consumed. Principal Component Analysis was used as a data reduction method to derive dietary patterns. Furthermore, this chapter applied a Multiple-Indicator, Multiple-Cause (MIMIC) model in which NCD's is dealt with as an unobserved construct or latent variable to be determined by its causes and indicators and to be estimated in a system of structural equations. SEM was used to assess the association between the prevalence of NCD's and food types consumed. Fruits, foods such as condiments/tea/coffee and potatoes, yams, cassava, or any foods made from roots and tubers accounted for majority of the variation. The SEM showed that food types such as local grains, meat and food made from oil or were found to be significant at 5% level.

Keywords: Structural Equation Models (SEM), Principal Component Analysis (PCA), Non-Communicable Diseases (NCD)

[4]* Corresponding Author Email: inambongo@gmail.com

6.1. Introduction

Non-communicable diseases (NCDs) kill about 41 million people each year, equivalent to 71% of all deaths globally. Each year, more than 15 million people die from a non-communicable disease between the ages of 30 and 69 years; 85% of these "premature" deaths occur in low- and middle-income countries (WHO, 2021). Non-communicable diseases, sometimes referred to chronic diseases, tend to be of long duration and are the result of a combination of genetic, physiological, environmental and behavioural factors. The main types of NCD are cardiovascular diseases (such as heart attacks and stroke), cancers, chronic respiratory diseases (such as chronic obstructive pulmonary disease and asthma) and diabetes (WHO, 2021).

The prevalence of multimorbidity is increasing worldwide. A systematic review in WHO Eastern Mediterranean countries in 2013 showed that the high mortality of NCDs is partially related to their multimorbidity. More than half of the adults with NCDs have multimorbidity or multiple concurrent morbid conditions, and not one single chronic disease (Khorrami, et al., 2020). The NCD's are mostly driven by forces that include rapid unplanned urbanization, globalization of unhealthy lifestyles and population ageing. Increased prevalence of obesity, increased consumption of poor-quality diets, and pervasive undernutrition are contributing to this epidemic (UNSCN, 2018).

Poor quality diets are found to be among the top 6 risk factors contributing to the global burden of disease (Global Pattern, 2016). According to Global Pattern (2016), the NCD burden is specifically associated with diets that are low in fruits and vegetables, high in sodium, low in nuts and seeds, low in whole grains, and low in seafood-derived omega-3 fatty acids. The type of dietary pattern followed can easily influence one's health and the risk of contracting a chronic illness. Three (3)

categories of dietary pattern analysis approaches exist, namely the theoretical methods, empirical methods and the hybrid methods.

During the past few decades, quantifying the disease's burden over the population's health has been a topic of great interest to researchers as well as policymakers. A great deal of research has been conducted in the developed world to quantify the disease burden (communicable and non-communicable) on the population's health (El-Saadani, Saleh, & Ibrahim, 2021). One type of approach is based on Multiple-Indicator, Multiple-Cause (MIMIC) models in which NCD's is dealt with as an unobserved construct or latent variable to be determined by its causes and indicators and to be estimated in a system of structural equations. Measurement of Multiple-Indicator, Multiple-Causes variables such as prevalence of Non-Communicable Diseases or number of food groups consumed in a household can be a challenge to compute. In most instances, traditional analysis using a multivariate normal approximation for such type of variables can be misleading due to the nature of the data (small marginal means with a lot of zero counts) (Karlis & Meligkotsidou, 2007). Structural Equation Models combines both measurement and structural considerations. They integrate psychometric concepts (i.e., measurement approaches) and econometric ideas (structure approaches). Thus, this method has the ability to take into account measurement errors. As for the structure approaches in SEM, path analysis is applied to estimate the relationships among latent constructs. The ability to combine these two analyses is one of the advantages of SEM. By specifying and describing the plausible relationships between latent concepts and manifest variables, associated measurement errors, and proposed structural relationships among latent structures in SEM can effectively estimate parameters simultaneously, which mirror the fact that the variables coexist in reality.

Theoretical methods are also known as a priori methods and are used to assess diets based on prior knowledge and scientific evidence such as the dietary guideline index (Castro, Baltar, & Marchioni, 2016). Dietary indices are the most common hypothesis-oriented approaches that evaluate the adherence of population intake to nutritional recommendations. The common dietary indices include the Healthy Eating Index (HEI) that was developed to investigate American eating habits and their compliance with the dietary guidelines as provided by the Recommended Dietary Allowance (RDA) (de Calvalho, Dutra, Pizato, Gruezo, & Ito, 2014); the Original Diet Quality index that was developed to assess the intake of eight food groups and the recommendations of the committee on diet and health (Patterson et al., 1994); the Mediterranean diet score that is characterized by high intake of olive oil, non-starchy vegetables, legumes, whole grains, fruits and the low intake of whole milk and dairy products and red meats; and low to moderate intake of wine as the main source of alcohol during the meals (de Calvalho, Dutra, Pizato, Gruezo, & Ito, 2014); the Overall Nutritional Quality Index for assessing the overall nutritional quality of foods, and the Dietary Approaches to Stop Hypertension (DASH) which is a lifelong approach to healthy eating that is designed to help treat or prevent high blood pressure.

Empirical methods, sometimes referred to as a posteriori, uses statistical approaches to deduce information about existing dietary patterns within the population (Thorpe, Milte, Crawford, & McNaughton, 2016). Exploratory factor analysis is used to analyse interrelationships among a large number of variables and to explain these variables in terms of smaller number of common underlying dimensions. It involves finding a way of shrinking the information contained in some of the original variables into a smaller set of implicit variables with a minimal loss of information (Zaiontz, 2018). Principal Component Analysis (PCA) and Cluster Analysis (CA) are the other commonly used empirical methods for dietary patterns. PCA uses the correlation matrix of food

intake variables to identify common patterns of food consumption within the data to account for the largest amount of variation in diet (Thorpe, Milte, Crawford, & McNaughton, 2016). Both PCA and factor analysis are most suitable when confronted with a large number of correlated variables, and the desire is to reduce them into a small set of non-correlated variables that contains the same information of the larger one. Other reduction methods include the Cluster analysis, the Least Absolute Shrinkage and Selection Operator (LASSO), Reduced Rank Regression (RRR), and the partial least-squares regression.

This chapter thus aims to apply structural equation models to multiple-indicator, multiple causes dataset. The model is used to find the relationship between non-communicable diseases (NCD's) and the type of diets consumed in Namibia. Additionally, the chapter explored other data reduction method, PCA, to explain the type of foods consumed.

6.2. Materials and Methods

6.2.1. The NHIES 2015/16

The study used cross-sectional survey data of the Namibian Household and Income Expenditure (NHIES) of 2015/2016. The primary sampling frame that was used for this survey is a list of Primary sampling Units (PSUs) based on the 2011 Population and Housing Census Enumeration Areas (EAs). A secondary sampling frame for each of the selected PSUs was created for the purpose of selecting the sample households through a listing procedure. The sample design for the survey was a stratified two-stage cluster sample, where the first stage units were geographical areas designated as the Primary Sampling Units (PSUs) and the second stage units were the households. The up-to-date list of households in the selected PSU were prepared during the listing stage of fieldwork, and 12 households were systematically selected in each PSUs.

For this analysis, five (5) non-Communicable diseases; Diabetes (0.8%), High Blood Pressure (6.7%), Cancer (0.2%), Cardiac/Heart diseases (0.8%) and respiratory diseases (including asthma) (1.5%) were selected for analysis due to their high prevalence. Structural Equation Models (SEM) were used to model for NCD's, and the type of foods consumed. The food groups in the NHIES 2015/2016 were re-grouped and re-arranged in order to make up the 12 food groups. Principal Component analysis was used to reduce the 12 food groups to a few principal components. SPSS & statistical R Version 3.6 was used to compute PCA and SEM, respectively.

6.2.2. Statistical Methods

6.2.2.1. Principal Component Analysis (PCA)

Two extensively used empirical methods for food pattern analysis are principal component analysis (PCA) and cluster analysis (CA) (Thorpe, Milte, Crawford, & McNaughton, 2016). In order to find common patterns of food consumption within the data and account for the most variation in diet, PCA uses the correlation matrix of food intake variables (Thorpe, Milte, Crawford, & McNaughton, 2016). PCA and factor analysis are mostly used when there are a large number of potential variables to analyze and there is a need to summarize the information contained in those variables as efficiently as possible (Gleason, Boushey, Harris, & Zoellner, 2015). The following definition for PCA is derived from (Zaiontz, 2018)

Let $X = [x_i]$ be any $k \times 1$ random vector. We now define a $k \times 1$ vector $Y = [y_i]$, where for each i the i_{th} principal component of X is

$$y_i = \sum_{j=1}^{k} \beta_{ij} x_j \quad (94)$$

for some regression coefficients β_{ij}. Since each y_i is a linear combination of the x_j, Y is a random vector.

Let $\sum = [\sigma_{ij}]$ be the $k \times k$ population covariance matrix for X. Since the column vectors β_j are orthonormal, $\beta_i . \beta_j = \beta_i^T \beta_j = 0$ if $j \neq i$ and $\beta_i^T \beta_j = 1$ if $j = i$. Then the covariance matrix for Y is given by:

$$var(y_i) = \sum_{p=1}^{k} \sum_{m=1}^{k} \beta_{ip} \beta_{im} \sigma_{pm} = \beta_i^T \left(\sum_{j=1}^{k} \lambda_j \beta_j \beta_j^T \right) \beta_i = \sum_{j=1}^{k} \lambda_j \left(\beta_i^T \beta_j \right) \left(\beta_j^T \beta_i \right) = \lambda_i \quad (95)$$

$$cov(y_i, y_j) = \sum_{p=1}^{k} \sum_{m=1}^{k} \beta_{ip} \beta_{jm} \sigma_{pm} = \beta_i^T \left(\sum_{r=1}^{k} \lambda_r \beta_r \beta_r^T \right) \beta_j = \sum_{r=1}^{k} \lambda_r \left(\beta_i^T \beta_r \right) \left(\beta_r^T \beta_j \right) = 0 \quad (96)$$

It is also worth noting that the first principal component is the combination that accounts for the largest variance in the sample. The second component accounts for the next largest amount of variance and is uncorrelated with the first. Successive components thus explain progressively smaller portions of the sample variance and are uncorrelated with each other (Suresh., 2014).

Since one can calculate as many principal components as there are variables, the researcher does not gain any additional insight if all the variables are replaced by their principal components. Thus, one needs to determine how many factors are needed to represent the data, i.e., to reproduce the original correlations. There are two main criteria for deciding how many factors to extract. One by examining Eigenvalues whereby a criterion of eigenvalue greater than 1 suggests that only factors that account for variances greater than 1 should be included. Factors with a variance of less than 1 are not better than individual variables, since each variable has a variance of 1. Additionally, they can be studied using a scree plot, which plots the eigenvalues versus the number of variables in the order of extraction. The curve's point where the slope changes to a horizontal angle determines

how many factors can be derived. The maximum number of components that can be extracted is indicated at this stage (Suresh ., 2014).

The other recommended method is the Varimax Orthogonal Rotation. The VARIMAX method of rotation is the most frequently used rotation method (Hair et al., 1998, as cited in (Suresh., 2014)). It minimizes the number of variables that have high loadings on a factor, so that the factors can be interpreted more easily. The relationship between the test points remains the same as before. However, the axes are altered to interpret the factors more easily (Suresh., 2014).

6.2.2.2. Structural Equation Model (SEM)

Structural Equation Modelling abbreviated as SEM, is a very general statistical modelling technique, which is widely used in the behavioural sciences (Hox, Moerbeek, & Van De Schoot, 2017). It can be viewed as a combination of factor analysis and regression or path analysis. The interest in SEM is often on theoretical constructs, which are represented by the latent factors. The relationship between the theoretical constructs is represented by regression or path coefficients between the factors. The structural equation model implies a structure for the covariances between the observed variables, which provides the alternative name covariance structure modelling. It should be noted that the model can be extended to include means of observed variables or factors in the model, which makes covariance structure modelling a less accurate name (Hox et al., 2017).

Bardenheier, et al., (2013) used structural equation modeling with factor analysis, which groups inter-correlated variables into a single factor or latent construct, and path analysis, which includes the direct and indirect effects of factors previously reported associated with prediabetes. Direct effects are depicted as an arrow emanating from an independent variable (exposure) leading and pointing to a dependent variable (outcome). An indirect effect is depicted as a mediating variable

having an arrow pointing to it from an independent variable but also pointing to yet another dependent variable. A confounder is depicted as a variable with direct effects on both the exposure and the dependent variable. Correlations between the measurement errors of two variables are represented by two-headed curving arrows, in which case only the measurement error terms are correlated.

Latent variable models typically have several indicators for each latent construct, the capacity to test models with multiple dependent variables, and the advantage of testing multiple integrated models simultaneously rather than factors one at a time. Additionally, structural equation modeling studies the direct and indirect impacts of mediators on dependent variables as well as complex associations between various mediators (Bardenheier, et al., 2013). Equally, in a traditional regression model, mediators would not be included because they would block the pathway between the independent variable of interest and the dependent variable. Thus, in the structural-equation model, the independent factors and combined mediated relationships can be examined simultaneously, determining the impact of each of the dependent variables in the appropriate order. Thus, the SEM includes mediating effects without sacrificing indirect effects of interest. For each relationship in the SEM model, only data missing for either the independent or dependent variable would be missing from that equation (Bardenheier, et al., 2013).

The latent variable is divided into two parts namely the latent variable model and the measurement model. The latent variable is defined as follows:

$$\eta_i = \alpha_\eta + \boldsymbol{B}\eta_i + \boldsymbol{\Gamma}\xi_i + \varsigma_i \qquad (97)$$

Whereby η_i is a vector of latent endogenous variables for unit i, α_η is a vector of intercept terms for equations, $\boldsymbol{B}$ is the matrix of coefficients giving the expected effects of the latent endogenous

variables (η) on each other, ξ_i is the vector of latent exogenous variables, **Γ** is the coefficient matrix giving the expected effects of the latent exogenous variables (ς) on the latent endogenous variables (η), and ς_i is the vector of disturbances. The i subscript indexes the ith case in the sample.

The measurement model links the latent to the observed responses (indicators). It has two equations as outlined below:

$$y_i = \alpha_y + \Lambda_y \eta_i + \varepsilon_i \text{ and} \quad (98)$$

$$x_i = \alpha_x + \Lambda_x \xi_i + \delta_i \quad (99)$$

Where y_i and x_i are vectors of the observed indicators of η_i and ξ_i, respectively, α_y and α_x are intercept vectors, Λ_y and Λ_x are matrices of factor loadings or regression coefficients giving the impact of the latent $\eta_i$and ξ_i on y_i and x_i, respectively, and ε_i and δ_i are the unique factors of y_i and x_i.

6.2.2.3. Model Selection

Chi-square test statistics is the most used when modelling latent variables to measure/quantify model fit; however, it is sensitive to large sample size. Methodologists developed numerous fit indices to adjust the chi-square test statistics with the information in the model, such as degrees of freedom, sample size, and/or the number of variables. Chi-square can be calculated as follows:

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected}$$

Depending on the elements in the formula, fit indices in latent variable models can be categorized into three types (Chang, Gardiner, Houang, & Yu, 2020): 1) relative fit indices, Comparative Fit

Index and the absolute fit indices). R packages: “lavaan” and “semPlot” were used to model the structural equations.

6.3. Results

6.3.1. Prevalence of Non- Communicable Diseases

Non-communicable diseases are a concern in Namibia. High Blood pressure was found to be highest (6.7%) NCD among the population in Namibia (Table 34). Other NC diseases that most people are suffering from diseases include Asthma and epilepsy (1.0%), diabetes and cardiac/heart diseases (0.8%) respectively, and cancer (0.2%).

Table 34: non-communicable diseases in Namibia

Disease	Frequency	%
NCD1: Diabetes	348	0.8
NCD2: High blood pressure	2785	6.7
NCD3: Cancer	72	0.2
NCD4: Cardiac / Heart	336	0.8
NCD5: Respiratory disease (asthma, etc.)	641	1.5
NCD6: No Chronic illness	37399	89.9
Total	**41581**	**100**

6.3.2. Types of Food Consumed

Increased prevalence of obesity, increased consumption of poor-quality diets, and pervasive undernutrition are contributing to the NCD epidemic. Specifically, the NCD burden is associated with diets low in fruits and vegetables, high in sodium, low in nuts and seeds, low in whole grains, and low in seafood-derived omega-3 fatty acids (UNSCN, 2018). Table 35 shows that 23.4% of the food consumed are local foods mostly made from wheat or grain, 19.6% of the food consumed were from foods made with oil, fat or butter, 18.3% is meat products and 17.9% from sugar or honey. High consumption of these foods is associated with NCD’s (UNSCN, 2018).

Table 35: Type of foods consumed.

Food Type	Yes		No	
	Frequency	%	Frequency	%
Any (local food) bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice, wheat, or (any other local grain)	9739	23.40%	351	0.80%
Beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or other organ meats	7663	18.30%	2427	5.80%
Foods made with oil, fat or butter	8157	19.60%	1933	4.60%
Sugar/honey	7428	17.90%	2662	6.40%

6.3.3. Association of Type of Foods Consumed and Non-Communicable Diseases

6.3.3.1. Local grain foods

At least 6.9% of the households with High Blood pressure have indicated that they consume "Any (local food) bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice, wheat, or (any other local grain)". Other NCD with A high percentage in consumption of local foods/grain/wheat was the respiratory diseases (including asthma) (1.4%). Additionally, the Pearson's Chi-square test indicated that there was no association between NCDs and the food type ("Any (local food) bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice, wheat, or (any other local grain)"), P-value 0.780. (Table 36).

Table 36: Association of NCD and Local Food

Disease	Any local food/grain/wheat		Total	Pearson Chi-Square	
	No	Yes		Value	Asymptotic Significance (2-sided)
Diabetes	0.90%	0.80%	0.80%	3.228	0.78
High Blood Pressure	6.00%	6.90%	6.90%		
Cancer	0.00%	0.20%	0.20%		
Cardiac or Heart	1.40%	0.90%	0.90%		
Respiratory Disease (Inc. Asthma)	2.00%	1.40%	1.50%		
Does not have a Chronic illness	87.70%	87.40%	87.40%		
Total	**100.00%**	**100.00%**	**100.00%**		

6.3.3.2. Meat Products

At least 7.1 percent of individuals with high blood pressure consumed meat/chicken products. Meat products included beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or other organ meats. The Pearson's Chi-square test shows that there is a significant relationship between the NCDs and Meat/Chicken products (P-value- 0.0034).

Table 37: Association of NCD and Meat

Disease	Meat/Chicken Products		Total	Pearson Chi-Square	
	No	Yes		Value	Asymptotic Significance (2-sided)
Diabetes	0.90%	0.70%	0.80%	10.837	0.034
High blood pressure	6.20%	7.10%	6.90%		
Cancer	0.10%	0.20%	0.20%		
Cardiac/Heart	0.90%	0.90%	0.90%		
Respiratory disease (asthma, etc.)	1.00%	1.60%	1.50%		
Does not have a Chronic illness	88.70%	87.00%	87.40%		
Total	**100.00%**	**100.00%**	**100.00%**		

6.3.3.3. Foods made with Oil, Fat or Butter

Table 38 indicates that 6.9%, 1.4%, 0.8%, 0.8% and 0.2% of the households had High Blood pressure, Respiratory diseases, Cardiac/ Heart, Diabetes, and cancer respectively (Table 38). The Pearson Chi-square was however not significant at 5% and indicated no association between the NCDs and foods made with oil, fat or butter.

Table 38: Association of NCD with Fats/Oils

Disease	Foods made with oil, fat or butter		Total	Pearson Chi-Square	
	No	Yes		Value	Asymptotic Sign. (2-sided)
Diabetes	0.80%	0.80%	0.80%	8.537	0.201
High blood pressure	6.80%	6.90%	6.90%		
Cancer	0.30%	0.20%	0.20%		
Cardiac/Heart	1.40%	0.80%	0.90%		
Respiratory disease (asthma, etc.)	1.60%	1.40%	1.50%		
Does not have a chronic illness	86.50%	87.70%	87.40%		
Total	100.00%	100.00%	100.00%		

6.3.3.4. Sugar/Honey

Table 39 shows that at least 7.0 percent of the high blood pressure, 1.5% respiratory diseases, 0.9% Cardiac/Heart, 0.8% Diabetes and 0.2% Cancer individuals consumed sugar or honey products. According to the Pearson Chi-Square test, there was no association between NCDs and Sugar/Honey (P-value greater than 0.005).

Table 39: Association of NCD and Sugar/Honey

Disease	Sugar/Honey		Total	Pearson Chi-Square	
	No	Yes		Value	Asymptotic Sign. (2-sided)
Diabetes	0.60%	0.80%	0.80%	12.273	0.056
High blood pressure	6.50%	7.00%	6.90%		
Cancer	0.10%	0.20%	0.20%		
Cardiac/Heart	1.00%	0.90%	0.90%		
Respiratory disease (asthma, etc.)	1.30%	1.50%	1.50%		
Does not have a chronic illness	87.30%	87.50%	87.40%		
Total	**100.00%**	**100.00%**	**100.00%**		

6.3.4. Principal Component Analysis (PCA)

Principal Component analysis was used to reduce the 12 food groups to a few principal components. The PCA extracted three (3) components with eigen values greater than 1, explaining 49.4% of the total variance in the data set. The first, second and third components explained 29.7%, 10.4% and 9.4% respectively of all variations (Table 40).

Table 40: PCA components

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		Rotation Sums of Squared Loadings[a]	
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	3.560	29.670	29.670	3.560	29.670	29.670	3.085
2	1.246	10.383	40.053	1.246	10.383	40.053	2.038
3	1.125	9.374	49.426	1.125	9.374	49.426	1.500
4	.902	7.516	56.943				
5	.852	7.103	64.046				
6	.750	6.247	70.294				
7	.735	6.126	76.419				
8	.629	5.243	81.662				
9	.581	4.842	86.504				
10	.555	4.624	91.128				
11	.536	4.470	95.598				
12	.528	4.402	100.000				
Extraction Method: Principal Component Analysis.							
a. When components are correlated, sums of squared loadings cannot be added to obtain a total variance							

Component 1, 2 and 3 accounted for majority of the variances and had eigen values of 3.6, 1.2 and 1.1 respectively. This implies the PCA explained 49.9% of the food types summarized as three (3) underlying dimensions coined from the food types loaded significantly in the 3 extracted Components. Table 41 indicates that the variables can be grouped into three (3) components with three (3) factor loadings each. Food made with oil, fat and butter, and vegetables food items overlaps across the components but with their strongest loading of 0.516 and 0.477 respectively in the 1st component.

Table 41: Component Matrix of the PCA

Food Types	Component		
	1	2	3
Fruits	.682	-.334	
Any foods such as condiments/tea/coffee	.672		
Potatoes, yams, cassava, or any foods made from roots and tubers	.670		
Eggs	.657		
Cheese, yoghurt, milk or other milk products	.605		-.322
Sugar/honey	.597	.450	
Beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or other organ meats	.552		
Foods made with oil, fat or butter	.516	.439	.308
Vegetables	.477	-.320	.380
Food made from beans, peas, lentils or nuts	.319	-.448	
Any (local food) bread, rice, noodles, biscuits or any other foods made from millet, sorghum, maize, rice, wheat, or (any other local grain)		.413	
Fresh or dried fish or shellfish			.733
Extraction Method: Principal Component Analysis.			
a. 3 components extracted.			

Figure 3 shows graphic dimensions to determine the number of components to be extracted. The Scree plot suggests taking the first 3 components.

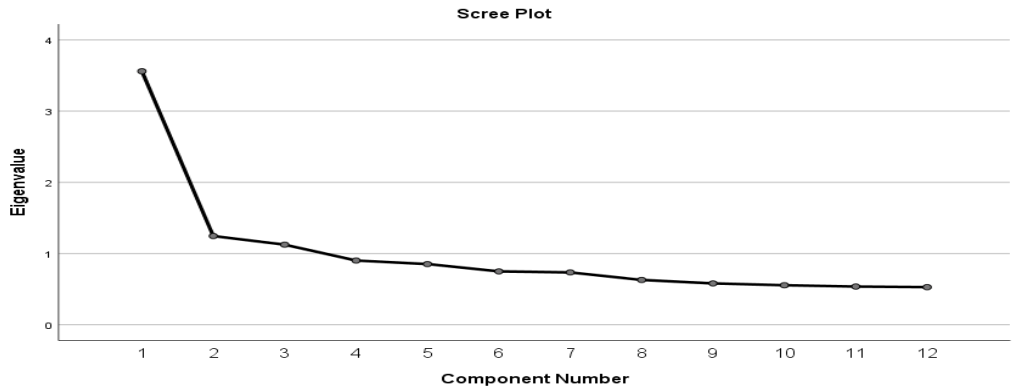


Figure 3: Scree Plot of food types

6.3.5. Structural Equation Modelling (SEM)

Our model is estimated by maximum likelihood (ML). A likelihood ratio statistic comparing the fitted model (with 75 parameters) to the unconstrained saturated model produces a p value of less

than 0.001 with df = 166. The chi square test is significant suggesting good fit. Standardized Root Mean Square Residual (SRMR), Comparative Fit Index (CFI), and Tucker-Lewis Index (TLI) are used to assess model fit. In this study, SRMR = 0.063, CFI= 0.850 and TLI= 0.799, which are near the recommended cutoffs of less than 0.08 and more than 0.95, indicating that the model is a good-fitting model. Comparative indices, such as AIC and BIC, are used to compare competing models, that is, with different co-variance structures. Since there is no competitive model in our example, AIC and BIC is not used in this case.

Table 42: SEM Model Specifications

Estimator	**ML**
Optimization method	NLMINB
Model Test User Model	
Test Statistic	6211.565
D.F	168
P-Value	<0.001
Model Test Baseline Model	
Test Statistic	40403.294
D.F	225
P-Value	<0.001
User Model versus Baseline Model	
Comparative Fit Index (CFI)	0.850
Tucker-Lewis Index (TLI)	0.799
Log likelihood and Information criterion	
Log likelihood User Model (H0)	-9079.044
Log likelihood unrestricted model (H1)	-5973.261
Akaike (AIC)	18308.087
Bayesian (BIC)	18849.535
Sample-size adjusted Bayesian	18611.196
Root Mean Square Error of Approximation	
RMSEA	0.060
90% CU – Lower	0.058
90% CI- Upper	0.061
P-value RMSEA <=0.05	0.000
Standardized Root Mean Square Residual	
SRMR	0.063
Parameter Estimates	
Standard errors	Standard
Information	Expected
Information saturated (h1) model	Structured

R packages 'lavaan" and "semPlot" used to run structural equation modeling in this chapter (Table 43, 44, 45, 46), and the latter one is for generating the diagram (Figure 4). The "Std.lv" column reported the estimates when the latent variables "FS" (Food Security) and "NCD" (Non-Communicable Diseases) were standardized. The last column "Std.all" reported the parameter estimates when both the latent variables and the observed variables were standardized (also called the 'completely standardized solution'). The function semPaths is used to plot the SEM diagrams (Figure 4).

Table 43: Parameter Estimates: Latent Variables

Latent variables. FS=~	**Estimate**	**Std. err**	**P value**	**Std. lv**	**Std. all**
Food Type 1	1.000			0.038	0.206
Food Type 2	7.975	0.435	<0.001	0.301	0.629
Food Type 3	5.303	0.310	<0.001	0.200	0.405
Food Type 4	7.829	0.427	<0.001	0.295	0.636
Food Type 5	5.443	0.308	<0.001	0.205	0.481
Food Type 6	7.179	0.393	<0.001	0.271	0.613
Food Type 7	2.718	0.203	<0.001	0.103	0.205
Food Type 8	2.537	0.168	<0.001	0.096	0.265
Food Type 9	7.263	0.402	<0.001	0.274	0.553
Food Type 10	4.450	0.257	<0.001	0.168	0.427
Food Type 11	6.071	0.339	<0.001	0.229	0.520
Food Type 12	0.154	0.446	<0.001	0.308	
Ncd1=~ncd_1 (Diabetes)	1.000			0.087	1.000
Ncd2=~ncd_2 (High blood pressure)	1.000			0.253	1.000
Ncd3=~ncd_3 (Cancer)	1.000			0.044	1.000
Ncd4=~ncd_4 (Cardiac/Heart)	1.000			0.095	1.000
Ncd5=~ncd_5 (Respiratory illness)	1.000			0.120	1.000
Ncd6=~ncd_6 (No NCD)	1.000			0.331	1.000

Table 44 shows regression estimates. Educational level of a household and residence type was found to have a statistically significant relationship (p-value less than 0.001) with diabetes (NCD1). Other variables that were significant at 5% are residence (Diabetes (NCD1), Cardiac/Heart (NCD4) and No NCD (NCD6)). Table 45 show covariances of NCD's.

Table 44: Parameter Estimates: Regression

Variable	Estimate	Std. err	P value	Std. lv	Std. all
Ncd1 (Diabetes)~					
Food Security	-0.042	0.026	0.100	-0.018	-0.018
Attain (educational level)	0.005	0.001	<0.001	0.060	0.060
li_urbrur (residence)	-0.000	0.000	0.011	-0.001	-0.026
q04_20 (Smoking)	-0.005	0.003	0.105	-0.056	-0.017
Q04_22 (Alcohol consumption)	0.002	0.002	0.315	0.025	0.011
Ncd2 (High Blood Pressure) ~					
Food Security	0.022	0.073	0.760	0.003	0.003
Attain (educational level)	0.035	0.003	<0.001	0.140	0.139
li_urbrur (residence)	0.000	0.000	0.167	0.000	0.014
q04_20 (Smoking)	-0.040	0.009	<0.001	-0.157	-0.048
Q04_22 (Alcohol consumption)	0.007	0.0060	0.260	0.028	0.012
Ncd3 (Cancer) ~					
Food Security	0.021	0.013	0.109	0.018	0.018
Attain (educational level)	0.002	0.000	<0.001	0.038	0.038
li_urbrur (residence)	-0.000	0.000	0.796	-0.000	-0.003
q04_20 (Smoking)	-0.003	0.002	0.045	-0.069	-0.021
Q04_22 (Alcohol consumption)	0.003	0.001	0.045	0.058	0.025
Ncd4 (Cardiac/heart) ~					
Food Security	0.013	0.028	0.635	0.005	0.005
Attain (educational level)	0.002	0.001	0.085	0.018	0.017
li_urbrur (residence)	0.000	0.000	0.003	0.001	0.030
q04_20 (Smoking)	-0.001	0.003	-0.373	-0.013	-0.004
Q04_22 (Alcohol consumption)	-0.000	0.002	-0.083	-0.002	-0.001
Ncd5 (Respiratory illness) ~					
Food Security	0.044	0.035	0.208	0.014	0.014
Attain (educational level)	0.002	0.001	0.144	0.015	0.015
li_urbrur (residence)	0.000	0.000	0.248	0.000	0.012
q04_20 (Smoking)	-0.008	0.004	0.046	-0.069	-0.021
Q04_22 (Alcohol consumption)	0.004	0.003	0.224	0.030	0.012
Ncd6 (No NCD) ~					
Food Security	0.009	0.096	0.092	0.001	0.001
Attain (educational level)	-0.047	0.003	<0.001	-0.142	-0.141
li_urbrur (residence)	-0.000	0.000	0.021	-0.000	-0.023
q04_20 (Smoking)	0.063	0.011	<0.001	0.190	0.058
Q04_22 (Alcohol consumption)	-0.017	0.008	0.037	-0.051	-0.022

Table 45: Parameter Estimates: Covariances

Variable (NCD)	Estimate	Std.err	P value	Std.lv	Std.all
Ncd1~~					
Ncd2	-0.001	0.000	**0.001**	-0.033	-0.033
Ncd3	-0.000	0.000	0.505	-0.007	-0.007
Ncd4	-0.000	0.000	0.384	-0.009	-0.009
Ncd5	-0.000	0.000	0.245	-0.012	-0.012
Ncd6	0.006	0.000	**0.001**	-0.225	-0.225
Ncd2~~					

Ncd3	-0.000	0.000	0.064	-0.018	-0.018
Ncd4	-0.001	0.000	0.004	-0.029	-0.029
Ncd5	-0.001	0.000	**<0.001**	-0.036	-0.036
Ncd6	-0.058	0.001	**<.0001**	-0.710	-0.710
Ncd3~~					
Ncd4	-0.000	0.000	0.642	-0.005	-0.005
Ncd5	-0.000	0.000	0.506	-0.007	-0.007
Ncd6	-0.002	0.000	<0.001	-0.113	-0.113
Ncd4~~					
Ncd5	-0.000	0.000	0.225	-0.012	-0.012
Ncd6	-0.008	0.000	**<0.001**	-0.252	-0.252
Ncd5~~					
NCd6	-0.013	0.000	**<0.001**	-0.321	-0.321

Table 46 shows variances of food types and NCD's. all the food types were significant at 5%.

Table 46: Parameter Estimates: Variances

Variable	Estimate	Std. err	P value	Std. lv	Std. all
Food Type 1	0.032	0.000	<0.001	0.032	0.958
Food Type 2	0.138	0.002	<0.001	0.138	0.604
Food Type 3	0.204	0.003	<0.001	0.204	0.836
Food Type 4	0.129	0.002	<0.001	0.129	0.596
Food Type 5	0.140	0.002	<0.001	0.140	0.769
Food Type 6	0.122	0.002	<0.001	0.122	0.624
Food Type 7	0.239	0.003	<0.001	0.239	0.958
Food Type 8	0.121	0.002	<0.001	0.121	0.930
Food Type 9	0.170	0.003	<0.001	0.170	0.694
Food Type 10	0.127	0.002	<0.001	0.127	0.818
Food Type 11	0.142	0.002	<0.001	0.142	0.730
Food Type 12	0.155	0.002	<0.001	0.155	0.621
Ncd1=~ncd_1 (diabetes)	0.000			0.000	0.000
Ncd2=~ncd_2 (High blood pressure)	0.000			0.000	0.000
Ncd3=~ncd_3 (Cancer)	0.000			0.000	0.000
Ncd4=~ncd_4 (Cardiac/Heart)	0.000			0.000	0.000
Ncd5=~ncd_5 (Respiratory illness)	0.000			0.000	0.000
Ncd6=~ncd_6 (No NCD)	0.000			0.000	0.000
FS (Food Security)	0.001	0.000	<0.001	1.000	1.000
NCDs	0.008	0.000	<0.001	0.995	0.995
NCD2	0.063	0.001	<0.001	0.979	0.979
NCD3	0.002	0.000	<0.001	0.998	0.998
NCD4	0.009	0.000	<0.001	0.999	0.999
NCD5	0.014	0.000	<0.001	0.999	0.999
NCD6	0.107	0.002	<0.001	0.978	0.978

Figure 4 shows structural equation modelling pathways between type of foods consumed, non-communicable diseases and other socio-economic variables.

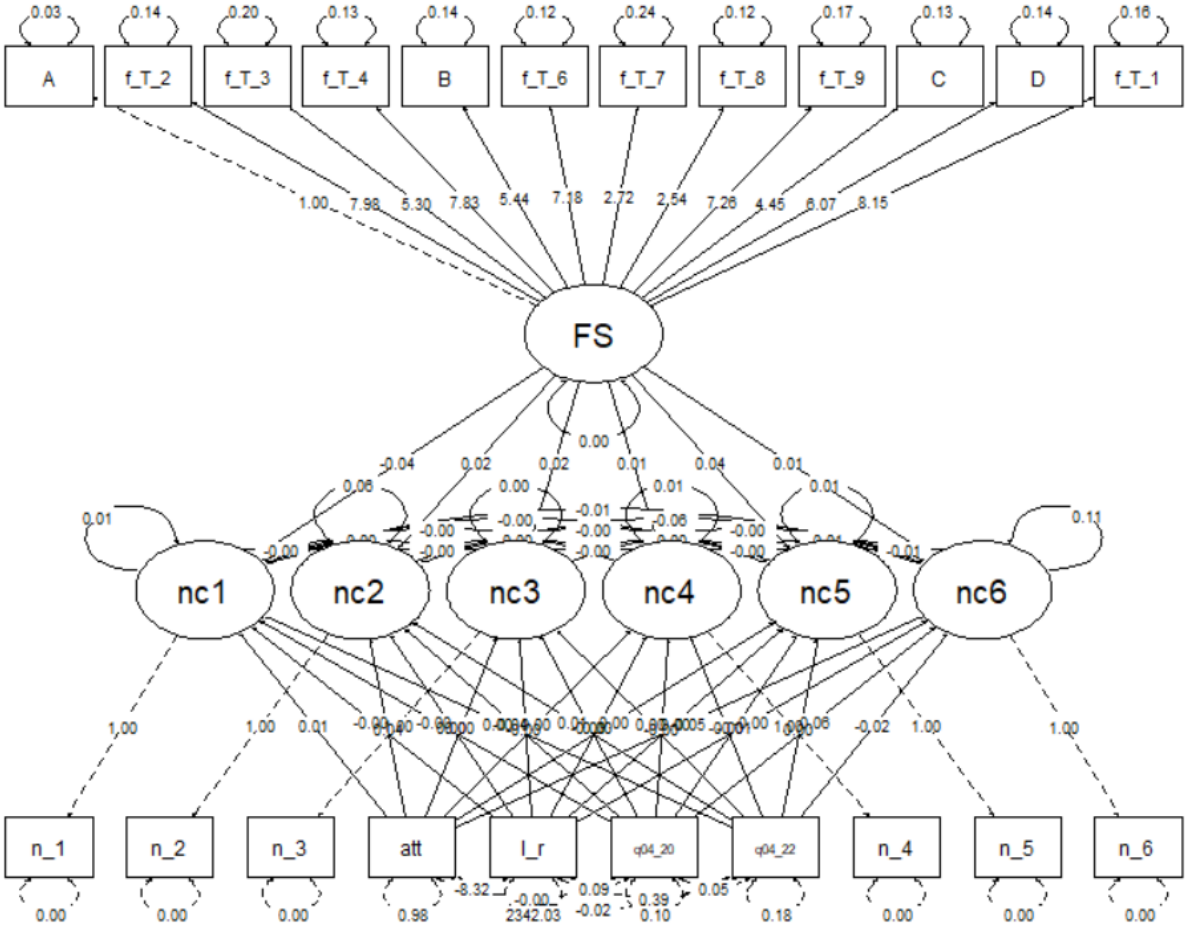


Figure 4: SEM: Foods Consumed, NCDs and Socio-economic variable

6.4. Discussion

This chapter modelled multiple indicator-multiple causes regression using Structured Equation Models (SEM). The study first looked at the prevalence of NCD's in the country. Non-communicable diseases have been on a rise, especially in low-middle income countries for the past

decades (UNSCN, 2018). High blood pressure was particularly found to contribute largely to NCDs in the country. High Blood Pressure or sometimes referred to as hypertension, is a common disorder that affects a large proportion of the community. It is mostly asymptomatic and is detected on routine exams or after the occurrence of a complication such as a heart attack or stroke (Sunil & Gregory, 2021). Globally, the overall prevalence of hypertension in adults is estimated to be between 30-45%, with a higher prevalence in men than women (24% and 20%) respectively (Williams, Mancia, & Spiering, 2018). Other types of NCDs are respiratory illnesses, diabetes and cardiac or heart disease and cancer.

The types of food consumed has a significant contribution to the presence of an NCD in an individual. A diet that lacks fruits and vegetables and has a high intake of sodium, low intake of nuts and seeds, low in whole grains as well as in seafood-derived omega-3 fatty acids is specifically associated with a high prevalence of NCDs (Global Pattern, 2016). This chapter analysed the association between NCDs and foods types such as local food (bread, rice, noodles, biscuits or any other food made from millet, sorghum, maize, rice, wheat, or any other local grain), Meat (beef, pork, lamb, goat, rabbit, wild game, chicken, duck, other birds, liver, kidney, heart or other organ meats), Foods made with oil, fat or butter. Among all the food types, the meat was found to have a significant effect to NCDs. An analysis done by the American Heart Association found that each serving per day of processed meat was associated with a 42% higher risk of coronary heart disease and a 19% higher risk of diabetes, while total meat intake was associated with a 25% higher risk of coronary heart disease (Micha, Wallace, & Mozaffarian, 2010).

Empirical methods, particularly the Principal Component Analysis (PCA) as a data reduction method was used to derive dietary patterns. PCA used the correlation matrix of food intake groups

to identify common patterns of food consumption in the dataset to account for the largest amount of variation in diet. Fruits, foods such as condiments/tea/coffee and potatoes, yams, cassava, or any foods made from roots and tubers accounted for majority of the variation. Furthermore, the SEM was derived through the packages “lavaan” and “semPlot” to analyse multivariate data with multiple indicators. Food types such as local grains, meat and food made from oil were found to be significant at 5% level and associated with NCDs. This concurs with the findings by Micha, Wallace, & Mozaffarian (2010).

6.5. Conclusion

Multiple data analysis has been on an increase for researchers for the past decades, but models accounting for multiple-indicators, multiple cause data are rarely applied for its computational difficulties. Mostly, for single counts, the Poisson regression model is used. The limitations of Poisson regression model are the assumption of equal mean and variance and restricting the count variables to positive. The aim of this study was to model multiple-indicator, multiple-cause to examine the relationship between foods consumed and NCDs. The study concluded that the type of food consumed has a significant contribution to the incidence of an NCD in an individual. In this chapter, we employed Structural Equation Models. The SEM analyses structural relationships between Non-Communicable Diseases and Foods Consumed and other Socio-economic variables.

6.6. Acknowledgements

The authors had support from the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)’s Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa’s Development Planning and

Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [grant 107754/Z/15/Z- DELTAS Africa Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) programme] and the UK government.

CHAPTER 7: CONCLUSIONS AND RECOMMENDATIONS

7.1. Introduction

The analysis of nutritional epidemiology is of importance in determining the health effects of various food groups on individuals and households. The recommendation and consumption of certain diets may influence the health and risk of developing Non-Communicable Diseases. With high levels of urbanization in developing countries, NCD's has been on an increase for the past decades. The rise of NCDs has also been attributed to sedentary lifestyles, eating between main meals, buying street or processed foods and eating outside the home. Comparison of different models is a very important aspect in identifying the best possible fit model for different data types.

7.2. Review and Evaluation of the Objectives

7.3.1. Count Models Application on Dietary Diversity in Namibia

There are several food security and dietary diversity measurement approaches. The Household dietary Diversity Score (HDDS) adopts the food groups approach and asks how many of the 12 different groups were consumed in the household over a specific recall period (24 hours). This study concluded that households consumed mostly local foods from grains, foods made with oil, fat or butter and sugar/honey. Additionally, there was less consumption in foods made from beans, peas, lentils, or nuts, eggs, fruits and cheese. Rural households particularly had a low dietary diversity (75%) and consumed a monotonous diet, while households headed by individuals with salaries and wages had a higher dietary diversity.

Count of foods or food groups has been widely practiced in dietary diversity analysis. Poisson regression model provides a basis for the analysis of count data; however, the classical Poisson

regression model cannot be used alone especially for complex data types as it is often of limited use because empirical count data sets typically exhibit over-dispersion. Various models including the Poisson regression, negative binomial regression, Poisson inverse gaussian regression, Poisson logit hurdle, zero inflated Poisson inverse gaussian and the Conway-Maxwell Poisson were fitted, and their AIC and BIC analyzed to find the fitness of each model. The Poisson inverse gaussian regression fitted the data better. The Poisson inverse gaussian has the advantage that it is a mixture model, specifically a mixture of Poisson and Inverse gaussian distributions. It is mostly used to model count data that have a higher initial peak and that may be skewed to the right as well as data that are highly over dispersed.

7.3.2. Convenience and Non-Convenience Consumption Food Preference in Windhoek, Namibia: A Bivariate Count System Approach

The nutritional transition in the world has greatly affected the dietary pattern and nutrient intake and has led to a rise in the purchases and consumption of processed and convenience foods. Event counts such as the number of convenience and non-convenience foods consumed are likely to be jointly dependent. Bivariate Poisson models are appropriate for modelling count data exhibiting correlation and require joint estimation.

The study found that households purchased convenience foods from street sellers/traders/hawkers and spaza/tuck-shops more on a weekly basis. On a monthly period, households purchased more non-convenient foods. Bivariate Poisson regressions, the truncated and untruncated models were fitted to find the relationship between the convenience and non-convenience versus weekly and monthly intakes. The untruncated bivariate Poisson model fitted the data best. The study further

concluded that the variables age, marital status, work status and educational level of head of household influenced the choices of foods a household makes.

7.3.3. Copula Joint Modelling of Food Insecurity Indicators using Copulas: FIP, HDDS and MIHFP

Food insecurity indicators such as Food Insecurity prevalence (FIP), Household Dietary Diversity Score (HDDS) and Months of Inadequate Household Food Provisioning (MIHFP) are used to estimate food insecurity levels of a household. The Generalized Joint Regression Models (GJRM) through Copulas were used to estimate the relationship between food security outcomes/indicators and other exposure variables.

The Copula approach is defined as a useful method for deriving joint distributions. The approach relates an arbitrary joint distribution to its corresponding univariate marginal distributions via Copula. Specifically, five (5) Copula families namely the Bivariate normal, Frank, Survival, Clayton, Gumbel and the Survival Gumble were used in this analysis. The Frank Copula was identified to fit the data between FIP and HDDS better while the Bivariate normal better fitted the data between FIP and MIHFP. Sample selection and Partial Observability were conducted to observe specific responses between the three indicators. The socio-demographic variables were all not significant at 5% indicating a non-relationship between the exposure and outcome variables.

7.3.4. Multiple-Indicator, Multiple-Cause Modelling to Examine the Relationship between Foods Consumed and Non-Communicable diseases

The usage of Structural Equation Models (SEM) for the measurement of multiple-indicator-multiple-causes variables such as prevalence on non-communicable diseases and the number of food groups consumed in a household helps with dealing with measurement errors by combining

both measurement (Psychometric approaches) and structural (econometric ideas) considerations. Principal Component Analysis (PCA) is a data reduction method and was used to reduce the 12 food groups to a few individual components. The study found that fruits, foods such as condiments/tea/coffee and potatoes, yams, cassava, or any food made from roots and tubers account for majority of the variation. Furthermore, foods such as local grains, meat and foods made from oil were found to have an association with NCDs.

7.3. Recommendations

The study found that non-communicable diseases are prevalent in the country. Of particular, high blood pressure, respiratory diseases such as asthma, cardiac/heart and diabetes were found to be high. The type of foods consumed contributed to the prevalence of NCDs. The consumption of meat products such as beef, pork, lamb, goat, rabbit, wild game, chicken, duck and other birds, liver, kidney, heart or other organ meats was associated with NCDs. This study thus recommends that the government and private should intensify programs on health education, particularly on the risk factors of NCDs and strengthen advocacy of healthy and diverse diet needs by individuals and households in the country to prevent a further rise in NCD's.

An increase in the number of different types of food consumed reflects an improvement in the household's diet. The study found that households consumed most local foods from wheat such as bread, rice, noodles, and biscuits, meats and foods made with oil, fat or butter and sugar/honey. The household dietary diversity assessment found that if the head of the household's main source of income comes from salary or wages had a higher dietary diversity. Furthermore, households with only primary education, their dietary diversity was low compared to those with secondary and tertiary education. This study thus recommends that the government find strategies to invest

in programmes that will increase agriculture production for cereal horticulture especially in rural areas where monotonous diets are common. The government and private sector should further devise targeted strategies that enhance wealth redistribution and household income. Additionally, it is recommended that the government creates an enabling environment and policy reforms that ensure greater food insecurity reduction is reflected in national planning processes.

The usage of different models and approaches in food security, dietary diversity and food consumption allows for flexibility to explore the association among variables in different models and allow for comparison of best fit modelling. Analysis in nutritional epidemiology traditionally focused on examining diseases in relation to a single or a few nutrients or foods. But people do not consume isolated nutrients. The higher degree of inter-correlation among nutrients as well as among foods make it difficult to attribute effects to single dietary component. It is thus this study recommend the usage of different methodological approaches in the analysis of food insecurity needs to be strengthened. Different data types require specific data analysis approaches.

References

Colón-Ramos, U., Kabagambe , E., & Baylin, A. (2007). Socio-economic status and health awareness are associated with choice of cooking oil in Costa Rica. *Public Health Nutrition*, 1214–1222.

Abegunde, D., Mathers, C., Adam, C., Ortegon, M., & Strong, K. (2007). The burden and costs of chronic diseases in low-income countries. *Lancet*.

Achim, Z., Christian , K., & Simon, J. (2008). Regression Models for Count Data in R. *Journal of statistical software*.

Ademora, A. J., & Ahamefula, M. U. (2012). Multivariate Generalized Poisson Distribution for Interference on selected Non-Communicable Diseases in Lagos Status . *Journal of Modern Applied Statistical Methods*, 524-529.

Ajieroh, V. (2009). *A quantitative analysis of determinants of Child and Malnutrition in Nigeria.* Washington, DC: IFPRI.

Anyadike, R. (2009). *Statistical Methods for the Social Sciences.* Ibadan: Spectrum Books Limited.

Arimond, M., & Ruel, M. (2002). *Summary indicators for infant and child feeding practice: an example from the Ethiopia demographic and Health survey 200.* Washington: International Food Policy Research Institute.

Arimond, M., & Ruel, M. T. (2004). Dietary diversity is associated with child nutritional status: Evidence from 11 demographic and health surveys. . *J Nutri*, 2579-2585.

Baiocchi de Carvalho, K., Dutra, E., Pizato, N., Gruezo, N., & Ito, M. (2014). Diet quality assessment indexes. *Scielo*.

Bardenheier, B. H., Bullard, K. M., Caspersen, C. J., Cheng, Y. J., Gregg, E. W., & Geiss, L. S. (2013). A Novel Use of Structural Equation Models to Examine Factors Associated With Prediabetes Among Adults Aged 50 Years and Older. *PMCID*, 2655–2662.

Battersby, J. (2013). Hungry Cities: A critical review of urban food security research in sub-Saharan African cities. *Geography Compass*, 452-463.

Becquey, E., Mathilde , S., Peggy , D., Hubert , D. B., Sylvestre , T., & Yves , M.-P. (2010). Dietary patterns of adults living in Ouagadougou and their association with overweight. *Nutrition Journal*, 13.

Bennett, D., Landry, D., Little, J., & Minelli, C. (2017). Systematic review of statistical approaches to quantify, or correct for, measurement error in a continuous exposure in nutritional epidemiology. *BMC Medical Research Methodolofy*.

Bernal, R. J., & Lorenzana, A. P. (2003). Dietary diversity and associated factors among beneficiaries of 77 child care Centers: Central Regional. *Venezuela*, 52-81.

Bilinsky, P., & Swindale, A. (2010). *Months of Adequate Household Food Provisioning (MAHFP) for measurement of Household Food Access: Indicator Guide .* Washington, DC: FANTA.

Black, E. (2016). Globalization of the Food Industry: Transnational Food Corporations, the Spread of Processed Food, and Their Implications for Food Security and Nutrition. *Spring*.

Black, R. E., Victora, C., Walker, S. P., Bhutta, Z. A., Christian, P., & de, O. M. (2013). Maternal and child undernutrition and overweight in low-income and middle-income countries. *The Lancet*, 427-51.

Broatch, J., & Karl, A. (2017). Multivariate Generalized Mixed Models for Joint Estimation of Sporting Outcomes. *Italian Journal of Applied Statistics*.

Burggraf, C., Teuber, R., Brosig, S., & Meier, T. (2018). Review of a priori dietary quality indices in relation to their construction criteria. *Nutrition Reviews*, 747-764.

Cameron, C., & Trivedi , P. (1999). Essentials of Count Data Regression.

Castro, B. H., Baltar, V. T., & Marchioni, D. M. (2016). Examining associations between dietary patterns and metabolic CVD risk factors: a novel use of structural equation modelling. *British Journal of Nutrition*, 1586-1597.

Castro, M. A., Baltar, V. T., Marchioni, D. M., & Fisberg, R. M. (2016). Examining associations between dietary patterns and metabolic CVD risk factors: a novel use of structural equation modelling. *British Journal of Nutrition*, 1586–1597.

Chang, C., Gardiner, J., Houang, R., & Yu, Y.-L. (2020). Comparing Multiple Statistical Software for multiple-indicator, Multiple-Cause Modelling: an application of gender Disparity in adult cognitive functioning using MIDUS 2 dataset. *BMC Medical research*.

Chesnaye, N., Tripepi, G., Dekker, F., Zoccali, C., Zwinderman, A., & Jager, K. (2020). An introduction to joint models- applications in nephrology. *Clinical Kidney*, 143-149.

Chopra, M., Galbraith, S., & Darnton-Hill, I. (2002). A global response to a global problem: the epidemic of overnutrition. *Bull*, 952-958.

Chou, N.-T., & Steenh, D. (2011). *Bivariate Count Data Regression Models – A SAS® Macro Program.*

Chowdhury, I. R., & Islam, M. A. (2019). bpglm: R package fo Bivariate Poisson GLM with covariates.

Chowdrhurry, R. I., & Islam, M. A. (2016). A bivariate Poisson models with Covariate Dependence. *Applied Mathematics*, 1589-1598.

Coates , J., Swindale, A., & Bilinsky, P. (2007). *Household Food Insecurity Access Scale (HFIAS) for measurement of food access: indicator guide.* Washington, DC.

Consul, P. C. (1989). *Generalized Poisson Distributions: Properties and Applications.* New York: Marcel Dekker.

Cordain, L., Eaton, S., Sebastian, A., Mann, N., Lindeberg, S., & Watkins, B. (2005). Origins and evolution of the Western diet: health implications for the 21st century. *Am J Clin Nutr*, 81:341–54.

de Calvalho, K., Dutra, E., Pizato, N., Gruezo, N., & Ito, M. (2014). Diet Quality assessment indexes. *Scieli*.

de Carvalho, K., Dutra , E. S., Pizato, N., Gruezo, N. D., & Ito, M. K. (2014). Diet quality assessment indexes. *sciELO*.

de Oliveira Otto, M., Anderson, C., Dearborn, J., Ferranti, E., Mozaffarian, D., Rao, G., . . . Lichtenstein, A. (2018). Dietary Diversity: Implications for Obesity Prevention in Adult Populations. *Circulation*, 160-168.

Dekker, L. H., Rijnks, R. H., Strijker, D., & Navis, G. J. (2017). A spatial analysis of dietary patterns in a large representative population in the north of The Netherlands – the Lifelines cohort study. *PMCID*.

Devlin, U. M., Mcnulty, B. A., Nugent, A. P., & Gibney, M. J. (2012). The use of cluster analysis to derive dietary patterns: methodological considerations, reproducibility, validity and the effect of energy mis-reporting. *Proceedings of the Nutrition Society*, 599-609.

El-Saadani, S., Saleh, M., & Ibrahim, S. (2021). Quantifying non-communicable diseases' burden in Egypt using State-Space model. *Plos one*.

Estruch , R., Ros, E., Salas-Salvadö , J., Covas , M., Corella , D., & Arós , F. (2013). Primary prevention of cardiovascular disease with a Mediterranean diet. *N Engl J Med*, 368:1279–90.

Famoye, F. (1993). Restricted Generalized Poisson Regression Model. In *Communications in Statistics, Theory and Methods* (pp. 1335-1354).

Famoye, F. (2010). On the Bivariate Negative Binomial Regression Model. *Research Gate*, 969-981.

Famoye, F. (2010b). A new Bivariate Generalized Poisson Distribution. *Statistica Neerlandica*, 112-124.

Famoye, F. (2015). A Multivariate Generalized Poisson Regression Model. *Communications in Statistics-Theory and Methods*, 497-511.

Famoye, F., Okafor, R., & Adamu, M. O. (2011). A Multivariate generalized Poisson distribution. *J. Stat.Theory Appl*, 519-531.

FAO. (1997). *Agriculture, food and nutrition for Africa. A resource book for teachers of agriculture Food and Nutrition Division.* Rome, Italy: Food and Agriculture Organization (FAO).

FAO. (2002). *The State of Food Security in the world.* Italy.

FAO. (2004). *Globalization of food systems in developing countries: impact on food security and nutrition.* Rome.

FAO. (2010). *Guidelines for measuring household and individual dietary diversity.*

FAO/WHO. (2002). *Vitamin and mineral requirements in Human Nutrition.* Rome: FAO.

Gibson, R., Charrondiere, R., & Bell, W. (2017). Measurement Errors in Dietary Assessment Using Self-Reported 24-Hour Recalls in Low-Income Countries and Strategies for Their Prevention. *Advances in Nutrition*, 980-991.

Gleason, P. M., Boushey, C. J., Harris, J. E., & Zoellner, J. (2015). Publishing Nutrition Research: A Review of Multivariate Techniques- Part 3: Data Reduction Methods. *Science Direct*.

Global Panel on Agriculture and Food Systems. (2017). *Healthy diets for all: A Key to meeting the SDG's.* London, U.K.

Global Pattern. (2016). *Food Systems and Diets: Facing the Challenges of the 21st Century.* London, UK: Global Panel on Agriculture and Food Systems for Nutrition.

Guenther, P. M., Reedy, J., Krebs-Smith, S., & Reeve, B. (2008). Evaluation of the Healthy Eating Index-2005. *JM Diet Assoc*.

Guenther, P., Casavale, C., Reedy, J., Kirkpatrick, S., Hiza, H., & Kucyzynski, K. (2013). Update of the Healthy Eating Index: HEI 2010. *J Acad Nutr Diet*.

Gurmu, S., & Elder, J. (2000). Generalized Bivariate Count Data Regression Models. *Ecoonomic letters*, 31-36.

Gurmu, S., & Elder, J. (2008). A Bivariate Zero-Inflated Count Data Regression Model with Unrestricted Correlation. *Economics Letters*, 245-248.

Haines, P., Siega-Riz, A., & Popkin, B. (1999). The diet Quality Index Revised: A measurement instrument for populations. *J AM Diet Assoc*, 697-704.

Harris, T., Yang, Z., & Hardin, J. W. (2012). Modelling Underdispersed Count Data with Generalized Poisson Regression. *Stata Journal*, 736-747.

Henderson., V. (2002). *Urbanisation in Developing countries (English).* The World Bank Researcher Observer.

Henderson., V. (2010). Cities and Development. *Journal of Regional Science*, 89-112.

Hervé, A., & Michel, B. (2014). Correspondence Analysis. In *"Encyclopedia of Social Networks and Mining.".* New York: Springer Verlag.

Hilbe, J. M. (2014). *Modelling Count Data.* New York: Cambridge University Press.

Hillbrunner, C., & Egan, R. 2. (2008). Seasonality, Household food Security and nutritional Status in Dinajpur, Bangladesh. *Food and Nutrition Bulletin*, 221-231.

Hoax, J. J., Moerbeek, M., & Van De Schoot, R. (2017). Multilevel analysis: Techniques and Applications. *Routledge*.

Hoddinot, J., & Yohannes, Y. (2002). Dietary diversity as household food security indicator. *FoodcCons Nutr Div Discu Paper No. 136. International Food Policy Research Institute (IFPRI).* .

Hoffmann, K., Schulze, M. B., Schienkiewitz , A., Nothlings, U., & Boeing, H. (2004). Application of a new statistical method to derive dietary patterns in nutritional epidemiology. *Am J Epidemiol*, 935-944.

Holgate, P. (1964). Estimation for the bivariate Poisson distribution. *Biometrica*, 241-245.

Hox, J. J., Moerbeek, M., & Van De Schoot, R. (2017). Multilevel analysis: Techniques and applications. *Routledge*.

Hu, F. B. (2002). *Dietary pattern analysis: a new direction in nutritional epidemiology.* Havard School of Public Health, Department of Nutrition, Boston.

Hwang, Y., & Choe , Y. (2016). *The consumption pattern of convenience food: A comparison of different income levels in South Korea .* Boston, Massachusetts.

IBM knowledge center. (2018). *K-Means Cluster Analysis*. Retrieved from https://www.ibm.com/support/knowledgecenter/en/SSLVMB_22.0.0/com.ibm.spss.statistics.help/spss/base/idh_quic.htm

IFPRI. (2015). *Actions and Accountability to Advance Nutrition & Sustainable Development.* Washington, DC: International Food Policy Research Institute.

Islam , M. A., & Chowdhurry, R. I. (2017). A generalized right truncated bivariate Poisson regression model with applications to health data. *PLOS ONE*.

Islam, M. A., & Chowdhurry, R. I. (2015). A Bivariate Poisson models with Covariates Dependence. *Bulletin of Calcutta Mathematical society*, 11-20.

Jaadi, Z. (2021). *A step-by-step explanation of Principal Component Analysis (PCA) .* Retrieved from Builtin.

Jackson, P., Brembeck, H., Everts, J., Fuentes, M., Halkier, B., Hertz, F. D., . . . Wenzl, C. (2018). *A Short History of Convenience Food*. Retrieved 2 27, 2020, from https://link.springer.com/chapter/10.1007/978-3-319-78151-8_2

Jill , R., Amy , S. F., Stephanie , G. M., & Susan , K.-S. M. (2018). Extending Methods in Dietary Patterns Research. *MDPI*, 10.

Johnson, N. L., & Kotz, S. (1969). *Discrete Distributions.* New York: John Wiley and Sons.

Jung, W. (1993). Two aspects of Labor mobility: A Bivariate Poisson Regressioon approach. *Empirical Economics*, 543-556.

Kant, A. (1996). Indexes of overall diet quality: A review. *J Am Diet Assoc*.

Karan, R. (2021). *An Introduction to Principal Component Analysis.* Retrieved from naukri learning.

Karlis, D. (2003). An EM algorithm for multivariate Poisson distribution and related models. *Journal of Applied Statistics*, 63-77.

Karlis, D., & Meligkotsidou, L. (2007). Finite mixtures of Multivariate Poisson distributions with Application. *Science Direct. Journal of Statistical Planning ans Inference*, 1942-1960.

Karlis, D., & Ntzoufras, I. (2005). Bivariate Poisson and Diagonal inflated Bivariate Poisson Regression Models in R. *Journal of Statistical Software*.

Katz, D., Njike, V., Faridi, Z., Rhee, L., Reeves, R., & Jenkins, D. (2009). The stratification of foods on the basis of overall nutritional quality: The overall nutritional quality index. *Am J Health Promo*.

Kazembe, L. (2016). Bivariate copula-based regression to model timing and frequenccy of antenatal care utilization.

Kazembe, L. (2021). Social Vulnerability and Childhood Health: Bayesian Spatial Models to Assess Risks from Multiple Stressors on Childhood Diarrhoea in Malawi. *Springer*.

Kazembe, L., Nickanor, N., & Crush, J. (2021). Food Insecurity, Dietary Patterns and Non-Communicable Diseases in Windhoek, Namibia. *Journal of Hunger & Environmental Nutrition*.

Kennedy, E. T., Ohls, J., Carlson, S., & Fleming, K. (1995). The Healthy Eating Index: design and applications. *Pub Med*.

Kennedy, G. L. (2009). *Evaluation of Dietary Diversity Scores for assessment of Micronutrient intake and food security in Developing Countries.* Netherlands.

Kennedy, G., Fanou, N., Seghieri, C., & Brouwer, I. D. (2009). Dietary diversity as a measure of the micronutrient adequacy of women's diets: results from Bamako, Mali site. *Food and Nutrition Technical Assistance II Project*.

Kennedy, G., Ballard , T., & Dop, M. (2010). *Guidelines for measuring household and individual dietary diversity Rome: Nutrition and Consumer Protection Division, Food and Agriculture Organization of the United Nations.* Retrieved from Guidelines for measuring household and individual dietary diversity Rome: Nutrition and Consumer Protection Division, Food and Agriculture Organization of the United Nations

Kennedy, G., Pedro, M. R., Seghieri, C., Nantel, G., & Brouwer, I. (2007). Dietary Diversity Score Is a Useful Indicator of Micronutrient Intake in Non-Breast-Feeding Filipino Children. *the Journal of Nutrition*, 472-477.

Kennedy, N. G., & Shetty, P. (2004). Globalization of food systems in developing countries: impact on food security and nutrition. *Google Scholar*, 1-25.

Khan, S. (2013). Food, Nutrition, Diet and Non-Communicable diseases key reasons to consider NCDs in policies to adress major nutritional challenges. *Research Gate*.

Khorrami, Z., Rezapour, m., Etemad, K., Yarahmadi, S., Khodakarim, S., Hezaveh, A., . . . Khanjani, N. (2020). The patterns of Non-communicable disease Multimorbidity in Iran: A Multilevel Analysis. *Scientific reports*.

Kiers, H. A., & Smilde, A. K. (2007). A comparison of various methods for multivariate regression with highly collinear variables. *Stat. Meth. & Appl*, 193-228.

Kocherladota, S., & Kocherlakota , K. (1992). Bivariate Discrete Distributions. *Marcel Dekker*.

Kuczmarski, M., Brewer, B., Rawal, R., Ryan, P., Zonderman, A., & Evans, M. (2019). Aspects of Dietary Diversity Differ in Their Association with Atherosclerotic Cardiovascular Risk in a Racially Diverse US Adult Population. *Nutrients*, 1034.

Labadarios, D., Steyn, N. P., & Nel, J. (2011). How diverse is the diet of adult South African. *Nutritional Journal*, 10(33).

Labadarios, D., Steyn, N. P., & Nel, J. (2011). How diverse is the diet of adult South African? *Nutrition Journal, 10*(33).

Lang, T. (2003). Food industrialization and food power: implications for food governance. *Development Policy Rev*, 555-568.

Leiter, R. E., & Hamdan, M. A. (1973). Some Bivariate Probability Models Applicable to Traffic Accidents and Fatalities . *International Statitistical Review*, 87-100.

Leon, A. R., & Wu, B. (2011). Copula- based Regression models for bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 175-185.

Lim, S. S., Vos, T., Flaxman, A. D., Danaei, G., Shibuya, K., Adair-Rohani, H., . . . Aryee, M. (2012). A comparative risk assessment of burden of disease and injury attributable to 67 risk factors clusters in 21 regions, 1990-2010: A systematic analysis for the Global burden of Disease study 2010. *lancet*, 2224-2260.

Luan, H., Law, J., & Quick, M. (2015). Identifying Food Swamps based on relative healthy food access: a Spatio-temporal Bayesian Approach. *International Journal of Health Geographics*.

Mardalena, S., Purhadi, P., Purnomo, J., & Prastyo, D. (2021). Bivariate Poisson Inverse Gaussian Regression Model with Exposure Variable: Infant and Maternal Death Case Study . *Journal of Physics: Conference series*.

Marra, G., & Radice, R. (2017). A joint regression modeling framework for analyzing bivariate binary data in R. *De Gruyer Open*, 268-294.

Martinez Steel, E., Popkin, B. M., Swinburn, B., & Monteiro, C. A. (2017). The share of ultra-processed foods and the overall nutritional quality of diets in the US: Evidence from a nationally representative cross-sectional study. *Popul. Health metr*, 15.

Mbanya, J., Motala, A., & Sobngwi, E. (2010). Diabetes in Sub-Saharan Africa. *PubMed*, 2254-2266. doi:375(9733)

Mbongo, L. T. (2017). *Food Insecurity and Quality of Life in the informal settlements of Katutura, Windhoek, Namibia.* Master's Thesis., University of Namibia, Department of Statistics and Population Studies, Windhoek.

Mesbahzadeh, T., Miglietta, M., Mirakbari, M., Soleimani Sardoo, S., & Abdolhoseini, M. (2019). Joint Modeling of precipitation and Temperature Using copula Theory for current and future prediction under Climate Change Scenarios in Arid Lands (Case Study, Kerman Province, Iran). *Hindawi*.

Micha, R., Wallace, S., & Mozaffarian, D. (2010). Red and Processed Meat Consumption and Risk of Incident Coronary Heart Disease, stroke and Diabetes Mellitus: A systematic Review and Meta-Analysis. *American Heart Association Journal*.

Mila-Villarroel, R., Bach-Faig, A., Puig, J., Puchal, A., Farran, A., & Serra-Majem, L. (2011). Comparison and evaluation of the reliability of indexes of adherence to the Mediterranean diet. *Public Health Nutr*.

Ministry of Health and Social Services. (2000). *Food and Nutrition guidelines for Namibia- Food choices for a healthy life.* Windhoek, Namibia.

Ministry of Health and Social Services. (2013). Food and Nutrition guidelines for Namibia- Food choices for a healthy life.

Mirmiran , P., Azadbakht, L., Esmaillzadeh, A., & Azizi, F. (2014). Dietary diversity score in adolescents - a good indicator of the nutritional adequacy of diets: Tehran lipid and glucose study. *Asia Pac J Clin Nutr*, 56-60.

Moe, A. M., Sigrunn, S., Hopstock, L., Carlsen, M., Lovsletten, O., & Ytterstad, E. (2022). Identifying dietary patterns across age, educational level and physical activity level in a cross-sectional study: the Tromsø Studyin a cross-sectional study: the Tromso Study. *BMC Nutrition*.

Mullahy, J. (1986). Specicification and Testing of some Modified Count Data Models. *Journal of Econometrics*, 341-365.

Mullahy, J. (1986). Specification and Testing in Some Modified Count Data Models. *Journal of Econometrics*, 341-365.

Muoka, A. K., Ngesa, O. O., & Waititu, A. G. (2016). Statistical Models for Count Data. *Science Journal of Applied Mathematics and Statistics*, 256-262.

Murawska, M., Lesaffre, E., & Rizopoulos, D. (2012). A two-stage Joint model for Nonlinear Longitudinal Response and a Time-to-event with application in Transplantation studies. *Journal of Probability and Statistics*, 18.

National Center for Health Statistics. (2020). *USDA Food and Nutrition services*. Retrieved from Healthy Eating Index (HEI): https://www.fns.usda.gov/healthy-eating-index-hei

National Planning Commission of Namibia. (2016). *Namibia Zero Hunger.* Windhoek: Republic of Namibia.

National Research Council, F. a. (1989). Recommended Dietary Allowances. *National Academy Press*.

Nickanor, N. (2014). *Food deserts and household food insecurity in the informal settlements of WIndhoek, Namibia.* Windhoek.

Nieman, M. (2015). Statistical Analysis of Strategic Interaction with Unobserved Player Actions: Introducing a Strategic Probit with Partial Observability. *Political Analysis*, 429-448.

NSA. (2018). *Namibia Household Income and Expenditure Survey, 2015/16, Interviewer Manual.* Windhoek, Namibia.

Pan American Health Organization (PAHO). (2015). *Ultra- Processed Food and Drink Products in Latin America: Trends, Impact on obesity, Policy Implications.* Washington, DC, USA: PAHO.

Patterson, R. E., Haines, P. S., & Popkin, B. M. (1994). Diet quality index: Capturing a multidimensional behavior. *J AM Diet Assoc*, 57-64.

Patterson, R. E., Haines, P. S., Popkin, B. M., & . (1994). Diet Quality index: Capturing a multidimensional behaviour. *J AM Diet Assoc*, 57-64.

Pazvakawambwa, L., & Nickanor, N. (2018). A Zero-Truncated negative Binomial regression model for Dietary Diversity in Namibian under-5 children.

Pervaiz, B., Ninghui, L., Manzoor, M. Q., & Altangerel , O. (2017). Determinants of Household Food Security in Punjab – Pakistan: A Binary Logistic Regression Analysis. *World Applied Sciences Journal*, 1021-1028.

Pillai, A., Kinabo, J., & Krawinkel, M. (2016). Effect of nutrition education on the knowledge scores of urban. *Agriculture & Food Security*.

Population Health Methods. (2018). Retrieved from Cluster analysis using K-Means: https://www.mailman.columbia.edu/research/population-health-methods/cluster-analysis-using-k-means

Poskitt, C., & Zhao, X. (2019). The bivariate Probit model, maximum likelihood estimation, Pseudo true parameters and partial identification . *Journal of econometrics*, 94-113.

Rafferty, A., Anderson, J., McGee, H., & Miller, C. (2002). A healthy diet indicator: Quantifying Compliance with the dietary guidelines using the BRFSS. *prev med*.

Rahkovsky, I., Jo, Y., & Carlson, A. (2018, July 24). What Drives Consumers to Purchase Convenience Foods?

Rajendran, S., Afari-Sefa, V., Shee, A., Bocher, T., Bekunda, M., & dominick, I. (2017). oes crop diversity contribute to dietary diversity? Evidence from integration of vegetables into maize-based farming systems. *Agriculture & Food Security*.

Rawat, A. (2017). *Toward data science*. Retrieved from Binary Logistic Regression: An overview and implementation in R: https://towardsdatascience.com/implementing-binary-logistic-regression-in-r-7d802a9d98fe

Reinsel, G. (2006). Encylopedia statistical sciences. *Wiley-Interscience*, 7015-7028.

Ruel, M. (2003). Is dietary diversity an indicator of food security or dietary quality? A review of measurement issues and research needs. *Food Nutri Bull*, 231-232.

Ruel, M. (2003). Is dietary diversity an indicator of food security or dietary quality? A review of measurement issues and research needs. *Food Nutrition Bull*, 231-232.

SAS Institute Inc. (2004). *SAS/STAT 9.1 User's Guide, SAS Institute Inc.* Cary, NC, USA.

Savy, M. M.-P., Traissac, P., & Delpeuch, F. (2007). Measuring dietary diversity in rural Burkina Faso: comparison of a 1-day and a 3-day dietary recall. *Public Health Nutrition*, 71-78.

Savy, M., Martin-Prevel, Y., Sawadogo, P., Kameli, Y., & Delpeuch, F. (2005). Use of variety/diversity scores for diet quality measurement: relation with nutritional status of women in rural area in Burkina Faso. *European Journal of Clinical Nutrition*, 703-716.

Shalmali, G. (2007). Globalization. *Development in practice*, 523-531.

Sharifi, A. (2016). Partial Least Squares-Regression (PLS-regression) in Chemometrics. *ResearchGate*, 305-308.

Smith, A. D., Emmett, P. M., Newby, P. K., & Northstone , K. (2011). A comparison of dietary patterns derived by cluster and principle components analysis in a UK cohort of children. *Journal of clinical nutrition*, 1102-1109. doi:http://dx.doi.org/10.1038/ejcn.2011.96

Smith, A. D., Emmett, P. M., Newby, P. K., & Northstone, K. (2011). How diverse is the diet of Adult South African? *Nutrition*.

Ssempiira, J., Kasirye, I., Kissa, J., Nambuusi, B., Mukooyo, E., & Opigo, J. (2018). Measuring health facility readiness and its effects on severe malaria outcomes in Uganda. *Scientific report*.

Styen, N. P., Nel, J. H., Nantel, G., Kennedy, G., & Labadarios, D. (2006). Food Variety and Dietary Diversity Scores in children: are they good indicators of dietary adequacy. *Public Health Nutrition*, 644-650.

Sunil, N., & Gregory, Y. (2021). *Evaluation of Hypertension.*

Suresh, B. C., Shailendra, G. N., & Prabuddha, S. (2014). *Food Security, Poverty, and Nutrition Policy Analysis: Statistical Methods and Applications.* Academic Press.

Suresh, B. C., Shailendra, G. N., & Prabuddha, S. (2014). *Food Security, Poverty,and Nutrition Policy Analysis. Statistical Methods and Applications.* (Second Edition ed.). Oxford, UK.: Elsevier Inc.

Swindale, A., & Bilinsky, P. (2005). Household Dietary diversity Score (HDDS) for Measurement of Household food Access: Indicator Guide, Food and Nutrition Technical Assistance. *The Journal of Nutrition*, 2448-53.

Swindale, A., & Bilinsky, P. (2006). *Household Dietary Diversity Score (HDDS) for measurement of Household Food Access: Indicator Guide.* Washington, DC: FANTA.

Swindale, A., & Bilinsky, P. (2006). *Household Dietary Diversity Score (HDDS) for Measurement of Household Food Access: Indicator Guide (v.2).* Washington, DC: Academy for Educational Development.

Thiele, S., & Weiss, C. (2003). Consumer demand for food diversity: evidence for Germany. *Elsevier Science*, 99-115.

Thorpe, M., Milte, C., Crawford, D., & McNaughton, S. (2016). A comparison of the dietary patterns derived by principal component analysis and cluster analysis in older Australians. *Internationa Journal of Behaviourak Nutrition and Physical Activity*.

Trivedi, P., & Zimmer, D. (2007). Copula Modeling: An Introduction for Practitioners. *Foundations and Trends in Econometrics*, 1-111.

Tso, M. (1981). Reduced-Rank Regression and Canonical Analysis. *Journal of the Royal Statistical Society*, 183-189.

UEF. (n.d.). *Dietary patterns and Health*. Retrieved October 02, 2018, from Ravitsemusepidemiologit: https://www.uef.fi/web/nutritionepidemiologists/dietary-patterns

Ukegbu , O. P., & Ogu , C. V. (2017). Assessment of Dietary Diversity Score, Nutritional Status and Socio-Demographic characteristics of under-5 Children in Some Rural Areas of Imo State, Nigeria. *Mal J Nutr*, 425 - 435.

UNSCN. (2018). *Non-Communicable Diseases, Diets and Nutrition.*

Vakili, M., Abedi, P., Sharifi, M., & Hosseini, M. (2013). Dietary Diversity and Its Related Factors among Adolescents: A Survey in Ahvaz-Iran. *Global Journal of Health Science, 5*(2).

Vellema, W., Desiere, S., & D'Haese, M. (2015). Verifying Validity of the Household Dietary Diversity Score: An application of Rasch model. *Sage*, 27-41.

Von Braun , J., Meinzen-Dick, R., Rosegrant, M., & Nin-Pratt, A. (2008). International agricultural research for food security, poverty reduction, and the environment: What to expect from scaling CGIAR investments and "Best Bet" programs. *International Food Policy Research Institute (IFPRI)*.

WHO. (2021). *Noncommunicable diseases.*

WHO. (2023). *Global Status Report on Noncommunicable Diseases.* Geneva: World Health Organiztion.

Willet, W., Sacks, F., Trichopoulou, A., Drescher, G., Ferroluzzi, A., & Helsing, E. (1995). Mediterranean diet pyramid: A cultural model for healthy eating. *Am J Clin Nutr*.

Williams, B., Mancia, G., & Spiering, W. (2018). ESC/ESH guidelines for the management of arterial hypertension. *Eur Heart J*.

World Food Programme. (2019). *WFP NAMIBIA Country Brief.*

World's urban population increasingly urban with more than half living in urban areas. (2019). Retrieved from UN.org.

Yan, J. (2007). Enjoy the Joy of Copulas:with a package copula. *Statistical software*, 1-21.

Yunteng , L., Yao-Jan, W., Jonathan, C., & Yinhai, W. (2011). Modeling animal-vehicle collisions using diagonal inflated bivariate Poisson regression. *Elsevier*, 220-227.

Zaiontz, C. (2018). *Principal Component Analysis*. Retrieved from Real Statistics Using Excel: http://www.real-statistics.com/multivariate-statistics/factor-analysis/principal-component-analysis/

Zaiontz, C. (2018). Principal Component Analysis. *Real Statistics using Excel*.

Zhang, F., Tapera, T. M., & Gou, J. (2018). Application of a new dietary pattern analysis method in nutritional epidemiology. *BMC Med Res*.

Zhang, Y., Zhou, H., Zhou, J., & Sun, W. (2017). Regression Models for Multivariate Count Data. *Journal of Computational and Graphical Statistics*, 1-13.

Zhao, J., Li, Z., Gao, Q., Zhao, H., Chen, S., Huang, L., . . . Wang, T. (2021). A review of statistical methods for dietary pattern analysis. *Nutrition Journal*.

Zhong , T., Si, Z., Crush, J., Xu, Z., Huang, X., Scott , S., . . . Zhang , X. (2018). The Impact of Proximity to Wet Markets and Supermarkets on Household Dietary Diversity in Nanjing City, China. *MDPI*.

Zimmer, D. M., & Trivedi, P. K. (2006). Using trivariate copulas to model sample selection and treatment effects: Application to family health care demand. *Journal of Business Econoomic Statistics*, 63-76.

Appendix: R codes and Output

1. COUNT MODELS

```
> library(foreign)

# Choosing a file

> file.choose()

[1] "C:\\Users\\mbongol\\Desktop\\Selected variables from NHIES2015_2016.sav"

# Importing Spss file into R

> dataset=read.spss("C:\\Users\\mbongol\\Desktop\\Selected variables from
NHIES2015_2016.sav", to.data.frame= TRUE)

re-encoding from UTF-8

# function for Poisson model

> poi<-glm(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head +
main_sourceincome + attain +  q05_04_13 + q05_04_15 + q05_04_16, family= poisson,
data=dataset)

> summary(poi)
```

Call:

glm(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + ain_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family = poisson, data = dataset)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-3.9196	-0.6865	-0.0758	0.5819	2.961

	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	1.805081	0.075329	23.963	<2e-16	***
ah_hh_members4-6Members	-0.009892	0.019213	-0.515	0.606651	
ah_hh_members7-9Members	-0.034939	0.028456	-1.228	0.219509	
ah_hh_members>10Members	0.044192	0.038159	1.158	0.246815	
li_urbrurRural	-0.152107	0.019137	-7.948	1.89E-15	***
sex_of_headMale	-0.083149	0.017166	-4.844	1.27E-06	***
age_of_head20-29Years	0.071301	0.065635	1.086	0.277334	

age_of_head30-39Years	0.036612	0.065302	0.561	0.575035	
age_of_head40-49Years	0.018003	0.066157	0.272	0.785526	
age_of_head50-59Years	0.065324	0.067779	0.964	0.335157	
age_of_head60+	0.132326	0.072088	1.836	0.066415	.
main_sourceincomePension	-0.182287	0.041004	-4.446	8.77E-06	***
main_sourceincomeSubsistencefarming	-0.14303	0.028692	-4.985	6.19E-07	***
main_sourceincomeBusinessincome	-0.010948	0.030214	-0.362	0.7171	
main_sourceincomeRemittances/grants	-0.140615	0.028332	-4.963	6.94E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.441736	0.042702	-10.345	<2e-16	***
main_sourceincomeCommercialfarming	0.203178	0.14297	1.421	0.15528	
main_sourceincomeOthers	-0.169056	0.051211	-3.301	0.000963	***
attainPrimary	0.090684	0.02318	3.912	9.14E-05	***
attainSecondary	0.241126	0.024451	9.862	<2e-16	***
attainTertiary	0.451718	0.046705	9.672	<2e-16	***
attainNotstated	0.249043	0.092988	2.678	0.007401	**
q05_04_13No	-0.036948	0.017841	-2.071	0.03836	*
q05_04_15No	-0.117	0.023765	-4.923	8.52E-07	***
q05_04_16No	-0.052153	0.035593	-1.465	0.142845	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 3725.7 on 3036 degrees of freedom

Residual deviance: 2920.3 on 3012 degrees of freedom

(7053 observations deleted due to missingness)

> AIC: 13103

Number of Fisher Scoring iterations: 4

> BIC(poi)

[1] 13253.67

> logLik(poi)

'log Lik.' -6526.604 (df=25)

Function for quasi Poisson regression model

> > quasipoi<-glm(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family= quasipoisson, data=dataset)

> summary(quasipoi)

Call:

glm(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family = quasipoisson, data = dataset)

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-3.9196	-0.6865	-0.0758	0.5819	2.961

Coefficients:					
	Estimate	Std.Error	t value	Pr(>|t|)	
(Intercept)	1.805081	0.071443	25.266	<2e-16	***
ah_hh_members4-6Members	-0.00989	0.018222	-0.543	0.587263	
ah_hh_members7-9Members	-0.03494	0.026988	-1.295	0.19555	
ah_hh_members>10Members	0.044192	0.03619	1.221	0.222139	
li_urbrurRural	-0.15211	0.01815	-8.38	<2e-16	***
sex_of_headMale	-0.08315	0.016281	-5.107	3.47E-07	***
age_of_head20-29Years	0.071301	0.062249	1.145	0.252127	
age_of_head30-39Years	0.036612	0.061933	0.591	0.554466	
age_of_head40-49Years	0.018003	0.062744	0.287	0.774189	
age_of_head50-59Years	0.065324	0.064283	1.016	0.309614	
age_of_head60+	0.132326	0.068369	1.935	0.053027	.
main_sourceincomePension	-0.18229	0.038889	-4.687	2.89E-06	***
main_sourceincomeSubsistencefarming	-0.14303	0.027211	-5.256	1.57E-07	***
main_sourceincomeBusinessincome	-0.01095	0.028655	-0.382	0.702454	
main_sourceincomeRemittances/grants	-0.14062	0.02687	-5.233	1.78E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.44174	0.0405	-10.907	<2e-16	***
main_sourceincomeCommercialfarming	0.203178	0.135594	1.498	0.134127	
main_sourceincomeOthers	-0.16906	0.048569	-3.481	0.000507	***
attainPrimary	0.090684	0.021984	4.125	3.81E-05	***
attainSecondary	0.241126	0.02319	10.398	<2e-16	***
attainTertiary	0.451718	0.044295	10.198	<2e-16	***
attainNotstated	0.249043	0.088191	2.824	0.004775	**
q05_04_13No	-0.03695	0.01692	-2.184	0.029066	*
q05_04_15No	-0.117	0.022539	-5.191	2.23E-07	***
q05_04_16No	-0.05215	0.033756	-1.545	0.122456	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

(Dispersion parameter for quasipoisson family taken to be 0.8994856)	
Null deviance: 3725.7 on 3036 degrees of freedom Residual deviance: 2920.3 on 3012 degrees of freedom (7053 observations deleted due to missingness)	

> library(gamlss)

Loading required package: splines

Loading required package: gamlss.data

Attaching package: 'gamlss.data'

> library(gamlss); library(COUNT)

Function for negative binomial

> negativebino<-glm.nb(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data=dataset)

> summary(negativebino)

Call:

glm.nb(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data = dataset, init.theta = 61318.93169, link = log)

Deviance Residuals:				
Min	1Q	Median	3Q	Max
-3.9195	-0.6865	-0.0758	0.5819	2.96

Coefficients:					
	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	1.805082	0.075333	23.961	<2e-16	***
ah_hh_members4-6Members	-0.00989	0.019214	-0.515	0.606698	
ah_hh_members7-9Members	-0.03494	0.028457	-1.228	0.21954	
ah_hh_members>10Members	0.044193	0.038161	1.158	0.24683	
li_urbrurRural	-0.15211	0.019138	-7.948	1.90E-15	***
sex_of_headMale	-0.08315	0.017167	-4.843	1.28E-06	***
age_of_head20-29Years	0.071302	0.065638	1.086	0.277353	
age_of_head30-39Years	0.036612	0.065305	0.561	0.57505	

age_of_head40-49Years	0.018003	0.06616	0.272	0.785538	
age_of_head50-59Years	0.065324	0.067782	0.964	0.335182	
age_of_head60+	0.132326	0.072091	1.836	0.066427	.
main_sourceincomePension	-0.18229	0.041006	-4.445	8.77E-06	***
main_sourceincomeSubsistencefarming	-0.14303	0.028693	-4.985	6.20E-07	***
main_sourceincomeBusinessincome	-0.01095	0.030215	-0.362	0.717105	
main_sourceincomeRemittances/grants	-0.14062	0.028333	-4.963	6.94E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.44174	0.042705	-10.344	<2e-16	***
main_sourceincomeCommercialfarming	0.20318	0.142979	1.421	0.155302	
main_sourceincomeOthers	-0.16906	0.051214	-3.301	0.000964	***
attainPrimary	0.090685	0.023181	3.912	9.15E-05	***
attainSecondary	0.241126	0.024452	9.861	<2e-16	***
attainTertiary	0.451718	0.046708	9.671	<2e-16	***
attainNotstated	0.249044	0.092992	2.678	0.007404	**
q05_04_13No	-0.03695	0.017841	-2.071	0.038364	*
q05_04_15No	-0.117	0.023767	-4.923	8.53E-07	***
q05_04_16No	-0.05215	0.035595	-1.465	0.14287	***
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 (Dispersion parameter for Negative Binomial(61318.93) family taken to be 1) Null deviance: 3725.4 on 3036 degrees of freedom Residual deviance: 2920.1 on 3012 degrees of freedom (7053 observations deleted due to missingness) AIC: 13105 Number of Fisher Scoring iterations: 1 Theta: 61319, Std. Err.: 209531, 2 x log-likelihood: -13053.23					

> BIC(negativebino)

[1] 13261.72

> logLik(negativebino)

'log Lik.' -6526.617 (df=26)

> library(gamlss)

#modelling Poisson Inverse Gaussian Model

> PIGMOD<-gamlss(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family=PIG, data=na.omit(dataset))

GAMLSS-RS iteration 1: Global Deviance = 12768.12

GAMLSS-RS iteration 2: Global Deviance = 12838.8

GAMLSS-RS iteration 3: Global Deviance = 12805.21

GAMLSS-RS iteration 4: Global Deviance = 12805.2

GAMLSS-RS iteration 5: Global Deviance = 12805.2

Warning message:

> summary (PIGMOD)

Family: c("PIG", "Poisson.Inverse.Gaussian")

Call: gamlss(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family = PIG, data = na.omit(dataset))

Fitting method: RS()

Mu link function: log					
Mu Coefficients:					
(Intercept)	1.805081	0.083135	21.713	<2e-16	***
ah_hh_members4-6Members	-0.00989	0.020513	-0.482	0.629684	
ah_hh_members7-9Members	-0.03494	0.031482	-1.11	0.267175	
ah_hh_members>10Members	0.044192	0.04106	1.076	0.281885	
li_urbrurRural	-0.15211	0.01952	-7.792	8.98E-15	***
sex_of_headMale	-0.08315	0.018153	-4.58	4.83E-06	***
age_of_head20-29Years	0.071301	0.073653	0.968	0.333089	
age_of_head30-39Years	0.036612	0.073302	0.499	0.617488	
age_of_head40-49Years	0.018003	0.073895	0.244	0.807531	
age_of_head50-59Years	0.065324	0.075572	0.864	0.387438	
age_of_head60+	0.132326	0.081993	1.614	0.106661	
main_sourceincomePension	-0.18229	0.047382	-3.847	0.000122	***
main_sourceincomeSubsistencefarming	-0.14303	0.032103	-4.455	8.68E-06	***
main_sourceincomeBusinessincome	-0.01095	0.031316	-0.35	0.726669	
main_sourceincomeRemittances/grants	-0.14062	0.031491	-4.465	8.29E-06	***
main_sourceincomeDrought/in-kindreceipts	-0.44174	0.043423	-10.173	<2e-16	***

main_sourceincomeCommercialfarming	0.202957	0.266218	0.762	0.445898	
main_sourceincomeOthers	-0.16906	0.045657	-3.703	0.000217	***
attainPrimary	0.090685	0.024836	3.651	0.000265	***
attainSecondary	0.241126	0.025818	9.339	<2e-16	***
attainTertiary	0.451719	0.048141	9.383	<2e-16	***
attainNotstated	0.249043	0.105009	2.372	0.017772	*
q05_04_13No	-0.03695	0.019508	-1.894	0.058315	.
q05_04_15No	-0.117	0.026367	-4.437	9.43E-06	***
q05_04_16No	-0.05215	0.037877	-1.377	0.168649	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Sigma link function: log Sigma Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) -36.05033 0.03196 -1128 <2e-16 *** --- Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 No. of observations in the fit: 3037 Degrees of Freedom for the fit: 26 Residual Deg. of Freedom: 3011 at cycle: 5 Global Deviance: 12805.2 AIC: 12857.2 SBC: 13013.69					

> BIC(PIGMOD)

[1] 13013.69

> logLik(PIGMOD)

'log Lik.' -6402.6 (df=26)

> library(pscl)

#Classes and Methods for R developed in the Political Science Computational Laboratory Department of Political Science Stanford UniversitySimon Jackman

hurdle and zeroinfl functions by Achim Zeileis

> library(pscl)

> library(COUNT)

#Modelling Poisson hurdle models

> hurdlemod<-hurdle(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, dist= "poisson", zero.dist="binomial", link= "logit", data=dataset)

> summary(hurdlemod); AIC(hurdlemod);BIC(hurdlemod)

Call:

hurdle(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head +

main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data = dataset,

dist = "poisson", zero.dist = "binomial", link = "logit")

Pearson residuals:				
Min	1Q	Median	3Q	Max
-2.6897	-0.6419	-0.07266	0.59594	3.78787

Count model coefficients (truncated poisson with log link):					
	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	1.800001	0.076945	23.393	<2e-16	***
ah_hh_members4-6Members	-0.01987	0.019601	-1.014	0.31061	
ah_hh_members7-9Members	-0.05127	0.029232	-1.754	0.07946	.
ah_hh_members>10Members	0.042417	0.03898	1.088	0.27651	
li_urbrurRural	-0.16268	0.019447	-8.365	<2e-16	***
sex_of_headMale	-0.07218	0.017515	-4.121	3.78E-05	***
age_of_head20-29Years	0.070115	0.067107	1.045	0.29611	
age_of_head30-39Years	0.034727	0.066803	0.52	0.60317	
age_of_head40-49Years	0.020368	0.067697	0.301	0.76351	
age_of_head50-59Years	0.070345	0.0694	1.014	0.31077	
age_of_head60+	0.148523	0.073861	2.011	0.04434	*
main_sourceincomePension	-0.20353	0.042216	-4.821	1.43E-06	***
main_sourceincomeSubsistencefarming	-0.1614	0.02954	-5.464	4.66E-08	***
main_sourceincomeBusinessincome	0.003321	0.030522	0.109	0.91336	
main_sourceincomeRemittances/grants	-0.15002	0.028904	-5.19	2.10E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.45232	0.04565	-9.908	<2e-16	***
main_sourceincomeCommercialfarming	0.189815	0.143695	1.321	0.18652	
main_sourceincomeOthers	-0.15516	0.052276	-2.968	0.003	**
attainPrimary	0.101478	0.023933	4.24	2.23E-05	***
attainSecondary	0.248203	0.025152	9.868	<2e-16	***
attainTertiary	0.47086	0.047162	9.984	<2e-16	***

attainNotstated	0.247379	0.09456	2.616	0.00889	**
q05_04_13No	-0.03204	0.018202	-1.76	0.07841	.
q05_04_15No	-0.10728	0.024064	-4.458	8.27E-06	***
q05_04_16No	-0.0516	0.035878	-1.438	0.15038	

Zero hurdle model coefficients (binomial with logit link):					
	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	5.85773	1.56279	3.748	0.000178	***
ah_hh_members4-6Members	0.72592	0.39459	1.84	0.065819	.
ah_hh_members7-9Members	1.78842	1.03064	1.735	0.082698	.
ah_hh_members>10Members	0.10843	0.77159	0.141	0.888247	
li_urbrurRural	0.40322	0.343	1.176	0.239768	
sex_of_headMale	-1.07775	0.37959	-2.839	0.004522	**
age_of_head20-29Years	0.2429	1.08823	0.223	0.823376	
age_of_head30-39Years	0.09578	1.07507	0.089	0.929006	
age_of_head40-49Years	-0.1261	1.07546	-0.117	0.90666	
age_of_head50-59Years	-0.14245	1.10543	-0.129	0.897466	
age_of_head60+	-0.55669	1.19915	-0.464	0.642476	
main_sourceincomePension	1.25112	0.94197	1.328	0.184112	
main_sourceincomeSubsistencefarming	1.74819	1.05405	1.659	0.097206	.
main_sourceincomeBusinessincome	-0.84979	0.46478	-1.828	0.067493	.
main_sourceincomeRemittances/grants	0.61046	0.76143	0.802	0.422714	
main_sourceincomeDrought/in-kindreceipts	-0.96872	0.47113	-2.056	0.039768	*
main_sourceincomeCommercialfarming	13.4157	2352.108	0.006	0.995449	
main_sourceincomeOthers	-0.94345	0.64752	-1.457	0.145108	
attainPrimary	-0.21213	0.40327	-0.526	0.598861	
attainSecondary	0.24603	0.46665	0.527	0.598039	
attainTertiary	-0.47501	0.82354	-0.577	0.564081	
attainNotstated	12.71222	1276.419	0.01	0.992054	
q05_04_13No	-0.47553	0.36867	-1.29	0.197107	
q05_04_15No	-1.69197	0.88459	-1.913	0.055785	.
q05_04_16No	0.38335	1.23453	0.311	0.75616	
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Number of iterations in BFGS optimization: 30 Log-likelihood: -6479 on 50 Df [1] 13058 [1] NA					

library(pscl); library(COUNT)

> poi<-glm(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family= poisson, data=dataset)

Modelling Zero inflated Poisson regression

> zip<-zeroinfl(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data=dataset)

> summary(zip)

Call:

zeroinfl(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data = dataset)

Pearson residuals:				
Min	1Q	Median	3Q	Max
-2.66512	-0.64281	-0.07423	0.59132	3.58188

Count model coefficients (poisson with log link):					
	Estimate	Std.Error	zvalue	Pr(>|z|)	
(Intercept)	1.810338	0.076872	23.55	<2e-16	***
ah_hh_members4-6Members	-0.01766	0.019333	-0.913	0.36099	
ah_hh_members7-9Members	-0.04697	0.028489	-1.649	0.09924	.
ah_hh_members>10Members	0.037365	0.039134	0.955	0.33968	
li_urbrurRural	-0.16378	0.019499	-8.399	<2e-16	***
sex_of_headMale	-0.07379	0.01731	-4.263	2.02E-05	***
age_of_head20-29Years	0.072834	0.067154	1.085	0.2781	
age_of_head30-39Years	0.034086	0.066795	0.51	0.60984	
age_of_head40-49Years	0.021845	0.067725	0.323	0.74703	
age_of_head50-59Years	0.068785	0.069326	0.992	0.3211	
age_of_head60+	0.138099	0.073468	1.88	0.06015	.
main_sourceincomePension	-0.18504	0.041354	-4.475	7.66E-06	***
main_sourceincomeSubsistencefarming	-0.14682	0.028916	-5.078	3.82E-07	***
main_sourceincomeBusinessincome	0.002321	0.030634	0.076	0.9396	
main_sourceincomeRemittances/grants	-0.14439	0.028438	-5.078	3.82E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.42528	0.044054	-9.654	<2e-16	***
main_sourceincomeCommercialfarming	0.191113	0.142997	1.336	0.18139	
main_sourceincomeOthers	-0.14844	0.051673	-2.873	0.00407	**
attainPrimary	0.092613	0.023449	3.949	7.83E-05	***

attainSecondary	0.235476	0.024795	9.497	<2e-16	***
attainTertiary	0.46212	0.04695	9.843	<2e-16	***
attainNotstated	0.239576	0.093016	2.576	0.01001	*
q05_04_13No	-0.0335	0.018042	-1.857	0.06333	.
q05_04_15No	-0.10819	0.023817	-4.542	5.56E-06	***
q05_04_16No	-0.05181	0.035687	-1.452	0.14658	

Zero-inflation model coefficients (binomial with logit link):				
	Estimate	Std.Error	zvalue	Pr(>|z|)
(Intercept)	-5.52E+00	2.54E+00	-2.172	0.0299
ah_hh_members4-6Members	-9.90E-01	6.60E-01	-1.5	0.1336
ah_hh_members7-9Members	-1.35E+01	1.03E+03	-0.013	0.9896
ah_hh_members>10Members	-6.32E-01	2.13E+00	-0.297	0.7661
li_urbrurRural	-1.18E+00	6.73E-01	-1.752	0.0798
sex_of_headMale	1.17E+00	6.97E-01	1.684	0.0922
age_of_head20-29Years	3.53E-02	2.16E+00	0.016	0.987
age_of_head30-39Years	-4.06E-01	2.17E+00	-0.187	0.8514
age_of_head40-49Years	1.96E-01	2.17E+00	0.09	0.9282
age_of_head50-59Years	5.90E-02	2.21E+00	0.027	0.9788
age_of_head60+	3.93E-01	2.31E+00	0.17	0.8649
main_sourceincomePension	-8.32E-01	1.49E+00	-0.558	0.5771
main_sourceincomeSubsistencefarming	-2.01E+01	1.78E+04	-0.001	0.9991
main_sourceincomeBusinessincome	1.19E+00	6.41E-01	1.854	0.0637
main_sourceincomeRemittances/grants	-1.75E+01	9.05E+03	-0.002	0.9985
main_sourceincomeDrought/in-kindreceipts	1.18E+00	8.66E-01	1.362	0.1733
main_sourceincomeCommercialfarming	-1.34E+01	3.01E+03	-0.004	0.9964
main_sourceincomeOthers	1.40E+00	7.99E-01	1.755	0.0793
attainPrimary	8.28E-02	6.56E-01	0.126	0.8996
attainSecondary	-7.31E-01	8.11E-01	-0.901	0.3675
attainTertiary	4.98E-01	9.35E-01	0.533	0.5942
attainNotstated	-1.27E+01	1.63E+03	-0.008	0.9938
q05_04_13No	4.38E-01	6.09E-01	0.719	0.4722
q05_04_15No	1.61E+00	1.17E+00	1.376	0.1687
q05_04_16No	-6.18E-01	1.45E+00	-0.426	0.6704
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Number of iterations in BFGS optimization: 67 Log-likelihood: -6493 on 50 Df				

> AIC(zip); BIC(zip)

[1] 13086.93

[1] NA

> zip<-zeroinfl(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, data=dataset)

> zip<-zeroinfl(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, dist="poisson",data=dataset)

#Zero inflated negative binomial

> nb2<-glm.nb(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16,data=dataset)

> zipb<-zeroinfl(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, dist="negbin",data=dataset)

> summary(zipb)

Call:

zeroinfl(formula = HDDS ~ ah.hh.members + li.urbrur + sex.of.head + age.of.head + main.sourceincome + attain + HDDS2, data = MYDATA, dist = "negbin")

Pearson residuals:				
Min	1Q	Median	3Q	Max
-2.04076	-0.43925	0.03155	0.46207	2.15582

Count model coefficients (negbin with log link):					
	Estimate	Std.Error	z.value	Pr(>|z|)	
(Intercept)	1.24704	0.053205	23.439	<2e-16	***
ah.hh.members4-6Members	0.016709	0.009296	1.798	0.072254	.
ah.hh.members7-9Members	0.019189	0.013359	1.436	0.150875	
ah.hh.members>10Members	0.007236	0.018546	0.39	0.696391	
li.urbrurRural	-0.09416	0.009621	-9.787	<2e-16	***
sex.of.headMale	-0.00331	0.008346	-0.397	0.691723	
age.of.head20-29Years	0.074728	0.051881	1.44	0.149765	
age.of.head30-39Years	0.078743	0.051533	1.528	0.126513	

age.of.head40-49Years	0.081999	0.051675	1.587	0.112553	
age.of.head50-59Years	0.107617	0.051921	2.073	0.0382	*
age.of.head60+	0.147153	0.052809	2.787	0.005327	**
main.sourceincomePension	-0.08318	0.019025	-4.372	1.23E-05	***
main.sourceincomeSubsistencefarming	-0.06178	0.016281	-3.795	0.000148	***
main.sourceincomeBusinessincome	-0.02198	0.014497	-1.516	0.12944	
main.sourceincomeRemittances/grants	-0.05112	0.015935	-3.208	0.001338	**
main.sourceincomeDrought/in-kindreceipts	-0.1902	0.032107	-5.924	3.14E-09	***
main.sourceincomeCommercialfarming	0.100733	0.048672	2.07	0.038486	*
main.sourceincomeOthers	-0.03941	0.026092	-1.51	0.130954	
attainPrimary	0.045264	0.013752	3.291	0.000997	***
attainSecondary	0.110764	0.013873	7.984	1.41E-15	***
attainTertiary	0.202055	0.017068	11.838	<2e-16	***
attainNotstated	-0.00479	0.056898	-0.084	0.932901	stated
HDDS2HighDiversity	0.720018	0.010717	67.182	<2e-16	***
Log(theta)	37.90415	NA	NA	NA	

Zero-inflation model coefficients (binomial with logit link):					
	Estimate	Std.Error	z.value	Pr(>|z|)	
(Intercept)	-6.1242	4.1298	-1.483	0.1381	
ah.hh.members4-6Members	-0.3584	0.4879	-0.735	0.4626	
ah.hh.members7-9Members	-7.2306	17.7383	-0.408	0.6835	
ah.hh.members>10Members	-6.795	20.2086	-0.336	0.7367	
li.urbrurRural	-3.1224	3.6016	-0.867	0.386	
sex.of.headMale	1.285	0.6155	2.088	0.0368	*
age.of.head20-29Years	1.5253	4.083	0.374	0.7087	
age.of.head30-39Years	1.6003	4.0652	0.394	0.6938	
age.of.head40-49Years	2.0255	4.0607	0.499	0.6179	
age.of.head50-59Years	1.5829	4.085	0.387	0.6984	
age.of.head60+	3.0574	4.142	0.738	0.4604	
main.sourceincomePension	-2.1984	1.3409	-1.64	0.1011	
main.sourceincomeSubsistencefarming	-6.0568	17.9572	-0.337	0.7359	
main.sourceincomeBusinessincome	0.8155	0.5113	1.595	0.1107	
main.sourceincomeRemittances/grants	-4.7155	12.5343	-0.376	0.7068	
main.sourceincomeDrought/in-kindreceipts	0.796	1.0117	0.787	0.4314	
main.sourceincomeCommercialfarming	-16.0001	28616.8	-0.001	0.9996	
main.sourceincomeOthers	0.3879	1.2504	0.31	0.7564	
attainPrimary	0.2454	0.7754	0.316	0.7516	
attainSecondary	0.2717	0.7106	0.382	0.7022	
attainTertiary	1.6108	0.7523	2.141	0.0323	*

attainNotstated	-16.9942	13300.52	-0.001	0.999	
HDDS2HighDiversity	-19.4629	1162.386	-0.017	0.9866	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Theta = 28944185624462492
Number of iterations in BFGS optimization: 51
Log-likelihood: -1.95e+04 on 47

> AIC(zipb); BIC(zipb)

[1] 39084.86

[1] NA

fitting zero inflated Poisson regression

> zpig<-gamlss(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family=ZIPIG,data=na.omit(dataset))

> zpig<-gamlss(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family=ZIPIG,data=na.omit(dataset))

> summary(zpig)

Family: c("ZIPIG", "Zero inflated Poisson inverse Gaussian")

Call: gamlss(formula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family = ZIPIG,

data = na.omit(dataset))

Fitting method: RS()

Mu link function: log
Mu Coefficients:

	Estimate	Std.Error	tvalue	Pr(>|t|)	
(Intercept)	1.805886	0.078225	23.086	<2e-16	***
ah_hh_members4-6Members	-0.01246	0.01989	-0.626	0.53105	
ah_hh_members7-9Members	-0.03862	0.029346	-1.316	0.18821	
ah_hh_members>10Members	0.044661	0.039523	1.13	0.25857	
li_urbrurRural	-0.1585	0.019954	-7.943	2.76E-15	***

sex_of_headMale	-0.07809	0.017833	-4.379	1.23E-05	***
age_of_head20-29Years	0.072511	0.068043	1.066	0.28666	
age_of_head30-39Years	0.035353	0.0677	0.522	0.60157	
age_of_head40-49Years	0.017586	0.068584	0.256	0.79764	
age_of_head50-59Years	0.067639	0.07027	0.963	0.33585	
age_of_head60+	0.145246	0.074751	1.943	0.0521	.
main_sourceincomePension	-0.19249	0.042546	-4.524	6.29E-06	***
main_sourceincomeSubsistencefarming	-0.1494	0.029642	-5.04	4.93E-07	***
main_sourceincomeBusinessincome	-0.00377	0.031563	-0.119	0.90498	
main_sourceincomeRemittances/grants	-0.14555	0.029279	-4.971	7.03E-07	***
main_sourceincomeDrought/in-kindreceipts	-0.4526	0.04424	-10.231	<2e-16	***
main_sourceincomeCommercialfarming	0.192029	0.149054	1.288	0.19773	
main_sourceincomeOthers	-0.17235	0.053302	-3.233	0.00124	**
attainPrimary	0.095406	0.023959	3.982	6.99E-05	***
attainSecondary	0.243713	0.025391	9.599	<2e-16	***
attainTertiary	0.468843	0.049189	9.532	<2e-16	***
attainNotstated	0.246051	0.095771	2.569	0.01024	*
q05_04_13No	-0.03519	0.018467	-1.906	0.05681	.
q05_04_15No	-0.11147	0.024589	-4.533	6.03E-06	***
q05_04_16No	-0.05179	0.037033	-1.399	0.16205	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					

Sigma link function: log
Sigma Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -4.4748 0.2391 -18.71 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Nu link function: logit
Nu Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.0990 0.3271 -15.59 <2e-16 ***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

No. of observations in the fit: 3037
Degrees of Freedom for the fit: 27

```
     Residual Deg. of Freedom:  3010
                     at cycle:  20

Global Deviance:     13055.02
           AIC:     13109.02
           SBC:     13271.52
```

> logLik(zpig)

'log Lik.' -6527.508 (df=27)

> library("CompGLM")

Warning message:

package ‘CompGLM’ was built under R version 3.6.1

#Conway-Maxwell regression model

> maxwel<-glm.comp(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16, family=ZIPIG,data=dataset)

Error in fn(par, ...) : unused argument (family = ZIPIG)

> maxwel<-glm.comp(HDDS~ah_hh_members + li_urbrur + sex_of_head + age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 + q05_04_16,data=dataset)

> summary(maxwel); BIC(maxwel); AIC(maxwel)

Call:

glm.comp(lamFormula = HDDS ~ ah_hh_members + li_urbrur + sex_of_head +

   age_of_head + main_sourceincome + attain + q05_04_13 + q05_04_15 +

   q05_04_16, data = dataset)

Beta:					
	Estimate	Std.Error	t.value	p.value	
(Intercept)	2.013494	0.02788	72.2206	<2.2e-16	***
ah_hh_members4-6Members	-0.01088	0.098318	-0.1107	0.911895	
ah_hh_members7-9Members	-0.03834	0.020133	-1.9042	0.056982	.
ah_hh_members>10Members	0.048409	0.029817	1.6235	0.104581	
li_urbrurRural	-0.16713	0.039995	-4.1788	3.01E-05	***
sex_of_headMale	-0.09131	0.020503	-4.4533	8.77E-06	***
age_of_head20-29Years	0.078341	0.018135	4.3199	1.61E-05	***

age_of_head30-39Years	0.040201	0.068787	0.5844	0.55898	
age_of_head40-49Years	0.019762	0.068414	0.2889	0.77271	
age_of_head50-59Years	0.071707	0.069303	1.0347	0.300897	
age_of_head60+	0.145193	0.071021	2.0444	0.041003	*
main_sourceincomePension	-0.20005	0.075596	-2.6463	0.00818	**
main_sourceincomeSubsistencefarming	-0.15694	0.043234	-3.6301	0.000288	***
main_sourceincomeBusinessincome	-0.012	0.030301	-0.3961	0.692053	
main_sourceincomeRemittances/grants	-0.15445	0.031683	-4.8749	1.15E-06	***
main_sourceincomeDrought/in-kindreceipts	-0.48356	0.029942	-16.1496	<2.2e-16	***
main_sourceincomeCommercialfarming	0.22351	0.046149	4.8432	farming	1.34E-06
main_sourceincomeOthers	-0.18572	0.15018	-1.2366	0.216316	
attainPrimary	0.099348	0.053856	1.8447	0.065179	.
attainSecondary	0.264603	0.024389	10.8495	<2.2e-16	***
attainTertiary	0.496713	0.026451	18.7785	<2.2e-16	***
attainNotstated	0.273185	0.050627	5.3961	7.34E-08	***
q05_04_13No	-0.04055	0.097696	-0.415	0.678164	
q05_04_15No	-0.12858	0.018721	-6.8684	7.85E-12	***
q05_04_16No	-0.05758	0.02513	-2.2913	0.022016	*

Zeta:
Estimate Std.Error t.value p.value
(Intercept) 0.104346 0.037375 2.7919 0.005274 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

AIC: 13091.81
Log-Likelihood: -6519.904
[1] 13248.29
[1] 13091.81
> BIC(maxwel)
[1] 13248.29

2. BIVARIATE POISSON REGRESSION MODELS

Model 1: Bivariate Model regression: Constant only model

R code:

The function shows a data frame with only 2 outcomes, which is the first 2 columns in the dataset and mtype 2 for untruncated model

> mod1<-bpglm(datsss[,1:2], mtype=2)

Table 1: Parameter Estimates

Variable Name	Coeff.	S.E	t.value	P.value	Adj. S.E	Adj. P. Value
Y1: Constant	-2.170	0.047	-45.921	0	0.055	0.001
Y2: Constant	0.22	0.042	5.327	0	0.049	0.006
Log likelihood	-1960.897					
AIC	3925.795					
AICC	3925.8					
BIC	3938.343					
Deviance	3087.749					
Phi1 1.355		Phi2 1.374				
Overdispersion: Z-test for Y1 and Y2						
	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-7.511	0				
Z(Y2) Marginal 2-tail	-8.864	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	281.9474	8	0.0028			
T2 (Dispersion Test)	207.6422	8	0.0015			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	1.304	18	1			

Model 2: Bivariate Poisson Regression Model with Covariates

R code:

The function shows untruncated model with all the six covariates

> mod2<-bpglm(datasss[,1:2], datasss[,25:41],datasss[,25:41], mtype=2)

Var.Names	Coeff	s.e	t.value	p.value	Adj.s.e	Adj.p.value
Y1: Constant	-2.217530	0.616655	-3.596066	0.000327	0.718723	0.002048
Education_1	0.064146	0.238752	0.268671	0.788198	0.278270	0.817704
Education_2	0.014404	0.218863	0.065814	0.947529	0.255089	0.954972
Education_3	-0.354755	0.206917	-1.714481	0.086526	0.241165	0.141378
work_1	0.113262	0.220228	0.514295	0.607078	0.256680	0.659053
work_2	-0.070519	0.147531	-0.477990	0.632686	0.171951	0.681750
sex_1	0.107760	0.100006	1.077542	0.281310	0.116559	0.355279
marital_1	-0.101984	0.603523	-0.168980	0.865822	0.703417	0.884732
marital_2	-0.133556	0.608781	-0.219382	0.826365	0.709546	0.850709
marital_3	0.674196	0.605022	1.114332	0.265211	0.705165	0.339094
age_1	0.134280	0.223290	0.601371	0.547630	0.260249	0.605908
age_2	0.336444	0.221054	1.521999	0.128097	0.257643	0.191685
age_3	0.067153	0.231799	0.289701	0.772061	0.270166	0.803715
age_4	-0.041984	0.238207	-0.176250	0.860108	0.277634	0.879811
income_2	-0.006929	-0.271913	-0.025483	0.979671	0.316919	0.982558
income_3	-0.421350	0.395111	-1.066410	0.286310	0.460509	0.360271
income_4	-0.250100	0.524555	-0.476785	0.633544	0.611379	0.682509
income_5	-0.770878	1.016024	-0.758720	0.448069	1.184195	0.515106
Y2:Constant	-1.387332	1.021533	-1.358088	0.174521	1.205194	0.249756
Education_1	0.275728	0.234964	1.173489	0.240677	0.277208	0.319968
Education_2	0.214912	0.219029	0.981200	0.326560	0.258408	0.405648
Education_3	0.214912	0.219029	0.981200	0.326560	0.258408	0.405648
work_1	0.226881	0.194813	1.164606	0.244256	0.229838	0.323645
work_2	-0.087023	0.136859	-0.635861	0.524907	0.161465	0.589946
sex_1	0.162306	0.096007	1.690563	0.091007	0.113268	0.151962
marital_1	0.408028	1.037399	0.393318	0.694108	1.223912	0.738867
marital_2	0.842926	1.026995	0.820769	0.411832	1.211637	0.486667
marital_3	0.342578	1.036893	0.330389	0.741126	1.223315	0.779463
age_1	0.854303	0.279967	3.051444	0.002294	0.330302	0.009736
age_2	0.781349	0.272382	2.868584	0.004147	0.321353	0.015087

age_3	0.926213	0.262892	3.523164	0.000432	0.310157	0.002843
age_4	0.661617	0.258321	2.561224	0.010471	0.304764	0.030002
income_2	0.004596	0.238947	0.019234	0.984656	0.281906	0.986994
income_3	0.213540	0.329851	0.647383	0.517425	0.389155	0.583226
income_4	0.636659	0.372444	1.709410	0.087461	0.439405	0.147449
income_5	1.567736	0.657837	2.383166	0.017216	0.776109	0.043458

Log Likelihood	-1787.488
AIC	3646.976
AICC	3648.446
BIC	3870.246
Deviance	2805.341

Phi1 1.358	Phi2 1.391

Overdispersion: Z-test for Y1 and Y2						
	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-7.740	0				
Z(Y2) Marginal 2-tail	-8.836	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	249.4620	8	0.0022			
T2 (Dispersion Test)	182.1084	8	0.0037			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	8.120	-16	NaN			

Model 3: Bi-Variate Poisson Regression: Constant only model (No covariates)

R code:

The function shows right truncated model reduced model

> Mod3<-bpglm(datasss[,1:2], mtype =5)

Parameter estimates

Variable Name	Coeff.	S.E	t.value	P.value	Adj. S.E	Adj. P. Value

Y1: Constant	-2.169	0.047	-45.921	0	0.055	0
Y2: Constant	-1.944	0.042	-46.021	0	0.049	0
Log likelihood	-3237.18					
AIC	6478.359					
AICC	6478.364					
BIC	6490.908					
Deviance	6171.408					
Phi1 1.355		Phi2 1.374				
Overdispersion: Z-test for Y1 and Y2						
	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-7.511	0				
Z(Y2) Marginal 2-tail	-8.864	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	281.994	8	0.0028			
T2 (Dispersion Test)	207.642	8	0.0015			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	1994.069	18	0			

Model 4: Bivariate Poisson Regression model with covariates

R code:

The function shows right truncated model with six covariates

> mod4<- bpglm(datasss[,1:2], datasss[, 25:41], mtype=5)

Parameter estimate

Var.Names	Coeff	S.E	t.value	P.value	Adj. S.E	Adj. P.value
Y1:Constant	-2.21754	0.616663	-3.59603	0.000327	0.718727	0.002048
Education_1	0.064167	0.238763	0.268746	0.78814	0.27828	0.817652
Education_2	0.014428	0.218873	0.065918	0.947447	0.255099	0.954901
Education_3	-0.35477	0.206924	-1.71451	0.086522	0.241173	0.141369
work_1	0.113286	0.220253	0.514344	0.607043	0.256707	0.65902

work_2	-0.07052	0.147541	-0.478	0.63268	0.17196	0.681742
sex_1	0.107767	0.10001	1.077554	0.281305	0.116563	0.35527
marital_1	-0.102	0.60353	-0.169	0.865805	0.703421	0.884717
marital_2	-0.13355	0.608787	-0.21938	0.826371	0.709548	0.850713
marital_3	0.674285	0.605032	1.114462	0.265155	0.705171	0.339035
age_1	0.134292	0.2233	0.601399	0.547612	0.260258	0.605888
age_2	0.33649	0.221069	1.522107	0.12807	0.257658	0.19165
age_3	0.067158	0.231811	0.289711	0.772054	0.270178	0.803708
age_4	-0.04199	0.238217	-0.17627	0.860091	0.277645	0.879796
income_2	-0.00694	0.271926	-0.02551	0.979653	0.316932	0.982542
income_3	-0.42136	0.395115	-1.06643	0.286301	0.460511	0.360259
income_4	-0.25011	0.524563	-0.47679	0.63354	0.611384	0.682503
income_5	-0.77088	1.016064	-0.7587	0.448084	1.184233	0.515116
Y2:Constant	-3.86048	1.02499	-3.76635	0.000168	1.206857	0.001392
Education_1	0.454971	0.23311	1.951745	0.051046	0.274471	0.097479
Education_2	0.276696	0.218246	1.267818	0.204945	0.25697	0.281657
Education_3	0.093259	0.204633	0.455735	0.648608	0.240942	0.698736
work_1	0.246981	0.187644	1.316224	0.188182	0.220938	0.263694
work_2	-0.11118	0.134509	-0.82656	0.40854	0.158375	0.482722
sex_1	0.262367	0.090029	2.91426	0.003587	0.106003	0.013366
marital_1	0.794061	1.013437	0.783532	0.433366	1.193254	0.5058
marital_2	1.018147	1.014177	1.003914	0.315487	1.194125	0.393921
marital_3	1.487476	1.013926	1.467045	0.142451	1.193829	0.212856
age_1	0.675116	0.233401	2.892514	0.003845	0.274814	0.014071
age_2	0.854872	0.230801	3.70393	0.000215	0.271753	0.00167
age_3	0.738029	0.233167	3.165236	0.001562	0.274538	0.007216
age_4	0.431625	0.240605	1.793911	0.072911	0.283297	0.127701
income_2	0.0158	0.23862	0.066216	0.94721	0.280959	0.955156
income_3	-0.20886	0.318908	-0.65494	0.512552	0.375493	0.578081
income_4	0.383435	0.362459	1.057871	0.290185	0.426771	0.369003
income_5	0.29729	0.598836	0.496445	0.61961	0.705089	0.673317

Log Likelihood	-2948.823
AIC	5969.646
AICC	5969.646
BIC	6192.916
Deviance	5631.857

Phi1	Phi2
1.358	1.386

Overdispersion: Z-test for Y1 and Y2

	Z-Value	P-Value				

Z(Y1) Marginal 2-tail	-7.741	0				
Z(Y2) Marginal 2-tail	-8.837	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	248.796	8	0.0031			
T2 (Dispersion Test)	181.673	8	0.0045			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	1503.987	-16	NaN			

Model 5: Bivariate Poisson Regression: Constant only model

R code:

The function shows a data frame with only 2 outcomes, which is the 3rd and 4th columns in the dataset and mtype 2 for untruncated model

> mod5<-bpglm(datasss [, 3:4], mtype=2)

Parameter estimates

Variable Name	Coeff.	S.E	t.value	P.value	Adj. S.E	Adj. P. Value
Y1: Constant	-1.948	0.042	-46.062	0	0.052	0
Y2: Constant	0.784	0.029	27.419	0	0.037	0
Log likelihood	-2222.053					
AIC	4448.106					
AICC	4448.111					
BIC	3584.559					
Deviance	3584.559					
Phi1 1.523		Phi2 1.633				
Overdispersion: Z-test for Y1 and Y2						
	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-12.373	0				

Z(Y2) Marginal 2-tail	-22.138	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	226.753	8	0.0014			
T2 (Dispersion Test)	147.246	8	0.0073			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	3.717	18	1			

Model 6: Bivariate Poisson regression model with covariates

R codes:

The function shows untruncated model with six covariates

<-mod6<-bpglm(datasss[,3:4],datasss[,25:41],datasss=2)

Parameter estimates

Var.Names	Coeff	S.E	t.value	P.value	Adj.S.E	Adj.P.value
Y1:Constant	-2.22167	0.608249	-3.65256	0.000263	0.745802	0.002912
Education_1	0.108627	0.211842	0.512772	0.608142	0.259749	0.675827
Education_2	-0.01609	0.194526	-0.08273	0.93407	0.238517	0.946209
Education_3	-0.3361	0.182439	-1.84225	0.06552	0.223697	0.133062
work_1	0.186944	0.191801	0.974676	0.329786	0.235176	0.426718
work_2	-0.03814	0.131196	-0.29071	0.77129	0.160865	0.812598
sex_1	0.11217	0.089648	1.25123	0.210931	0.109922	0.307579
marital_1	0.109983	0.59796	0.18393	0.854079	0.733186	0.880768
marital_2	0.056548	0.602259	0.093893	0.925199	0.738457	0.938965
marital_3	0.898547	0.59896	1.500179	0.133655	0.734412	0.221223
age_1	0.090103	0.198163	0.454692	0.649358	0.242977	0.710786
age_2	0.346109	0.194357	1.780787	0.075031	0.23831	0.146492
age_3	0.043449	0.203937	0.213052	0.831298	0.250056	0.862065
age_4	-0.1013	0.210898	-0.48034	0.631012	0.258592	0.695265
income_2	0.200827	0.221393	0.907107	0.364411	0.271459	0.459467

income_3	-0.20581	0.318305	-0.64658	0.517942	0.390288	0.597996
income_4	-0.25209	0.468677	-0.53788	0.590695	0.574666	0.660924
income_5	0.10814	0.599098	0.180505	0.856766	0.73458	0.882972
Y2:Constant	-0.02085	0.518551	-0.04021	0.967924	0.660865	0.974829
Education_1	0.195818	0.155583	1.258614	0.208251	0.198281	0.323425
Education_2	0.273251	0.144933	1.885367	0.059461	0.184708	0.13913
Education_3	0.375575	0.137313	2.735187	0.006265	0.174997	0.031925
work_1	0.140687	0.130398	1.07891	0.2807	0.166184	0.397289
work_2	-0.16816	0.096264	-1.74684	0.08075	0.122683	0.170563
sex_1	0.073974	0.06341	1.16659	0.243453	0.080813	0.360058
marital_1	-0.0371	0.529641	-0.07005	0.944158	0.674998	0.956169
marital_2	0.235375	0.522344	0.450613	0.652296	0.665698	0.723677
marital_3	-0.14439	0.529742	-0.27258	0.785195	0.675127	0.830655
age_1	0.626819	0.176891	3.54353	0.0004	0.225438	0.005456
age_2	0.431884	0.174832	2.470274	0.013547	0.222814	0.052663
age_3	0.673414	0.167537	4.019501	0.00006	0.213516	0.001624
age_4	0.504269	0.166319	3.031931	0.002447	0.211965	0.01741
income_2	-0.13576	0.168608	-0.80517	0.420775	0.214882	0.52757
income_3	0.157354	0.208822	0.75353	0.45118	0.266132	0.554382
income_4	0.743584	0.255744	2.90753	0.003665	0.325932	0.022582
income_5	1.037496	0.35	2.964271	0.003054	0.446056	0.020077

Log likelihood	-2011.454
AIC	4094.909
AICC	4096.379
BIC	4318.178
Deviance	3230.644

Phi1	Phi2
1.503	1.624

Overdispersion: Z-test for Y1 and Y2

	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-12.171	0				
Z(Y2) Marginal 2-tail	-21.195	0				

Goodness of fit (T1) Overdispersion (T2):

	Chi-Square	d.f	P.Value			
T1 (GOF)	198.208	8	0.0015			

T2 (Dispersion Test)	129.562	8	0.0034			
Pearson Chi-square GOF using predicted probability						
	Chi. Square	d.f	P.Value			
	9.042	-16	NaN			

Model 7: Bivariate Poisson regression: Constant only (no covariates)

R code:

The function shows a data frame with only 2 outcomes, which is the 3rd and 4th columns in the dataset and mtype 2 for right truncated model

> model7<-bpglm(datasss[,3:4],mtype=5)

Parameter estimates

Variable Name	Coeff.	S.E	t.value	P.value	Adj. S.E	Adj. P. Value
Y1: Constant	-1.948	0.042	-46.062	0	0.052	0
Y2: Constant	-1.161	0.029	-40.403	0	0.037	0
Log likelihood	-4786.915					
AIC	9577.829					
AICC	9577.835					
BIC	9590.378					
Deviance	8439.631					
Phi1 1.523		Phi2 1.627				
Overdispersion: Z-test for Y1 and Y2						
	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-12.373	0				
Z(Y2) Marginal 2-tail	-22.138	0				
Goodness of fit (T1) Overdispersion (T2):						
	Chi-Square	d.f	P.Value			
T1 (GOF)	227.129	8	0.0012			
T2 (Dispersion Test)	147.246	8	0.0074			
Pearson Chi-square GOF using predicted probability						

	Chi. Square	d.f	P.Value			
	2319.843	18	0			

Model 8: Bivariate Poisson Regression Model with covariates

R code:

The function shows right truncated model with six covariates

> mod8<-bpglm(datasss[,3:4],datasss[,25:41],datasss[,25:41],mtype=5)

Parameter estimates

Var.Names	Coeff	S.E	t.value	P.value	Adj.S.E	Adj.P.value
Y1:Constant	-2.22168	0.608262	-3.6525	0.000263	0.745803	0.002912
Education_1	0.108675	0.211871	0.51293	0.608032	0.25978	0.675727
Education_2	-0.01606	0.194553	-0.08256	0.934207	0.238545	0.94632
Education_3	-0.33616	0.18246	-1.84236	0.065505	0.223719	0.133033
work_1	0.187039	0.191859	0.974876	0.329687	0.235243	0.426614
work_2	-0.03814	0.131216	-0.2907	0.771299	0.160887	0.812603
sex_1	0.112188	0.089658	1.251287	0.210911	0.109932	0.307549
marital_1	0.109953	0.59797	0.183876	0.854121	0.733185	0.8808
marital_2	0.05655	0.602267	0.093896	0.925197	0.738453	0.938962
marital_3	0.89878	0.598975	1.500531	0.133564	0.734417	0.221107
age_1	0.090128	0.198187	0.454764	0.649307	0.243001	0.710737
age_2	0.346212	0.194392	1.780997	0.074997	0.238349	0.146437
age_3	0.04345	0.203967	0.213026	0.831319	0.250088	0.86208
age_4	-0.10134	0.210924	-0.48047	0.630922	0.258619	0.695184
income_2	0.200884	0.221445	0.90715	0.364388	0.271519	0.459438
income_3	-0.20583	0.318319	-0.64663	0.517913	0.390298	0.597964
income_4	-0.25212	0.468697	-0.53791	0.590675	0.57468	0.660902
income_5	0.108221	0.599211	0.180605	0.856688	0.734707	0.882905
Y2:Constant	-2.37561	0.522024	-4.55076	0.000006	0.662564	0.000341
Education_1	0.386417	0.158772	2.433784	0.01499	0.201517	0.055248
Education_2	0.309772	0.147877	2.094787	0.036259	0.187689	0.098938
Education_3	0.063782	0.138978	0.458938	0.646306	0.176394	0.717679
work_1	0.292146	0.126979	2.300738	0.021463	0.161165	0.069959
work_2	-0.15531	0.093682	-1.65784	0.097437	0.118903	0.191573
sex_1	0.164376	0.061227	2.684712	0.007292	0.07771	0.034478
marital_1	0.367703	0.512135	0.717981	0.472815	0.650013	0.571642
marital_2	0.516187	0.513122	1.005972	0.314496	0.651266	0.428069
marital_3	1.077787	0.512924	2.101262	0.035687	0.651014	0.097899

age_1	0.501014	0.14797	3.385911	0.000717	0.187807	0.007671
age_2	0.582839	0.147065	3.963142	0.000075	0.186658	0.001807
age_3	0.499604	0.148895	3.355419	0.000801	0.18898	0.008236
age_4	0.216982	0.153627	1.412397	0.157919	0.194986	0.265866
income_2	0.030105	0.167818	0.179393	0.85764	0.212998	0.887609
income_3	0.017108	0.203213	0.084188	0.932911	0.257923	0.947118
income_4	0.446453	0.252779	1.766179	0.07745	0.320832	0.164146
income_5	0.856331	0.326866	2.619825	0.008834	0.414865	0.039078

Log likelihood	-4346.053
AIC	8764.107
AICC	8765.577
BIC	8987.376
Deviance	7606.437

Phi1	Phi2
1.503	1.611

Overdispersion: Z-test for Y1 and Y2

	Z-Value	P-Value				
Z(Y1) Marginal 2-tail	-12.171	0				
Z(Y2) Marginal 2-tail	-21.194	0				

Goodness of fit (T1) Overdispersion (T2):

	Chi-Square	d.f	P.Value			
T1 (GOF)	198.235	8	0.0015			
T2 (Dispersion Test)	129.405	8	0.0037			

Pearson Chi-square GOF using predicted probability

	Chi. Square	d.f	P.Value			
	1770.581	-16	NaN			

3. GENERALIZED POISSON REGRESSION MODELLING (GJRM)

```
> mode1<-
Insecurity~as.factor(Sx_1)+as.factor(Sx_2)+as.factor(Ms_1)+as.factor(MS_2)+as.factor(MS_3)
+as.factor(MS_4)+as.factor(MS_5)+as.factor(MS_6)+as.factor(MS_7)+as.factor(ED_1)+as.factor(ED_2)+as.factor(ED_3)+as.factor(ED_4)+as.factor(Work_1)+as.factor(Work_2)+as.factor(Work_3)+as.factor(Work_4)+
as.factor(Tenure_1)+as.factor(Tenure_2)+as.factor(Tenure_3)+as.factor(Structure_1)+as.factor(Structure_2)+as.factor(Structure_3)+as.factor(Structure_4)+as.factor(Structure_5)+as.factor(Water_1)+as.factor(Water_2)+as.factor(Electricity_1)+as.factor(Electricity_2)+as.factor(Toilet_1)+as.factor(Toilet_2)

> model2<-
MIFP~as.factor(Sx_1)+as.factor(Sx_2)+as.factor(Ms_1)+as.factor(MS_2)+as.factor(MS_3)+as.factor(MS_4)+as.factor(MS_5)+as.factor(MS_6)+as.factor(MS_7)+as.factor(ED_1)+as.factor(ED_2)+as.factor(ED_3)+as.factor(ED_4)+as.factor(Work_1)+as.factor(Work_2)+as.factor(Work_3)+as.factor(Work_4)+
as.factor(Tenure_1)+as.factor(Tenure_2)+as.factor(Tenure_3)+as.factor(Structure_1)+as.factor(Structure_2)+as.factor(Structure_3)+as.factor(Structure_4)+as.factor(Structure_5)+as.factor(Water_1)+as.factor(Water_2)+as.factor(Electricity_1)+as.factor(Electricity_2)+as.factor(Toilet_1)+as.factor(Toilet_2)

> f.list<-list(modz1,modz2)

> mr<-c("probit","PO")

> bpN<-gjrm(f.list,data=PhD,Model="B",margins=mr)

# “B” for the bivariate model with or without endogenous variable

> bpF<-gjrm(f.list,data=PhD,BivD="F",Model="B",margins=mr)

# BivD="F” denoting bivariate distribution for Frank copula

> bpC0<-gjrm(f.list,data=PhD,BivD="C0",Model="B",margins=mr)

# BivD="CO” denoting bivariate distribution for Clayton copula

> bpC180<-gjrm(f.list,data=PhD,BivD="C180",Model="B",margins=mr)

# BivD="C180” denoting bivariate distribution for Survival Clayton

> bpG0<-gjrm(f.list,data=PhD,BivD="G0",Model="B",margins=mr)

# BivD="G0” denoting bivariate distribution for Gumble

> bpG180<-gjrm(f.list,data=PhD,BivD="G180",Model="B",margins=mr)

# BivD="G180” denoting bivariate distribution for Survival Gumble
```

> AIC(bpN,bpF,bpG0,bpG180)

	d.f	AIC	
bpN	65	2072.708	
bpF	65	2074.352	
bpG0	65	2108.451	
bpG180	65	2108.451	

> Model3<-
glm(MIFP~Sx_1+Sx_2+Ms_1+MS_2+MS_3+MS_4+MS_5+MS_6+MS_7+ED_1+ED_2+ED_3+ED_4+Work_1+Work_2+Work_3+Work_4+Tenure_1+Tenure_2+Tenure_3+Structure_1+Structure_2+Structure_3+Structure_4+Structure_5+Water_1+Water_2+Electricity_1+Electricity_2+Toilet_1+Toilet_2, data=PHD,family=poisson)

> summary(Model3)

Call:

glm(formula = MIFP ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2, family = poisson, data = PHD)

	Estimate
(Intercept)	1.94E+14
Sx_1Male	3.00E-01
Sx_2Female	NA
Ms_1Married (living with spouse)	-5.40E-02
MS_2q16=Married not (living with spouse)	-4.90E-01
MS_3q16=Not married (living with partner)	4.62E-01
MS_4Going steady (in a relationship)	-5.45E-01
MS_5Single (not in a relationship)	1.22E-02
MS_6Divorced / separated	2.48E-01
MS_7Widower/Widow	NA
ED_1None	7.98E-01
ED_2Primary	3.88E-01
ED_3Secondary	-2.63E-02
ED_4Tertiary	NA
Work_1Working full – time	-2.23E-01
Work_2Working part-time/Casual	-1.54E-01
Work_3Not working - looking	-6.07E-02
Work_4Not working- not looking	NA
Tenure_1Owner/Purchaser/Family/Accommodation	1.34E+00
Tenure_2Tenant/Lodger	1.24E+00
Tenure_3Tied accommodation	NA
Structure_1Female centered (No husband/male partner in household, may include relatives, children, friends)	-1.94E+14
Structure_2Male centered (No wife/female parner in household, may include relatives, children, friends)	-1.94E+14
Structure_3Nuclear (Husband/male partner wife/female partner with or without children)	-1.94E+14
Structure_4Extended (Husband/male partner and wife/female partner and children and relatives)	-1.94E+14
Structure_5Under 18 headed households male centered	-1.94E+14
Water_1No piped water- private	-1.60E-01
Water_2Piped Water – Private	NA
Electricity_1No electricity	-3.51E-01
Electricity_2Electricity	NA
Toilet_1No flush toilet	-1.41E-01
Toilet_2Flush toilet	NA
	Std. Error
(Intercept)	7.68E+13
Sx_1Male	9.69E-02
Sx_2Female	NA
Ms_1Married (living with spouse)	2.23E-01
MS_2q16=Married not (living with spouse)	2.89E-01

MS_3q16=Not married (living with partner)	2.13E-01
MS_4Going steady (in a relationship)	2.20E-01
MS_5Single (not in a relationship)	2.04E-01
MS_6Divorced / separated	3.40E-01
MS_7Widower/Widow	NA
ED_1None	2.93E-01
ED_2Primary	2.84E-01
ED_3Secondary	2.92E-01
ED_4Tertiary	NA
Work_1Working full – time	1.32E-01
Work_2Working part-time/Casual	1.30E-01
Work_3Not working - looking	1.24E-01
Work_4Not working- not looking	NA
Tenure_1Owner/Purchaser/Family/Accommodation	6.82E-01
Tenure_2Tenant/Lodger	6.97E-01
Tenure_3Tied accommodation	NA
Structure_1Female centered (No husband/male partner in household, may include relatives, children, friends)	7.68E+13
Structure_2Male centered (No wife/female parner in household, may include relatives, children, friends)	7.68E+13
Structure_3Nuclear (Husband/male partner wife/female partner with or without children)	7.68E+13
Structure_4Extended (Husband/male partner and wife/female partner and children and relatives)	7.68E+13
Structure_5Under 18 headed households male centered	7.68E+13
Water_1No piped water- private	1.13E-01
Water_2Piped Water – Private	NA
Electricity_1No electricity	2.06E-01
Electricity_2Electricity	NA
Toilet_1No flush toilet	5.78E-01
Toilet_2Flush toilet	NA
	z value
(Intercept)	2.53
Sx_1Male	3.101
Sx_2Female	NA
Ms_1Married (living with spouse)	-0.242
MS_2q16=Married not (living with spouse)	-1.693
MS_3q16=Not married (living with partner)	2.168
MS_4Going steady (in a relationship)	-2.474
MS_5Single (not in a relationship)	0.06
MS_6Divorced / separated	0.729
MS_7Widower/Widow	NA
ED_1None	2.722

ED_2Primary	1.366
ED_3Secondary	-0.09
ED_4Tertiary	NA
Work_1Working full – time	-1.693
Work_2Working part-time/Casual	-1.183
Work_3Not working - looking	-0.49
Work_4Not working- not looking	NA
Tenure_1Owner/Purchaser/Family/Accommodation	1.965
Tenure_2Tenant/Lodger	1.785
Tenure_3Tied accommodation	NA
Structure_1Female centered (No husband/male partner in household, may include relatives, children, friends)	-2.53
Structure_2Male centered (No wife/female parner in household, may include relatives, children, friends)	-2.53
Structure_3Nuclear (Husband/male partner wife/female partner with or without children)	-2.53
Structure_4Extended (Husband/male partner and wife/female partner and children and relatives)	-2.53
Structure_5Under 18 headed households male centered	-2.53
Water_1No piped water- private	-1.416
Water_2Piped Water – Private	NA
Electricity_1No electricity	-1.7
Electricity_2Electricity	NA
Toilet_1No flush toilet	-0.244
Toilet_2Flush toilet	NA
	Pr(>|z|)
(Intercept)	0.01141
Sx_1Male	0.00193
Sx_2Female	NA
Ms_1Married (living with spouse)	0.80897
MS_2q16=Married not (living with spouse)	0.09038
MS_3q16=Not married (living with partner)	0.03018
MS_4Going steady (in a relationship)	0.01335
MS_5Single (not in a relationship)	0.95221
MS_6Divorced / separated	0.46598
MS_7Widower/Widow	NA
ED_1None	0.00649
ED_2Primary	0.17206
ED_3Secondary	0.92812
ED_4Tertiary	NA
Work_1Working full – time	0.09036
Work_2Working part-time/Casual	0.23691
Work_3Not working - looking	0.62405

Work_4Not working- not looking	NA
Tenure_1Owner/Purchaser/Family/Accommodation	0.04946
Tenure_2Tenant/Lodger	0.07431
Tenure_3Tied accommodation	NA
Structure_1Female centered (No husband/male partner in household, may include relatives, children, friends)	0.01141
Structure_2Male centered (No wife/female parner in household, may include relatives, children, friends)	0.01141
Structure_3Nuclear (Husband/male partner wife/female partner with or without children)	0.01141
Structure_4Extended (Husband/male partner and wife/female partner and children and relatives)	0.01141
Structure_5Under 18 headed households male centered	0.01141
Water_1No piped water- private	0.15689
Water_2Piped Water – Private	NA
Electricity_1No electricity	0.08912
Electricity_2Electricity	NA
Toilet_1No flush toilet	0.80722
Toilet_2Flush toilet	NA
(Intercept)	*
Sx_1Male	**
Sx_2Female	
Ms_1Married (living with spouse)	
MS_2q16=Married not (living with spouse)	.
MS_3q16=Not married (living with partner)	*
MS_4Going steady (in a relationship)	*
MS_5Single (not in a relationship)	
MS_6Divorced / separated	
MS_7Widower/Widow	
ED_1None	**
ED_2Primary	
ED_3Secondary	
ED_4Tertiary	
Work_1Working full – time	.
Work_2Working part-time/Casual	
Work_3Not working - looking	
Work_4Not working- not looking	
Tenure_1Owner/Purchaser/Family/Accommodation	*
Tenure_2Tenant/Lodger	.
Tenure_3Tied accommodation	
Structure_1Female centered (No husband/male partner in household, may include relatives, children, friends)	*

Structure_2Male centered (No wife/female parner in household, may include relatives, children, friends)	*
Structure_3Nuclear (Husband/male partner wife/female partner with or without children)	*
Structure_4Extended (Husband/male partner and wife/female partner and children and relatives)	*
Structure_5Under 18 headed households male centered	*
Water_1No piped water- private	
Water_2Piped Water – Private	
Electricity_1No electricity	.
Electricity_2Electricity	
Toilet_1No flush toilet	
Toilet_2Flush toilet	

> AIC,BIC(Model3)

Error: unexpected ',' in "AIC,"

> BIC(Model3)

[1] 1838.909

> AIC(Model3)

[1] 1752.019

4. STRUCTURAL EQUATION MODELLING

#Write out the measurement model and the structurw model

> sema<-'fs=~A + food_Type_2 + food_Type_3 + food_Type_4 + B + food_Type_6 + food_Type_7 + food_Type_8 + food_Type_9 + C + D + food_Type_12

+ ncd1=~ncd_1

+ ncd2=~ncd_2

+ ncd3=~ncd_3

+ ncd4=~ncd_4

+ ncd5=~ncd_5

+ ncd6=~ncd_6

+ #regression

+ ncd1~fs

+ ncd2~fs

+ ncd3~fs

+ ncd4~fs

+ ncd5~fs

+ ncd6~fs'

fit SEM using sem function in the “lavaan” package

> library(lavaan)

> sema1<-sem(sema,data=SEMD)

> summary(sema1,fit.measures=TRUE,standardized=TRUE)

Latent Variables:						
	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
FS =~						
A	1				0.038	0.206
food_Type_2	7.975	0.435	18.326	0	0.301	0.629
food_Type_3	5.303	0.31	17.124	0	0.2	0.405
food_Type_4	7.829	0.427	18.344	0	0.295	0.636
B	5.443	0.308	17.689	0	0.205	0.481
food_Type_6	7.179	0.393	18.276	0	0.271	0.613
food_Type_7	2.718	0.203	13.4	0	0.103	0.205

food_Type_8	2.537	0.168	15.063	0	0.096	0.265
food_Type_9	7.263	0.402	18.056	0	0.274	0.553
C	4.45	0.257	17.309	0	0.168	0.427
D	6.071	0.339	17.905	0	0.229	0.52
food_Type_12	8.154	0.446	18.284	0	0.308	0.616
ncd1 =~						
ncd_1	1				0.087	1
ncd2 =~						
ncd_2	1				0.253	1
ncd3 =~						
ncd_3	1				0.044	1
ncd4 =~						
ncd_4	1				0.095	1
ncd5 =~						
ncd_5	1				0.12	1
ncd6 =~						
ncd_6	1				0.331	1
Regressions:						
	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
ncd1 ~						
FS	-0.042	0.026	-1.643	0.1	-0.018	-0.018
attain	0.005	0.001	5.905	0	0.06	0.06
li_urbrur	0	0	-2.544	0.011	-0.001	-0.026
q04_20	-0.005	0.003	-1.623	0.105	-0.056	-0.017
q04_22	0.002	0.002	1.005	0.315	0.025	0.011
ncd2 ~						
FS	0.022	0.073	0.305	0.76	0.003	0.003
attain	0.035	0.003	13.888	0	0.14	0.139
li_urbrur	0	0	1.38	0.167	0	0.014
q04_20	-0.04	0.009	-4.579	0	-0.157	-0.048
q04_22	0.007	0.006	1.127	0.26	0.028	0.012
ncd3 ~						
FS	0.021	0.013	1.605	0.109	0.018	0.018
attain	0.002	0	3.719	0	0.038	0.038
li_urbrur	0	0	-0.259	0.796	0	-0.003
q04_20	-0.003	0.002	-2.009	0.045	-0.069	-0.021
q04_22	0.003	0.001	2.307	0.021	0.058	0.025
ncd4 ~						
FS	0.013	0.028	0.475	0.635	0.005	0.005
attain	0.002	0.001	1.723	0.085	0.018	0.017
li_urbrur	0	0	2.935	0.003	0.001	0.03

q04_20	-0.001	0.003	-0.373	0.709	-0.013	-0.004
q04_22	0	0.002	-0.083	0.934	-0.002	-0.001
ncd5 ~						
FS	0.044	0.035	1.258	0.208	0.014	0.014
attain	0.002	0.001	1.46	0.144	0.015	0.015
li_urbrur	0	0	1.156	0.248	0	0.012
q04_20	-0.008	0.004	-1.997	0.046	-0.069	-0.021
q04_22	0.004	0.003	1.217	0.224	0.03	0.013
ncd6 ~						
FS	0.009	0.096	0.092	0.927	0.001	0.001
attain	-0.047	0.003	-14.061	0	-0.142	-0.141
li_urbrur	0	0	-2.308	0.021	0	-0.023
q04_20	0.063	0.011	5.54	0	0.19	0.058
q04_22	-0.017	0.008	-2.08	0.037	-0.051	-0.022
Covariances:						
	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.ncd1 ~~						
.ncd2	-0.001	0	-3.328	0.001	-0.033	-0.033
.ncd3	0	0	-0.667	0.505	-0.007	-0.007
.ncd4	0	0	-0.871	0.384	-0.009	-0.009
.ncd5	0	0	-1.162	0.245	-0.012	-0.012
.ncd6	-0.006	0	-22.089	0	-0.225	-0.225
.ncd2 ~~						
.ncd3	0	0	-1.851	0.064	-0.018	-0.018
.ncd4	-0.001	0	-2.864	0.004	-0.029	-0.029
.ncd5	-0.001	0	-3.645	0	-0.036	-0.036
.ncd6	-0.058	0.001	-58.162	0	-0.71	-0.71
.ncd3 ~~						
.ncd4	0	0	-0.465	0.642	-0.005	-0.005
.ncd5	0	0	-0.664	0.506	-0.007	-0.007
.ncd6	-0.002	0	-11.253	0	-0.113	-0.113
.ncd4 ~~						
.ncd5	0	0	-1.213	0.225	-0.012	-0.012
.ncd6	-0.008	0	-24.542	0	-0.252	-0.252
.ncd5 ~~						
.ncd6	-0.013	0	-30.726	0	-0.321	-0.321
Variances:						
	Estimate	Std.Err	z-value	P(>|z|)	Std.lv	Std.all
.A	0.032	0	70.286	0	0.032	0.958
.food_Type_2	0.138	0.002	59.93	0	0.138	0.604

.food_Type_3	0.204	0.003	67.724	0	0.204	0.836
.food_Type_4	0.129	0.002	59.555	0	0.129	0.596
.B	0.14	0.002	65.972	0	0.14	0.769
.food_Type_6	0.122	0.002	60.843	0	0.122	0.624
.food_Type_7	0.239	0.003	70.291	0	0.239	0.958
.food_Type_8	0.121	0.002	69.761	0	0.121	0.93
.food_Type_9	0.17	0.003	63.593	0	0.17	0.694
.C	0.127	0.002	67.287	0	0.127	0.818
.D	0.142	0.002	64.787	0	0.142	0.73
.food_Type_12	0.155	0.003	60.711	0	0.155	0.621
.ncd_1	0				0	0
.ncd_2	0				0	0
.ncd_3	0				0	0
.ncd_4	0				0	0
.ncd_5	0				0	0
.ncd_6	0				0	0
FS	0.001	0	9.411	0	1	1
.ncd1	0.008	0	71.023	0	0.995	0.995
.ncd2	0.063	0.001	71.028	0	0.979	0.979
.ncd3	0.002	0	71.023	0	0.998	0.998
.ncd4	0.009	0	71.028	0	0.999	0.999
.ncd5	0.014	0	71.025	0	0.999	0.999
.ncd6	0.107	0.002	71.028	0	0.978	0.978

Estimator	ML
Optimization method	NLMINB
Number of model parameters	75
Number of observations	10090
Model Test User Model:	
Test statistic	6211.565
Degrees of freedom	168
P-value (Chi-square)	0
Model Test Baseline Model:	
Test statistic	40403.294
Degrees of freedom	225
P-value	0

User Model versus Baseline Model:	
Comparative Fit Index (CFI)	0.85
Tucker-Lewis Index (TLI)	0.799
Loglikelihood and Information Crit	
Loglikelihood user model (H0)	-9079.044
Loglikelihood unrestricted model	-5973.261
Akaike (AIC)	18308.087
Bayesian (BIC)	18849.535
Sample-size adjusted Bayesian (B	18611.196
Root Mean Square Error of Approxim	
RMSEA	0.06
90 Percent confidence interval -	0.058
90 Percent confidence interval -	0.061
P-value RMSEA <= 0.05	0
Standardized Root Mean Square Resi	
SRMR	0.063
Parameter Estimates:	
Standard errors	Standard
Information	Expected
Information saturated (h1) model	Structured

```
> fitMeasures(sema1,c("chisq","rmsea","srmr","gfi","ecvi"))
  chisq   rmsea    srmr     gfi    ecvi
3236.751   0.051   0.036   0.960   0.331
```

#Plot the diagram

```
>
semPaths(semmodel7,what="mod",whatLabels="parameters",style="OpenMx",edge.width=0.1,edge.color="black")
```

5. SAMPLE SELECTION AND PARTIAL OBSERVABILITY

```
> model1<-
Insecurity~Sx_1+Sx_2+Ms_1+MS_2+MS_3+MS_4+MS_5+MS_6+MS_7+ED_1+ED_2+ED_3
+ED_4+Work_1+Work_2+Work_3+Work_4+Tenure_1+Tenure_2+Tenure_3+Structure_1+Stru
cture_2+Structure_3+Structure_4+Structure_5+Water_1+Water_2+Electricity_1+Electricity_2+
Toilet_1+Toilet_2

> M1<-
Insecurity~Sx_1+Sx_2+Ms_1+MS_2+MS_3+MS_4+MS_5+MS_6+MS_7+ED_1+ED_2+ED_3
+ED_4+Work_1+Work_2+Work_3+Work_4+Tenure_1+Tenure_2+Tenure_3+Structure_1+Stru
cture_2+Structure_3+Structure_4+Structure_5+Water_1+Water_2+Electricity_1+Electricity_2+
Toilet_1+Toilet_2

> M2<-
HDDS~Sx_1+Sx_2+Ms_1+MS_2+MS_3+MS_4+MS_5+MS_6+MS_7+ED_1+ED_2+ED_3+E
D_4+Work_1+Work_2+Work_3+Work_4+Tenure_1+Tenure_2+Tenure_3+Structure_1+Structu
re_2+Structure_3+Structure_4+Structure_5+Water_1+Water_2+Electricity_1+Electricity_2+Toi
let_1+Toilet_2

> M3<-
MIFP~Sx_1+Sx_2+Ms_1+MS_2+MS_3+MS_4+MS_5+MS_6+MS_7+ED_1+ED_2+ED_3+E
D_4+Work_1+Work_2+Work_3+Work_4+Tenure_1+Tenure_2+Tenure_3+Structure_1+Structu
re_2+Structure_3+Structure_4+Structure_5+Water_1+Water_2+Electricity_1+Electricity_2+Toi
let_1+Toilet_2

> f.list(M1,M2)

Error in f.list(M1, M2) : could not find function "f.list"

> f.list<-list(M1,M2)

> marg<-c("probit","probit")

#  BivD="F" denoting bivariate distribution for Frank. “BSS” denoting bivariate sample
selection

> SS<-gjrm(f.list, data=Paper3, BivD="F", Model="BSS", margins=mar)

> summary(SS)


COPULA:   Frank

MARGIN 1: Bernoulli

MARGIN 2: Bernoulli
```

EQUATION 1

Link function for mu.1: probit

Formula: Insecurity ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coefficients		:		
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.808	7082.429	-0.001	0.999
Sx_1	-0.1583	0.2114	-0.749	0.454
Sx_2	-0.5476	3148.404	0	1
Ms_1	-0.7336	3148.404	0	1
MS_2	-0.7902	3148.404	0	1
MS_3	-0.6932	3148.404	0	1
MS_4	-1.1697	3148.404	0	1
MS_5	-6.9342	8190.41	-0.001	0.999
MS_6	-1.3326	3148.404	0	1
MS_7	-7.2977	2524.724	-0.003	0.998
ED_1	-6.6746	2524.724	-0.003	0.998
ED_2	-6.2952	2524.724	-0.002	0.998
ED_3	-5.7056	2524.724	-0.002	0.998
ED_4	-5.9945	2524.724	-0.002	0.998
Work_1	-6.2268	2524.724	-0.002	0.998
Work_2	-6.4088	2524.724	-0.003	0.998
Work_3	-5.8204	2524.724	-0.002	0.998
Work_4	-0.1158	5205.962	0	1
Tenure_1	-0.6445	5205.962	0	1
Tenure_2	-5.6318	7740.023	-0.001	0.999
Tenure_3	6.3425	3918.062	0.002	0.999
Structure_1	6.173	3918.062	0.002	0.999
Structure_2	5.9895	3918.062	0.002	0.999
Structure_3	5.9448	3918.062	0.002	0.999
Structure_4	-0.7766	6819.299	0	1
Structure_5	-42.6258	5414.556	-0.008	0.994
Water_1	-42.0931	5414.556	-0.008	0.994
Water_2	68.1962	5414.556	0.013	0.99
Electricity_1	68.1474	5414.556	0.013	0.99
Electricity_2	-10.6151	6431.781	-0.002	0.999

Toilet_1	-3.9586	6817.927	-0.001	1
Toilet_2	0.1745	0.1221	1.429	0.153

EQUATION 2

Link function for mu.2: probit

Formula: HDDS ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.4731	8192	0	1
Sx_1	-14.4234	8192	-0.002	0.999
Sx_2	-14.5227	8192	-0.002	0.999
Ms_1	-14.7238	8192	-0.002	0.999
MS_2	-14.6326	8192	-0.002	0.999
MS_3	-14.6418	8192	-0.002	0.999
MS_4	-14.7109	8192	-0.002	0.999
MS_5	34.62872	8192	0.004	0.997
MS_6	-14.6379	8192	-0.002	0.999
MS_7	-14.6921	8192	-0.002	0.999
ED_1	-14.5743	8192	-0.002	0.999
ED_2	-14.5415	8192	-0.002	0.999
ED_3	-14.2063	8192	-0.002	0.999
ED_4	-14.3854	8192	-0.002	0.999
Work_1	-14.2894	8192	-0.002	0.999
Work_2	-14.4048	8192	-0.002	0.999
Work_3	-14.3302	8192	-0.002	0.999
Work_4	-14.4423	8192	-0.002	0.999
Tenure_1	-14.5537	8192	-0.002	0.999
Tenure_2	40.99843	8192	0.005	0.996
Tenure_3	8.88075	8192	0.001	0.999
Structure_1	8.79503	8192	0.001	0.999
Structure_2	8.85997	8192	0.001	0.999
Structure_3	8.7937	8192	0.001	0.999
Structure_4	-15.0566	8192	-0.002	0.999
Structure_5	-18.3289	8192	-0.002	0.998

Water_1	-18.2091	8192	-0.002	0.998
Water_2	8.74104	8192	0.001	0.999
Electricity_1	8.71489	8192	0.001	0.999
Electricity_2	66.76216	8192	0.008	0.993
Toilet_1	67.29619	8192	0.008	0.993
Toilet_2	0.04005	8192	0	1
n = 371 n.sel = 150 theta = 100(87,100)				
tau = 0.961(0.955,0.961) total edf = 65				

> AIC(SS)

[1] 473.073

"BSS" denoting bivariate function for partial observability

> PO<-gjrm(f.list, data=Paper3, Model="BPO", margins=mar)

> summary(PO)

COPULA: Gaussian

MARGIN 1: Bernoulli

MARGIN 2: Bernoulli

EQUATION 1

Link function for mu.1: probit

Formula: Insecurity ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.04422	66.73568	0.001	0.9995
Sx_1	0.48371	0.21045	2.298	0.0215 *
Sx_2	-0.59522	32.72433	-0.018	0.9855
Ms_1	-0.89994	32.72499	-0.028	0.9781

MS_2	-0.97615	32.72382	-0.03	0.9762
MS_3	-0.71401	32.72442	-0.022	0.9826
MS_4	-1.34048	32.72378	-0.041	0.9673
MS_5	-5.85783	124.0662	-0.047	0.9623
MS_6	-1.43116	32.76436	-0.044	0.9652
MS_7	-6.30563	24.43378	-0.258	0.7964
ED_1	-5.58831	24.43447	-0.229	0.8191
ED_2	-5.06296	24.43426	-0.207	0.8358
ED_3	-4.55699	24.43379	-0.187	0.852
ED_4	-4.79465	24.45488	-0.196	0.8446
Work_1	-5.1807	24.45536	-0.212	0.8322
Work_2	-5.31955	24.4549	-0.218	0.8278
Work_3	-4.72545	24.45467	-0.193	0.8468
Work_4	-0.14631	55.28766	-0.003	0.9979
Tenure_1	-0.60513	55.28728	-0.011	0.9913
Tenure_2	-4.20108	81.59361	-0.051	0.9589
Tenure_3	5.25813	53.35509	0.099	0.9215
Structure_1	5.03008	53.35319	0.094	0.9249
Structure_2	4.67581	53.35429	0.088	0.9302
Structure_3	4.97149	53.35435	0.093	0.9258
Structure_4	-0.71544	91.97737	-0.008	0.9938
Structure_5	-0.49817	53.58638	-0.009	0.9926
Water_1	0.01912	53.58538	0	0.9997
Water_2	14.92397	43.68023	0.342	0.7326
Electricity_1	15.06049	43.67351	0.345	0.7302
Electricity_2	-8.1045	70.97083	-0.114	0.9091
Toilet_1	-2.80961	102.7098	-0.027	0.9782
Toilet_2	0.08517	0.1193	0.714	0.4753

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*	’ 0.05 ‘.’ 0.1 ‘ ’ 1

EQUATION 2

Link function for mu.2: probit

Formula: HDDS ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	87.0818	59.9021	1.454	0.14602
Sx_1	-24.1477	20.3961	-1.184	0.236438
Sx_2	-27.6327	20.338	-1.359	0.17425
Ms_1	-17.342	20.3976	-0.85	0.395215
MS_2	-18.2659	20.3443	-0.898	0.369271
MS_3	-27.8198	20.3327	-1.368	0.17124
MS_4	-27.3259	20.3326	-1.344	0.178965
MS_5	-28.3335	8192	-0.003	0.99724
MS_6	-19.4556	20.3463	-0.956	0.338959
MS_7	-12.1	20.3059	-0.596	0.551251
ED_1	-11.0291	20.304	-0.543	0.586993
ED_2	-12.9384	20.3044	-0.637	0.523981
ED_3	-11.75	20.304	-0.579	0.562791
ED_4	-8.1275	20.3346	-0.4	0.689388
Work_1	-6.0477	20.343	-0.297	0.766248
Work_2	-4.9863	20.3419	-0.245	0.806362
Work_3	-7.7362	20.3402	-0.38	0.703693
Work_4	-1.5458	24.2064	-0.064	0.949081
Tenure_1	-2.3191	24.2178	-0.096	0.923709
Tenure_2	-4.7945	8192	-0.001	0.999533
Tenure_3	-20.0195	58.8618	-0.34	0.733773
Structure_1	-23.0471	58.8499	-0.392	0.695335
Structure_2	-20.7483	58.8809	-0.352	0.724555
Structure_3	-20.1484	58.8445	-0.342	0.732049
Structure_4	-18.2537	939.4957	-0.019	0.984499
Structure_5	1.2039	51.4382	0.023	0.981328
Water_1	0.6003	51.4383	0.012	0.990688
Water_2	1.1678	62.0525	0.019	0.984985
Electricity_1	0.9272	62.0464	0.015	0.988078
Electricity_2	-5.7911	69.2436	-0.084	0.933347
Toilet_1	4.4881	940.904	0.005	0.996194
Toilet_2	1.1381	0.3188	3.57	0.000358 ***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*	’ 0.05 ‘.’ 0.1 ‘ ’ 1
n = 371 , theta = 0.272(0.258,0.289), tau =0.175(0.166,0.187)				

> AIC (PO)

[1] 3093.055

“BSS” denoting bivariate sample selection

> SS2<-gjrm(M.list, data=Paper3c, Model="BSS", margins=margin)

> summary(SS2)

COPULA: Gaussian

MARGIN 1: Bernoulli

MARGIN 2: Poisson

EQUATION 1

Link function for mu.1: probit

Formula: Insecurity ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coe	fficients:			
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.73E+01	6.64E+03	-0.009	0.993115
Sx_1	-1.01E-01	2.74E-01	-0.37	0.711127
Sx_2	-1.65E+00	1.94E+03	-0.001	0.999323
Ms_1	-1.69E+00	1.94E+03	-0.001	0.999305
MS_2	-1.74E+00	1.94E+03	-0.001	0.999287
MS_3	-1.55E+00	1.94E+03	-0.001	0.999362
MS_4	-2.15E+00	1.94E+03	-0.001	0.999115
MS_5	-6.45E+00	2.28E+03	-0.003	0.997745
MS_6	-2.21E+00	1.94E+03	-0.001	0.999091

MS_7	-2.39E+00	1.79E+03	-0.001	0.998936
ED_1	-1.97E+00	1.79E+03	-0.001	0.999124
ED_2	-1.70E+00	1.79E+03	-0.001	0.999242
ED_3	-7.88E-01	1.79E+03	0	0.99965
ED_4	-7.09E-01	1.72E+03	0	0.999671
Work_1	-8.08E-01	1.72E+03	0	0.999625
Work_2	-1.03E+00	1.72E+03	-0.001	0.999523
Work_3	-3.48E-01	1.72E+03	0	0.999839
Work_4	7.22E+01	2.65E+03	0.027	0.978255
Tenure_1	7.17E+01	2.65E+03	0.027	0.978389
Tenure_2	6.85E+01	2.56E+03	0.027	0.978644
Tenure_3	6.87E+01	2.57E+03	0.027	0.978646
Structure_1	6.86E+01	2.57E+03	0.027	0.978654
Structure_2	6.85E+01	2.57E+03	0.027	0.978712
Structure_3	6.82E+01	2.57E+03	0.027	0.9788
Structure_4	6.28E+01	2.66E+03	0.024	0.9812
Structure_5	-5.25E-01	3.79E+03	0	0.99989
Water_1	1.73E-03	3.79E+03	0	1
Water_2	-7.37E+01	3.93E+03	-0.019	0.985046
Electricity_1	-7.44E+01	3.93E+03	-0.019	0.984897
Electricity_2	-4.95E+00	2.04E+03	-0.002	0.998063
Toilet_1	8.89E-01	2.12E+03	0	0.999665
Toilet_2	-6.59E-01	1.76E-01	-3.747	0.000179 ***

Signif. codes:	0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’ 0.1 ‘ ’ 1

EQUATION 2

Link function for mu.2: log

Formula: MIFP ~ Sx_1 + Sx_2 + Ms_1 + MS_2 + MS_3 + MS_4 + MS_5 + MS_6 + MS_7 + ED_1 + ED_2 + ED_3 + ED_4 + Work_1 + Work_2 + Work_3 + Work_4 + Tenure_1 + Tenure_2 + Tenure_3 + Structure_1 + Structure_2 + Structure_3 + Structure_4 + Structure_5 + Water_1 + Water_2 + Electricity_1 + Electricity_2 + Toilet_1 + Toilet_2

Parametric coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	9.69E-01	7.21E+03	0	0.99989
Sx_1	-2.49E-01	9.69E-02	-2.563	0.01036 *
Sx_2	2.76E-01	3.58E+03	0	0.99994

Ms_1	-1.16E-01	3.58E+03	0	0.99997
MS_2	-3.38E-01	3.58E+03	0	0.99992
MS_3	4.72E-01	3.58E+03	0	0.99989
MS_4	3.29E-02	3.58E+03	0	0.99999
MS_5	-1.79E-16	8.19E+03	0	1
MS_6	6.42E-01	8.19E+03	0	0.99994
MS_7	-6.34E-01	3.98E+03	0	0.99987
ED_1	2.27E-01	3.98E+03	0	0.99995
ED_2	5.80E-01	3.98E+03	0	0.99988
ED_3	7.97E-01	3.98E+03	0	0.99984
ED_4	4.14E-01	3.98E+03	0	0.99992
Work_1	3.58E-01	3.98E+03	0	0.99993
Work_2	-3.06E-01	3.98E+03	0	0.99994
Work_3	5.03E-01	3.98E+03	0	0.9999
Work_4	6.64E-01	5.46E+03	0	0.9999
Tenure_1	3.05E-01	5.46E+03	0	0.99996
Tenure_2	1.44E-16	8.19E+03	0	1
Tenure_3	4.71E-01	3.98E+03	0	0.99991
Structure_1	2.18E-01	3.98E+03	0	0.99996
Structure_2	5.12E-01	3.98E+03	0	0.9999
Structure_3	-2.32E-01	3.98E+03	0	0.99995
Structure_4	3.61E-17	8.19E+03	0	1
Structure_5	3.40E-01	5.46E+03	0	0.99995
Water_1	6.29E-01	5.46E+03	0	0.99991
Water_2	7.05E-01	7.21E+03	0	0.99992
Electricity_1	2.64E-01	8.19E+03	0	0.99997
Electricity_2	-6.42E-01	5.46E+03	0	0.99991
Toilet_1	1.61E+00	5.46E+03	0	0.99976
Toilet_2	-2.47E-01	8.06E-02	-3.069	0.00215 **

Signif. codes	: 0 ‘***’	0.001 ‘**’	0.01 ‘*’	0.05 ‘.’ 0.1 ‘ ’ 1
n = 314 n.sel = 91, theta = 0.535 (0.382,0.65), tau = 0.359(0.25,0.45), total edf = 65				