UNIVERSITY OF KWAZULU-NATAL

Flexible Statistical Modelling in Food Insecurity Risk Assessment

2015

LAILA BARNABA LOKOSANG

Flexible Statistical Modelling in Food Insecurity Risk Assessment

BY

LAILA BARNABA LOKOSANG

(BSc, MSc)

Submitted in fulfilment of the academic requirements for the degree of

Doctor of Philosophy

in

Statistics

in the

School of Mathematics, Statistics and Computer Science

University of KwaZulu–Natal

Pietermaritzburg

2015

DEDICATION

To my mother Mrs. Maria Yendu Lemi

DECLARATION

The research work described in this thesis was carried out in the School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Pietermaritzburg, under the supervision of Dr. Shaun Ramroop and Professor Temesgen Zewotir.

I, Laila Barnaba Lokosang, declare that this thesis is my own, unaided work. It has not been submitted in any form for any degree or diploma to any other University. Where use has been made of the work of others, it is duly acknowledged.

02 January 2016

22 January 2016

Laila Barnaba Lokosang — Date

Dr. Shaun Ramroop — Date

Professor Temesgen Zewotir — Date

NOTE

The following articles have been published, under peer review from this thesis:

1. Lokosang, L., Ramroop, S. & Zewotir, T., 2014. Indexing household resilience to food insecurity shocks: The case of South Sudan. *Agrekon:* 53(2), 137–159. DOI: 10.1080/03031853.2014.915486.
2. Lokosang, L., Ramroop, S. & Zewotir, T., 2016. The Effect of Weakened Resilience on Food Insecurity in Protracted Crisis: The Case of South Sudan. *Agriculture and Food Security*. DOI:10.1186/s40066-016-0051-y.
3. Lokosang, L., Ramroop, S. & Zewotir, T., n.d. The Effect of Weakened Resilience on "Double-Risk" of Food Insecurity: The Case of South Sudan. *Food Security*. (under review)

ACKNOWLEDGEMENT

I take pleasure in appreciating my Academic Supervisors Dr. Shaun Ramroop and Professor Temesgen Zewotir, both of the University of KwaZulu-Natal's School of Mathematics, Statistics and Computer Science for their guidance and encouragement. It also heartens me to acknowledge two of my colleagues: Mr. James Lemi Lugga (formerly of the National Bureau of Statistics in South Sudan), for providing me with the datasets used and Dr. Dawit Getnet Ayele who assisted me with some of the software used in the analysis of data. To God be the glory.

ABSTRACT

Food insecurity has remained a persistent problem in Sub-Saharan Africa. Conflict and other protracted crisis have rendered a significant proportion of Africa's populations to suffer the risk of food insecurity, as their resilience to livelihood shocks weakens. A significant and immense body of research in the past two decades has largely centred on describing the incidence of food insecurity and vulnerability. Limited research was done using statistical methods to determine the likelihood of food insecurity risk. The use of flexible statistical techniques for a sound and purposive monitoring, evaluation, planning and decision making in food security and resilience was limited.

The study aimed to extend the use of statistics into the expanding field of food security and resilience, and also to provide new direction for future research involving applications of the methods explored, such as adjustments in statistical methods, sampling and data collection. The study specifically aims at helping food security analysts with tested and statistically robust tools for use in the analyses of the likelihood of food insecurity risk in settings with structural food insecurity issues. Moreover, it aimed to inform practice, policy and analysis in monitoring and evaluation of food insecurity risk in protracted crisis; thus helping in improving risk aversion measures.

Utilising secondary data, the research examines relevant statistical techniques for determining predictors of food insecurity risk, namely; Principal Component Analysis; Multiple Correspondence Analysis; Classification and Regression Tree Analysis; Survey Logistic Regression, Generalized Linear Mixed Models for Ordered Categorical Data; and Joint Modelling. The study was conducted in the form of structured analysis of different datasets

collected in the conflict-ridden South Sudan. Assets owned by households, as well as availability of livelihood endowments, was used as proxy for determining the level of resilience in particular demographic unit or geographical setting.

The study highlighted the strengths and weaknesses of the techniques explored in the analysis as identifying or classifying potential predictors of food insecurity outcomes. Each technique is capable of generating a unique composite index for measuring the amount of resilience and predicting and classifying households according to food insecurity phase based on factor loadings.

In general, the study determined that each method explored has peculiar strengths as well as limitations. However, a noteworthy implication observed is that asset-based statistical analysis, whether based on composite index that can be used as proxy for measuring the amount of resilience to food insecurity eventualities or on regression modelling approaches, does assure sufficient rigour in drawing conclusions about the wellbeing of households or populations under study and how they might withstand food insecurity and livelihood shocks. As food insecurity and malnutrition continue to attract substantial attention, such flexible analytical approaches exert potential usefulness in determining food insecurity risks, especially in protracted crisis settings.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACRONYMS

AGFI	Adjusted Goodness of Fit Index
AIC	Akaike Information Criterion (Akaike 1973; Akaike 1987)
ANLA	Annual Needs and Livelihoods Analysis
CAADP	Comprehensive Africa Agriculture Development Programme
CART	Classification and Regression Tree
CHAID	Chi-squared Automatic Interaction Detection
CI	Confidence interval
CSI	Coping Strategies Index
DfID	Department for International Development of the United Kingdom
DHS	Demographic and Health Surveys
EAs	Enumeration areas (in a survey or census)
FAFS	Framework for African Food Security
FAO	Food and Agricultural Organisation of the United Nations
FCS	Food Consumption Score
FSMS	Food Security Monitoring Survey
GDP	Gross Domestic Product
GHI	Global Hunger Index
GLM	Generalized Linear Models
GLMM	Generalized Linear Mixed Models
HBS	Household Budget Surveys

HDI	Human Development Index
HH	Household
HHH	Head of household
HRI	Household Resilience Index
HWS	Health and Welfare Survey
IFAD	International Fund for Agricultural Development
MCA	Multiple Correspondence Analysis
MDG	Millennium Development Goals
MICS	Multiple Indicator Cluster Survey
MLE	Maximum Likelihood Estimator
NBHS	National Baseline Household Surveys
NBS	National Bureau of Statistics of South Sudan
NEPAD	New Partnership for Africa's Development
OR	Odds ratio
P	Probability value
PCA	Principal Component Analysis
PPP	Purchasing Power Parity
PPS	Probability proportion to size
PSU	Primary sampling units
REML	Restricted Maximum Likelihood Estimate
S.E	Standard error
SES	Socioeconomic status

UNDP	United Nations Development Programme
UNHCR	United Nations High Commissioner for Refugees
UNICEF	United Nations Children's Fund
VAM	Vulnerability Assessment and Mapping
WFP	World Food Programme of the United Nations
WHS	World Health Surveys

CHAPTER 1

Introduction

Over the years food insecurity has evolved as a major global problem that is deeply rooted in Africa. The 2012 Global Hunger Index (GHI) Report (Welthungerhilfe et al. 2012) depicts Africa to have extremely alarming hunger situation in 10 out of 12 countries of the world. With GHI value of 30 and above, hunger and food insecurity affects the most vulnerable of the population, leading to many debilitating social, economic and political problems. Causes of food insecurity – both natural and human inflicted – remain rampant, affecting millions of people each year in the poorest countries, especially when population growth keeps soaring. Two forms of new global trends, namely; climate change and burgeoning population, pose as notorious source of food insecurity shocks, especially in the developing countries.

Hunger, a pervasive problem in developing countries, as it causes substantial resources to be spent on food (Smith et al. 2006; FAO et al. 2013, p.27), should not be misconstrued to be due to lack of food alone. It is actually a function of several other factors that lead to lack of food such as flight from conflict that causes households to leave their food reserves behind in situations of emergency, crop failure, shortage of rains, dependency on food aid, animal disease, death or illness of economically productive household member and other shocks and strains. The 2013 State of the World Food Security, defines hunger as 'synonymous with chronic undernourishment' (FAO et al. 2013, p.50). Inadequacy of diets rich in micronutrients directly causes malnutrition and stunting, which is known as chronic form of food insecurity or 'hidden hunger' (FAO et al. 2013, p.32). In general, hunger, a result of severe food shortage and malnutrition, can be attributed to major causes, namely; poverty, protracted crises, social factors, emerging adverse economic trends such as price volatility of strategic commodities, inflation and transitory or recurrent effects such as environmental causes.

Hunger, food insecurity and malnutrition in Africa, especially in the Sub-Saharan region, are heightened by deeply ingrained poverty mainly in rural and peri-urban populations. Leading development organisations, on top of which are the World Bank and the United Nations Development Programme (UNDP), have consistently reported high poverty levels as measured by indices of 1.25 United States Dollars per day (World Bank 2008), Purchasing Power Parity (PPP) (Asian Development Bank 2008b), Gini Coefficients – a measure of income inequality (Deation 1997), per capita food supply of less than 2,200 calories (9,200 kilojoules) per day (World Bank 2008), food consumption ratio, i.e., the ratio of total expenditure on non-food to food items (Adongo & Deen-Swarray 2006) and other non-monetary proxy measures such as the Human Development Index (HDI) (Herrero et al. 2010; United Nations Development Programme 2009).

While efforts for eradicating poverty involve a complex of long-term development strategies, which become the main policy of governments in Africa, the race to finding solutions to reducing hunger has been on for many years. The United Nations estimates that there were close to one billion people worldwide affected by hunger (United Nations Development Programme 2010). A number of countries supported by multilateral development organisations have developed policies for reducing poverty, hunger and food insecurity; all interconnected problems (World Bank & IMF 2002). At the epitome of the global search for solutions, is the Millennium Development Goals (MDGs) (The United Nations 2002). The first of the MDGs spells out 'halving extreme poverty and hunger by the year 2015'.

One of the most important causes of hunger in Africa is protracted crises mainly as a result of conflict that also result in protracted food emergencies. Russo et al. (2008), citing Flores (2007) define protracted crises to be situations where large sections of populations are faced with acute threat to life and livelihoods over extended periods, especially when state and governance

institutions fail to provide adequate levels of protections. Protracted food emergencies can by themselves inadvertently become a cause for conflict and the vicious cycle continues. When a situation of structured food insecurity and malnutrition emergency is not addressed, it can by itself inadvertently cause or exacerbate tensions or conflict (Committee on World Food Security 2015). Severe food insecurity causes anxiety, which in turn causes desperation, which in turn causes households to resort to extreme or even unthinkable forms of survival or coping strategies. In situations where firearms are rampant, extreme coping strategies might be in the form of banditry, armed robbery and rustling of cattle – a practice existing amongst pastoralist communities of South Sudan.

Protracted or chronic food insecurity disables development and tear apart social fabrics of affected communities. Russo, et al. (2008) argue that achieving food security in crises of a complex and protracted nature can be a daunting task, as states become fragile. Schafer (2002) includes high vulnerability of livelihoods to external shocks and existence of serious poverty among several elements characterising protracted crises. This implies that vulnerability is most serious in protracted crisis and thus exacerbates poverty. It is on these grounds that development and relief need to go alongside each other such as the UN 'twin track' approach for intervening in crises in Sudan, Somalia and the Democratic Republic of Congo. The range of developmental interventions included livestock development, trade and veterinary services (Bishop et al. 2008) and developmental programmes aimed at sustaining local solutions with local community participation (Pantuliano 2008).

South Sudan, a country that recently acquired independence from Sudan – on 9th July 2011, is recovering from effects of protracted civil war that lasted nearly 22 years (from 1983 to 2005). Economic factors, such as general poverty, unemployment, lack of means of agricultural production (e.g. capital and inputs) and poor physical infrastructure, also worsened the

situation. Insecurity and displacement is also among the major causes of food insecurity and diminished livelihoods.

Another major factor causing structural food insecurity comprises of social factors, mainly poor health and low literacy are cited among the factors increasing vulnerability to food insecurity and poor livelihoods (World Food Programme 2007). For example, major epidemics such as HIV, and more recently the Ebola virus, interrupted economic activities of rural and poor communities and plunged the affected populations into food insecurity and undernutrition.

Transitory food insecurity is a function of recurrent crises such as those caused by environmental or geographical conditions. The most common form of transitory food insecurity in South Sudan is flooding and drought. Poor rainfall and flooding of lowland areas known as Eastern and Western Flood Plains on either side of the Nile (or the Sudd Region), have confounded to the country's structural food insecurity problems. The country has vast flood plains covering nearly half of its territory (Famine Early Warning Systems Network 2007). This has made the country to experience recurrent episodes of food insecurity, extreme hunger and humanitarian crises, even after it gained independence from Sudan in 2011.

Volatility of consumer prices and strategic commodity prices and other economic strains such as inflation compound to food insecurity of already poor consumers and producers (von Grebmer et al. 2011). Literature abounds on how strategic commodity price spikes and volatility have recently gained importance in poverty, livelihood and food security analysis. The Food and Agricultural Organisation (2012a) presents evidence on how price volatility affects the agricultural sector negatively, especially for importing countries.

All these causes of food insecurity amount to weakening the resilience of populations and thus the risk to food insecurity. Famine and food insecurity often hits households and communities

hardest when their resilience to such stresses is weak. Food insecurity resilience is technically defined as the ability of a household, community or a population to bounce back after experiencing a spell of food shortage, or when exposed to such form of shock. Resilience can also be defined as the ability of a household to resist, absorb, cope with and recover from the effects of shocks and to adapt to longer time changes in a timely and efficient manner. In brief, it is the capacity to endure food insecurity shocks and stressors and bounce back (Pasteur 2011).

The question that development and food security policy makers need to ask, therefore, is how can we build resilience, especially of populations with vulnerabilities related to income poverty, resource depravation and lack of certain geographical endowments and limited social and economic capital? Must we not start by assessing which population groups often become vulnerable due to low resilience?

Food insecurity resilience or its shortage is very much tied with the availability or lack of what households or communities have or lack. Resilience can surely be improved by the level of resources (assets), social, economic, natural and/or human capitals at a household's disposal or reach. Therefore, it is easy to argue that if the determinants of resilience are known, then why not use them to predict the outcomes of resilience in order to prepare adequately against the eventuality of food insecurity shocks? The International Food Policy Research Institute (2014) argues that building resilience makes it possible to prevent adverse stressors and shocks from bearing long-lasting negative development consequences to resource poor communities.

South Sudan is a typical case of weakened resilience, as it had just emerged from years of conflict and economic deprivation when its first household baseline survey was undertaken. Majority of the populations were still in transitional stages, with many living in displacement and transit camps. An equally large section of the population was still resettling in their homes

of origin, while urban centres were mushrooming with temporary settlements. It is on these grounds that this study is perceived to be of relevance and research significance.

Against the bleak background of high risk of food insecurity and weakened resilience to food insecurity risk, some continental initiatives surfaced in the last decade. In response to the persistence of structural food insecurity crises and low resilience on the continent, the break of the New Millennium led the African Union and its partners to conceive and endorse the Comprehensive Africa Agriculture Development Programme (CAADP). Later in 2009 the CAADP Framework for Africa's Food Security (FAFS) (New Partnership for Africa's Development 2009) was launched. The FAFS categorically targeted the chronically food-insecure and vulnerable populations affected by various crises and emergencies (New Partnership for Africa's Development 2009). FAFS' further aims to increase the resilience of vulnerable populations by reducing the risk of food insecurity. To fulfil this goal, the framework outlines the first of its four objectives as "to improve risk management at the household, community, national and regional levels" (New Partnership for Africa's Development 2003, p.46).

The framework ostensibly commits to the first Millennium Development Goal of cutting extreme poverty and hunger by half by the year 2015. It also aligns with the CAADP vision of attaining sustainable agriculture-led annual economic growth, with agriculture contributing an average of 6% to the Gross Domestic Product (GDP). To this end, African governments committed to increasing public investment in agriculture by allocating a minimum of 10% of the annual budget to the agricultural sector. Investment in agriculture through increased support to rural and smallholder farmers, has received considerable attention in the literature. Researchers such as Staatz and Dembélé (2007) have substantiated this need with empirical evidence. CAADP sufficiently offers the guiding lights on how the continent's smallholders

and the vulnerable can be targeted with a range of facilities to disentangle them out of poverty traps.

Despite the efforts to mitigate food insecurity and its causes on the continent, a major challenge has been at play - monitoring of the situation and use of reliable evidence for efficient decision making and effective response to hunger and malnutrition. Epitomising this challenge is in finding the standard measurement approaches. To-date existing measures are those mainly used in measuring vulnerability such as those used by the World Food Programme (WFP)'s Vulnerability Analysis and Mapping (VAM) and the Annual Needs and Livelihoods Analysis (ANLA). Since 2005 WFP and partner organisations conducted successive monitoring studies in South Sudan to identify the major factors influencing vulnerability and also the livelihood and humanitarian needs of the various population groups. The assessment is thus a tool for guiding in planning and implementing appropriate relief and humanitarian mitigation interventions.

The ANLA methods are pragmatic and analyses are based on stratified sampling of populations (World Food Programme 2007; FAO 2008). Data collection was often done by field workers of humanitarian agencies and national counterparts. In-country data analyses and reporting are mainly carried out by WFP-VAM units. However, these analyses usually employ descriptive methods rather than the robust statistical modelling techniques (Lokosang et al. 2010). This therefore uncovers the need to find ways of estimating the risk of food insecurity and mechanisms for predicting it. However, food security analysts point out the current dilemma of arriving at a standard measure for estimating and predicting food insecurity risk.

In order to intervene decisively toward mitigating food insecurity risks and improving livelihoods of populations trapped in protracted crises, the need for good information was paramount. Monitoring of crises and the need for robust measures has considerably increased

in the literature. Lokosang et al. (2010) elaborate the various tools, approaches and measurement indicators for monitoring food insecurity; both structural and transitory. The Status of Food Insecurity in the World (FAO et al. 2013) offers a "monitoring framework for the post-2015 development agenda", outlining the set of indicators to be monitored by focus area. None of these indicators, however, explicitly relate to the use of predictive statistics for early warning and preparedness.

For close to fifteen years there has been a growing interest in assessing household socioeconomic status (SES), identifying and profiling the economically poor (Hancioglu 2002). It seems that this surge of research interests has been evoked by the availability of data from Demographic and Health Surveys (DHS), Household Budget Surveys (HBS), Health and Welfare Survey (HWS) of Thailand (Prakongsai 2006), National Family Health Survey (NFHS) of India (Filmer & Pritchett 1998; Filmer & Pritchett 2001), the World Bank's Living Standards Measurement Study (LSMS), UNICEF's Multiple Indicator Cluster Survey (MICS) and WHO's World Health Surveys (WHS). These national surveys have been conducted in over 70 countries around the world by government statistical agencies and in collaboration with donor organisations (McKenzie 2004; Rutstein & Kiersten 2004). This is of course adding to the interest shown and the efforts made by countries developing and implementing poverty eradication (or pro-poor) strategies, which emerged in the last one and a half decade.

Prakongsai (2006) observes that some global concerns have led to a need to finding practical tools for the identification of socioeconomic status at both individual and household levels, particularly in developing countries. The commitment to the Millennium Development Goals (MDGs) by United Nations member countries (The United Nations 2002), must have also added momentum to these interests in finding rapid and robust measures for estimating socio-economic status and, by extension, poverty.

Instability in South Sudan had lasted close to 40 years in which period the country experienced some of the worst socioeconomic statistics according to various humanitarian reports. Although some efforts were exerted to produce some SES and poverty information based on a National Baseline Household Survey (NBHS) in 2009, measurements for predicting resilience to food insecurity shocks were not determined.

It is perceived that some semi-durable household assets, livelihoods capitals and certain household characteristics permit it to withstand improved or reduced resilience to food insecurity shocks and stressors (Prakongsai 2006). Availability of assets, livelihood constructs and certain household attributes is also considered a proxy for socio-economic welfare of the household and by extension poverty (Kumar 1989; Falkingham & Namzie 2001).

Semi-durable assets include transport (vehicles, bicycles, motor cycles and carts), means of communication and information (TV, Radio, telephone sets, computer and internet), and other appliances such as refrigerators, fans and air conditioners (Filmer & Pritchett 1998; Prakongsai 2006). Filmer and Pritchett (1998) and Prakongsai (2006) include in the analysis of data from India seven semi-durable asset indicators: bicycle, radio, television, sewing machine, motorcycle/scooter, refrigerator and car.

Livelihood capitals cover the domain of five categories of durable assets otherwise known as household endowments or property. These categories are physical, financial, human, natural and social. Elasha et al. (2005) describe physical capital to be that which is created by economic production. This is said to include infrastructure, reticulated equipment and housing. Human capital includes household size (number of household members), education, skills and health of household members (Elasha et al. 2005). Financial capital obviously covers liquid income such as wages and salaries, property income, pension, remittances, savings and access to credit. Natural capital is in the form of land, water, bioresources (e.g. trees, pasture and biodiversity).

Social capital consists of kinship support and other forms of social support, such as support from professional associations (Elasha et al. 2005).

It is rational to expect that resilience to food insecurity shocks can be improved or worsened by availability or absence of certain possessions or livelihood endowments. This understanding can be reconstructed from the so-called Sustainable Livelihoods Framework (Department for International Development 1999), which depicts the five livelihoods capitals to be determining, as well as being determined by, certain livelihood outcomes: sustainable use of natural resources, income, wellbeing, vulnerability and food security. There are other approaches or analytical frameworks for measuring resilience. For example Food and Agricultural Organisation (2012b) and Alinovi et al. (2010) estimate household resilience to food insecurity shocks as a function of seven livelihood constructs, namely; Income and food access; access to basic services; agricultural assets/non-agricultural assets; enabling institutional environment; climate change; agricultural practice and technology; and social safety nets.

The research problem for informing this study lies in that literature reviewed demonstrates that food security experts do not amply utilize the rigour of statistics. Despite the availability of robust statistical tools that have the rigour to satisfy the quest for assessing food insecurity risk, existing food insecurity and resilience analysis hardly, if at all, apply them. Rather, existing analysis of survey data heavily depend on the rudimentary, exploratory or descriptive statistics that lack depth. As a result, food insecurity and resilience analysis lack the efficiency of scientifically established evidence.

This research work is, therefore, aimed at achieving the following objectives:

- To extend applications of statistical methodology to a domain where the rigour and substance of statistics has for long been under-utilised. This work in particular is

expected to allow future analysts to acquire the confidence of using the methods explored in this research.

- As the field of food security is expanding rapidly due to the challenges described in the preceding sections, the need for more convincing and statistically established evidence is equally growing. There is need for exploiting statistical methods that have established robustness and efficiency and extending them into the domain of food insecurity and resilience measurement and informatics. In short, this work is aimed at enabling discipline of statistics cut into a field of knowledge (food insecurity and resilience) where it's much unknown or avoided. In other words, the study hopes to offer increased use of statistics in food insecurity and resilience.
- A number of academic institutions and agriculture economics disciplines, in particular, have started offering qualifications in food security studies. Statisticians are expected to extend or adapt applications to satisfying the growing demand for food in/security metrics in measuring or assessing vulnerability, resilience and risk. The study might also motivate some improvements in the methods explored, where need arises.
- By exploring two datasets (see Chapter 2) the study aims to validate conclusions and ascertain reliability of the measures.

In this light, the study sought an answer to three specific questions:

(i) Does the methodology for identifying the poor and profiling poverty proxies still apply in the context of the post-conflict South Sudan?

(ii) Based on the dataset collected pre- (2008 to 2010) and post-South Sudan's independence (after 2011), do the approaches examined arrive at similar conclusion of robustness of the asset-based methods and their validity for estimation and prediction of household food security outcomes?

(iii) Out of all the methods explored to the South Sudanese datasets, is there one that provides the 'best' estimates and predicted values as well as robustness tests?

To provide answers to these questions, this research project aims at exploring the rigour of models that may help in predicting the indications or outcomes of transitory or structured food insecurity, generating indices for measuring household resilience to food insecurity risk, strains and shocks, and profiling and mapping the predicted food insecurity levels. By so doing, the study hopes to strengthen food security monitoring, evaluation and reporting systems toward more robust, statistics-based predictive analysis. Traditional vulnerability analysis often shy away from such approaches in a misconception that statistical methods are complicated and user-unfriendly. The study specifically taps into recent work in constructing asset-based indices for estimating poverty and socioeconomic welfare of a given population.

The study, therefore, borrows from approaches by development analysts who constructed asset indices as proxy for estimating socio-economic wellbeing and, by extension, poverty. The study hopes to derive resilience indices from the set of household-based livelihood assets that are robust enough to profile inequalities in levels of resilience to food insecurity shocks and declining livelihoods. This desirable outcome will, however, depend on the quality of the data.

More specifically, the study derives its significance in attempting to find a measure of the amount of resilience households may exert in order to withstand food insecurity shocks. Statistical approaches explored attempt to utilise the rigor of factor analysis and modelling techniques for generating predictor variables that tend to associate with food insecurity and poverty-related outcomes. As the generated new predictor variable is assumed to determine future risk to food insecurity, and is based on weights of the combination of possible predictors, it is said to be indicative of the amount of resilience a household exerts. It is then referred to as "Household Food Insecurity Resilience Index" (or HRI in short).

Conceptually, if the strength of people's resilience could be determined (or classified) and predicted, it could be possible to influence decisions leading to preventive or early preparedness actions. Therefore, it is deemed of relevance and significance attempting to offer a viable tool for humanitarian and development programmes to intervene timely, and from an informed viewpoint, by targeting populations most at risk of food insecurity. Analysts and programme designers may also use the evidence for intensifying preventive action measures.

Over the last two decades poverty and socioeconomic welfare analysts and researchers have based their estimates of wellbeing and livelihoods on national household surveys, such as the Demographic and Health Survey (DHS), multi-cluster surveys, and Household Budget Survey (HBS). Due to lack of reliable data on household income and consumption, proxy indicators for estimating and profiling socio-economic status, poverty and household welfare were used (Filmer & Pritchett 1998; Falkingham & Namzie 2001; Prakongsai 2006). The proxy indicators or indices are often derived based on durable and semi-durable assets; also known as livelihood capitals and household characteristics (Elasha et al. 2005).

The study extends the use of statistical methods to analysis of asset-based and livelihood data to assess food insecurity risk in a protracted conflict setting. These methods are classified into three categories. In the first class of methods (see Chapter 3), Principal Component Analysis (PCA), Multiple Correspondence Analysis (MCA) (Clausen 1998; Greenacre 1984; Greenacre 1993) and Classification and Regression Tree (CART) Analysis (Breiman et al. 1984; Lemon et al. 2003), have featured highly in the derivation of indices for estimating, profiling and mapping poverty based on a set of variables related to household characteristics, durable assets owned and key livelihood capitals (Asselin 2002; Asselin & Vu Tuan 2008; Booysen et al. 2005).

Since the study explores several analytic approaches, a number of relevant assumptions guided inferences. Each statistical analytical approach comes with its own assumption or a set of assumptions. For PCA, three assumptions underlie it. First, a fundamental assumption of PCA is that relationships among observed variables are linear. In specific terms, it is assumed that the spectrum of points in p-dimensional space has linear dimensions that can be effectively summarised by the principal axes. If the structure in the data is nonlinear (i.e. the mass of points twists and curves its way through p-dimensional space), the principal axes will not be an efficient and informative summary of the data. Linearity also implies the assumption that the data are interpretable between data points, and also implies the constraint that PCA must re-express the data as a linear combination of its basis vectors. The second assumption is that the data are of a normal (or Gaussian) distribution. This implies that the mean and variance completely describe the probability distribution of the data. Third assumption is that the data are of high signal to noise ratio (SNR). A high SNR implies that large variances in the data represent important dynamics of the system.

Factor Analysis generally assumes that only one factor explains the variance in the observed variables, i.e. ownership of assets. The common factor is taken as the measure of socioeconomic status or welfare of the household. It is also assumed that ownership of the observed assets is a linear function of the unobserved common factor for each household and the unobserved noise component (Sahn & Stifel 2000). Assumptions of Multiple Correspondence Analysis (MCA) are those of multivariate factor analysis. As factor analysis starts with a linear modelling of the set of observed variables, the non-observable variables or "common factors" are assumed to be linearly dependent from a small set of $p<m$. Normality assumptions are required for optimal model estimation (Asselin 2002; Greenacre 2000).

For Classification and Regression Tree (CART) Analysis, no any assumptions of any kind are made. No variable in CART is assumed to follow any form of statistical distribution, which is common in *frequentist* statistical analysis (Yohannes & Hoddinott 1999). Lemon, et al. (2003) point out that since classification and regression tree analysis has the statistical advantage of being a nonparametric technique, it does not invoke assumptions about the functional form of the data. Therefore, CART can be used without constraints on the distributions of the variables being investigated.

In the second set of methods Logistic Regression models were explored (Chapters 4, 5 and 6). For ordered categorical data, the assumption that the parameters corresponding to each predictor are the same for each dichotomisation of the data holds true.

This work was entirely based on secondary data collected for purposes other than determining an asset-based Household Resilience Index. The second limitation is that the datasets are from large sample surveys that contained a lot of missing values, rendering the estimates inefficient to some extent. Missing values are assumed to have mainly arisen as a result of sampling, interviewing or data capturing. A third bottleneck is that survey data are usually prone to errors of inconsistency of data collection tools, questions and responses. Fourthly, comparison based on findings across datasets might be misleading. These limitations could have introduced biases in the results. There is therefore need for further investigation, using similar methods followed here, of new data sets from future household surveys. However, multiple imputation techniques in SPSS (2013) were used to replace some of the missing data cases. In order to account for complexity of the survey designs, alternative techniques were used based on the Survey Logistic procedure and the Structured Equation Modelling, which is known to handle such issue well.

The raw survey datasets can either be found in the online data repositories of the South Sudan National Bureau of Statistics or requested directly from WFP Juba Vulnerability Assessment and Mapping (VAM) Unit.

1. For the National Baseline Household Survey 2009 First Round:

   http://ssnbs.microdatahub.com/index.php/catalog/4

2. For the Food Security Monitoring Survey 2014, either request from WFP-VAM, as the dataset was transferred non-web-based means, through their email address: Juba.VAM@WFP.org

Table 1.1: South Sudan Key Country Profile as of December 2015

Area (kilometres square)	619,745
Population	2015 Estimate: 12,340,000 2008 Census: 8,260,490 (Female: 3,973,335; Male: 4,260,490) % Under 5 (2008): Females=7.5; Males= 8.3
Population Density	13.33/km^2
Gross Domestic Product	Total: US $22.880 billion Per capita: US $1,886
Human Development Index (2014)	0.467
Main Source of Livelihood among Households	Crop farming: 71%

Literacy among people aged 6 years and above (can read and write)	Total: 28% Urban: 52% Rural: 24% Male: 38% Female: 19%
Literacy among people aged 15 years and above (can read and write)	Total: 27% Urban: 53% Rural: 22% Male: 40% Female: 16%
School Attendance (is attending or ever attended school)	Total: 37% Urban: 64% Rural: 32% Male: 47% Female: 28%
Access to healthcare facilities	Total: 70% Urban: 93% Rural: 66%
Neonatal Mortality Rate (Number of deaths per 1000 live births)	52
Infant Mortality Rate (Number of deaths per 1000 live births)	102
Maternal Mortality Ratio by State (Number of deaths per 100,000 live birth)	2054
Prevalence of stunting (moderate and severe)	33.5%

Prevalence of underweight (moderate and severe)	33.6%
Prevalence of wasting (moderate and severe)	21.6%
Agro-ecological/Agro-climatic zones	Eastern Flood Plains; Western Flood Plain; Greenbelt; Hills and Mountains; Iron Stone Plateau; Nile-Sobat River Basin; Arid (Pastoral)

***Note**: Main single source for the data is the Statistical Yearbook 2010*

The main thesis has been organized into seven chapters. Chapter 2 presents an overview of the two sets of data explored in the analysis. It also covers the main focus of the study, which is food security resilience and the rationale of how asset-based indices are valid for determining how populations in distressful food insecurity settings are likely to withstand shocks. Chapters 3 through 6 present the different tools employed in analysing two sets of survey data. Each statistical technique is introduced. Chapter gives general discussions of the work and highlights strengths and limitations of each statistical technique explored.

CHAPTER 2

Data and Exploratory Analysis

2.1. Introduction

This Chapter is subdivided into five main sections. Section 2.2 describes the datasets explored for determining predictors of food insecurity outcomes and, where necessary, for constructing an asset-based index, which is used for profiling resilience to food insecurity shocks, measuring inequalities in food security-related outcomes and predicting the probability of food insecurity occurrence. Some descriptive statistics are given of the key variables in each dataset. Section 2.3 is on the methods for calculating the relevant dependent variables explored.

Section 2.4 attempts to deepen understanding on the concept of food insecurity resilience from the perspective of the household. It also describes the rationale for asset and livelihood-based indices for assessing food insecurity risk. It presents the concept of asset-based measures of socioeconomic welfare and their advantage over money metrics, especially in settings where collecting data on the latter is prone to errors. Section 2.4 further extends the argument presented in Chapter 1 on the need for household food insecurity resilience index and its potential to becoming an overall single measure for estimating the likelihood of a household (and by extension; a community) not being capable of withstanding food insecurity risk when faced with a major shock. Section 2.5 makes conclusions based on the data exploration.

2.2. Data

Analysis in the research is based on two nationwide sample datasets collected prior and after the independence of South Sudan in July 2011. Riddled with conflict and widespread displacement of populations, the humanitarian landscape of South Sudan was typical of food insecurity and livelihoods vulnerability and, therefore, fitting the motivation for this research. Two datasets were examined, namely; the National Baseline Household Survey (NBHS) 2009 and the South Sudan Food Security Monitoring Survey conducted in late 2014. The subsequent sections describe each of the datasets and present relevant descriptive statistics based on each of them.

2.2.1. National Baseline Household Survey 2009

The National Baseline Household Survey (NBHS) was conducted in April and May 2009, by the then Southern Sudan Centre for Census, Statistics and Evaluation with the sole objective of measuring poverty and current living standards of the population based on household's total consumption. The sample size of the survey was 5280 households, which covered all the ten states of South Sudan. The survey comprehensively covered information on a range of welfare dimensions such as housing conditions, education, healthcare access, nutrition and consumption (National Bureau of Statistics 2010).

Sampling of the NBHS adapted a two-stage stratified sampling design. It used the Sudan Population and Housing Census 2008 household counts and census cartography (mapping data) as its sampling frame. The census enumeration areas (EAs) were taken as the primary sampling units (PSUs). Each EA or PSU was comprised of households ranging from 184 in urban centres and 136 in rural areas. In the first stage, of sampling, EAs were stratified by state and urban and rural areas. This resulted in random selection of 44 EAs per state. Sampling involved

selection of EAs within each stratum based on the probability proportion to size (PPS) method of estimation. In this case the size of sampled EAs was determined based on the number of households in each EA as shown in the table of the 2008 Census preliminary results. However, due to insecurity in some parts of South Sudan, some EAs could not be enumerated and were replaced with random EAs within the same geographical areas.

In the second stage, households were selected from a list of selected EAs, resulting in 12 households per EA and sample of 5,280 households. The systematic selection of the 12 households per EA employed equal probability of selection from the listing of each sample EA. In order to improve the precision of urban estimates at the national level, a higher first stage sampling rate was used for the urban stratum of each state. This is considering the fact that there was very low proportion of households in urban areas (National Bureau of Statistics 2010).

During enumeration, a multi-module questionnaire was administered that collected baseline information for the different needs of stakeholders. This was done with an intention of supplementing the analysis of poverty by also looking at non-monetary deprivations and filling certain crucial data gaps. This resulted in collection of data on: health; education; labour; housing; asset ownership; access to credit; economic shocks; transfers to household; consumption and; agriculture. The questionnaire was pretested in a Pilot survey conducted in December 2008, after which some modifications were carried out and a final version was produced and used for the actual data collection. As this survey was comprehensive enough, its dataset received the special attention for comparison of asset-based index and consumption-based index of socio-economic welfare *vis-á-vis* resilience to food insecurity shocks.

The actual sample size of 4969 households included 1546 (31.1%) urban households and 3423 (68.9%) rural households. The mean household size in urban centres was 7.23 with a minimum

of 1 and maximum of 38 members per household. In rural areas the mean household size was 6.57 and range of 1 to 48. These extremely figures of large household sizes may be raise astonishment, but in fact they are substantiated by data from other nationwide surveys such as the 2006 Sudan Household Health Survey (2008), which showed that 20 households had 25 members, two had 28 and one had 32. The numbers can also be explained by the phenomenon of rampant polygamy and wife inheritance in some cultures of South Sudan, especially the pastoralist communities of South Sudan, where a man can marry as many wives as he can afford the bride price (Stern 2011; Beswick 2001). About a third (32%) or rural households were headed by women and a similar proportion in urban centres. An overwhelming majority of rural households (80%) and half of urban households (50.7%) were headed by persons who did not attend any form of schooling. Only a quarter of urban household heads ever attended primary schools. In rural areas there were only 14.5 *per cent* household heads who attended primary schools.

Modern housing (apartments, villa, brick and brick and concrete houses) were only in 1 *per cent* of rural households and 4.7 *per cent* of urban households. Drinking of water from unsafe sources was very common in 46 *per cent* of rural households and 31 *per cent* in urban households. These unsafe water sources included shallow wells, open dams or ponds, rivers/streams and water vendors. The data show that 60 *per cent* of urban dwellers and 51 *per cent* of rural households got their drinking water from boreholes or deep water hand pumps. Piped drinking water from a water filtering grid was very rare in both rural and urban settings, with only 1.2 *per cent* and 4.4 *per cent* of households respectively.

A meagre 5.4 *per cent* of urban households used public electricity and in rural areas it was use of electricity was even a rarity (0.1%). About eight *per cent* of urban households depended on private power generation for lighting and in rural areas 0.4 *per cent* of households used

generator power. This leaves the bulk of the population to depend on paraffin (14.8%), firewood (31.0%), grass (12.3%) and candle wax (9.9%). There were about 28 *per cent* rural households and 19 *per cent* of urban households that reported they did not use any lighting at all.

A large number (84%) of rural households had no toilet facility. In urban centres, half of the households had no toilet facility. Even the cheaper forms of toilet – pit latrines – were available in only 10.7 *per cent* of rural households and 32.4 *per cent* of urban households. Urban households that used their own flush toilets were only two *per cent*, while in rural areas these were 0.2 *per cent*.

Ownership of some semi-durable assets is shown in Table 2.1. It is obvious that the increase in the number of mobile telephone providers, which have extended to all major towns in South Sudan, has caused a marked increase in ownership of telephones to 60 *per cent* in urban areas. Households' ownership of radios also increased substantially from 28 *per cent* to 54.5 *per cent* in urban areas and from 13 *per cent* to 21.9 *per cent* in rural areas. In absence of automobiles, due to affordability, bicycles were an alternative to most urban and rural dwellers with 37 *per cent* and 25 *per cent* of households possessing them – up from 33 *per cent* and 18 *per cent* respectively. Electrical appliances such as computers, refrigerators, fans and air conditioners remained very low and the reason is obvious – lack of electricity.

It is to be noted that for the purpose of the study, "urban" refers to a setting in which households sampled were drawn from areas or towns inhabited by 5000 or more households, largely clustered in typically planned urban residential areas, which also have typical urban administrative authorities such as town councils, town clerks and others. In the case of South Sudan, urban centres usually had relatively larger markets with medium to large business enterprises. In contrast, "rural" households are those in settings with no typical urban

administrative governing authorities, are where households are largely sparse, are located in unplanned or demarcated residential areas, and seldom have households in thousands. An average rural household in South Sudan lives in huts with mud walls and thatch or straw roofs. It is worth mentioning that, as the data show, a substantial section of the South Sudanese population living in urban settings could not be distinguished in terms of means or sources of livelihood, as is the case in other stable countries. For example, even in the capital city of Juba, a sizeable number of households still depended on unsafe water sources, living in typical rural house characteristics, to mention but a few characteristics.

Table 2.1: *Per cent* distribution of ownership of semi-durable assets by residential setting (National Bureau of Statistics 2010)*

Assets owned (X_i)	**Rural**	**Urban**	**Total**	**Assets owned**(X_i)	**Rural**	**Urban**	**Total**
Motor vehicle (x_1)	1.2	7.1	3.0	Radio (x_7)	21.9	54.5	32.0
Motor cycle (x_2)	2.1	10.7	4.8	Phones (x_8)	10.1	60.3	25.7
Bicycle (x_3)	25.3	37.3	29.0	Computer (x_9)	0.4	3.1	1.2
Canoe/boat (x_4)	1.6	0.9	1.4	Refrigerator (x_{10})	0.4	4.5	1.7
Animal transport (x_5)	2.4	1.2	2.0	Fan (x_{11})	0.4	6.3	2.2
Television (x_6)	1.1	18.2	6.4	Air conditioner (x_{12})	0.4	1.9	0.9

* Sample size (n) = 4,968 (Rural =3,422; Urban=1,546)

As shown in Table 2.2, about three quarters of rural households earned their livelihoods mainly from crop farming – quite a formidable leap from 39 *per cent* in the 2006 survey, while wages and salaries represented the main source of income to about 45 *per cent* of urban households. Incredibly, employed labour in 2006 was 2.3 *per cent*. This could be due to the fact that the population was still resettling only one year after the end of the civil war. The situation could have changed for the better in 2010 as more households resettled, adjusted and recovered their lives. Also probably due relative stability, petty trade or holding business enterprises increased in urban areas from a meagre 4.2 *per cent* to 13.2 *per cent*. However, rural areas saw a decrease from 3.4 *per cent* to 2.4 *per cent*. It is difficult to explain this drop. However, it could be that after the war a substantial number of enterprising people moved to urban areas where market

was more available and population was rapidly increasing. It seemed that by 2010 South Sudanese were not investing adequately in property for income purposes. The reason could be lack of capital and savings, which is understandable as the population emerged from extreme poverty exacerbated by almost four decades of conflict.

Table 2.2: *Per cent* distribution of sources of income by residential setting (National Bureau of Statistics 2010)*

Source of income (X_i)	**Rural**	**Urban**	**Total**
Crop farming (x_1)	76.5	22.6	59.8
Animal husbandry (x_3)	6.8	1.6	5.2
Wages and salaries (x_4)	7.2	44.6	18.8
Business enterprise (x_5)	2.4	13.2	5.8
Property income (x_6)	0.6	2.2	1.1
Remittance (x_7)	0.1	1.0	0.4
Pension (x_8)	0.03	1.0	0.3
Aid (x_9)	0.3	1.0	0.5
Other (x_{10})	5.3	11.5	7.2

* Sample size (n) = 4,968 (Rural =3,422; Urban=1,546)

2.2.2. Food Security Monitoring Survey 2014

The Food Security Monitoring Survey (FSMS) (2014) was conducted in August 2014 at the peak of the conflict which raged from end of 2013 through 2015. Data were collected in all ten states of South Sudan and 145 clusters as determined during the national census of 2008. In a sample size of 3,692 households, 5.3 *per cent* were internally displaced as a result of the conflict. The stratified two-stage sample selection method was used based on the sampling frame of census enumeration areas and cartographic data.

The prime purpose of the FSMS was to provide essential and baseline information for monitoring the food security situation in South Sudan during the armed conflict, in order that informed decisions were made for mitigating the situation. The United Nations and other humanitarian organisations were mandated to intervene for the nearly two million people

displaced across the country. The survey was conducted with participation of World Food Programme (WFP), Food and Agricultural Organisation (FAO), UNICEF, UNHCR, the South Sudan National Bureau of Statistics and relevant government line ministries (World Food Programme 2014).

By April 2014 the food insecurity situation in South Sudan had reached its extreme low due to widespread displacement and reduced resilience (World Food Programme 2014). As shown in Figure 2.1, food consumption levels were unacceptably high in 2014, with poor food consumption ranging from 2 to 25 *per cent* in Upper Nile State (UNS); the epicentre of the conflict-related crisis. In fact, according to the report 41 *per cent* of the households in South Sudan have inadequate food based on a seven-day recall period, while 12 *per cent* of the households had '*poor*' food consumption. In general, it was evident that the conflict worsened food consumption levels. In July 2013 11 *per cent* of the household were classified to have '*poor*' consumption. It then increased to 12 *per cent*, which could be due to the conflict.

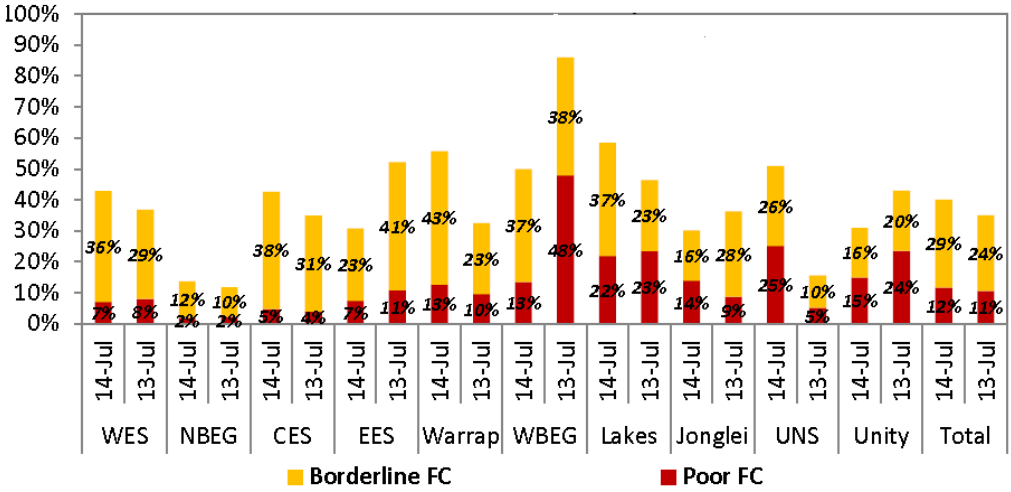

Figure 2.1: Food consumption by state in South Sudan (World Food Programme 2014)

The report also describes levels of acute malnutrition as 'critical' in most of the states affected by armed conflict: Unity, Jonglei and Upper Nile. Food consumption dominated over other household expenditure in South Sudan, reaching 76 *per cent*; an indication of distressful situation (Figure 2.2).

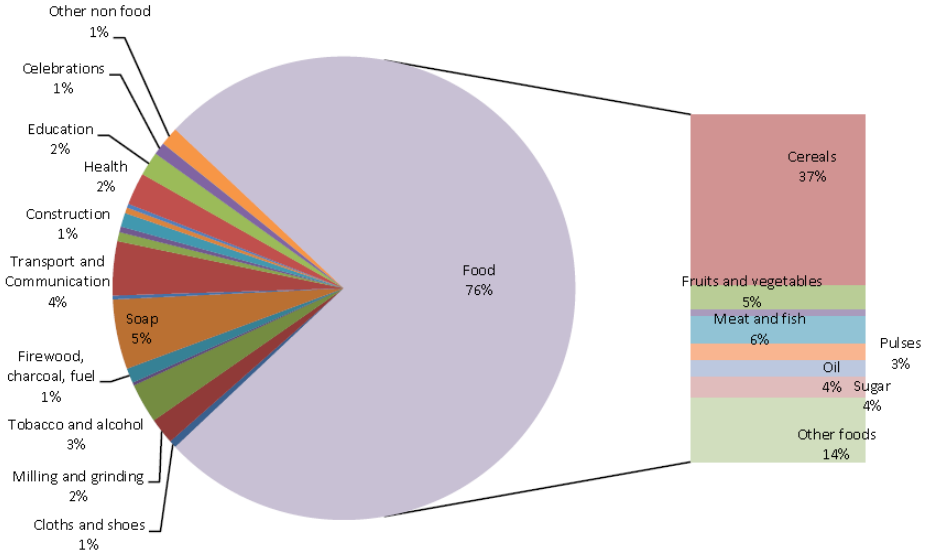

Figure 2.2: Share of household food expenditure as compared to non-food expenditure (World Food Programme 2014)

As expected, conflict was the most dominant form of shock in the four states where conflict raged with high prevalence, i.e., Unity (68%), Upper Nile (32%), Jonglei (33%) and Lakes (30%) (Table not shown). Overall, conflict represented 20 *per cent* of shocks affecting households in the country. Also as expected, soaring food prices presented another shock, aggravating the strains on households in those states. Even in the generalised low conflict affected states of Northern Bahr-el Ghazal, Warrap and Eastern Equatoria, high food prices were a major shock affecting a big number of households. Analysis showed high rates of severe to extreme forms of coping strategies (Maxwell et al. 2003), high rates of dependence on food aid (22%), with 64 *per cent* of internally displaced households receiving food assistance (World Food Programme 2014).

This study selected seven independent variables out of eight possible variables in the analysis as shown in Table 2.3 below. The seven variables were selected based on a sound rationale that they somehow affect the outcome variable under study, i.e., Food Consumption Score, Food Insecurity Coping, Share of Food Expenditure, etc. Gender of household head may be important in that food security or socioeconomic status indicator could be affected by whether a household is headed by a male or female. Intuitively, a household headed by a male is expected to cope better than one headed by a female, since typically the former combine complementary couple roles. A distinction by age group of household heads is also seen to be important in that some age groups such as that between 18 and 60 are typically economically active than the younger (less than 18 years) and older age (60 years and above) groups. Number of household members or household size is considered an important variable in the sense that a food security outcome might be influenced by contribution of some of its members to, say, access or consumption of food. A larger household size could be an advantage or a disadvantage. Ownership of a food source (livestock and crop farming) is also important variable that has variant effect on food security, nutrition or socioeconomic status outcome.

Conceptually, a household that owns cattle or other livestock is expected to exhibit better food security outcome indicator than one that does not own such asset. Main source of income from sale of food crops could tend to cope better than others. The variable not included in the analysis, which was asked in the survey is the *residential status of a household*, i.e., whether the household 'settled' before the survey, internally displaced, 'recently' returned or both internally displaced or recently returned. The reason for exclusion of the variable is that exploration of the data showed sharply skewed proportion toward the 'settled' status. Almost all households were 'settled' and only very few were displaced.

Table 2.3: Independent variables included in the analysis

Variable Description (X_{ij})	**Category (*J*)**	***n***	***Per cent***
Gender of the household head(x_{1j})	*Male*	2612	72.7
	Female	991	27.3
Age of the household head (x_{2j})	*1=(< 17 yrs)*	42	1.1
	2=(18-60 yrs)	3549	96.9
	3=(>60 yrs)	101	2.7
Size of household (x_3)	*Scale*	-	-
Cultivated crops (x_{4j})	*Yes*	2990	81.0
	No	702	19.0
Owned livestock (x_{5j})	*Yes*	3576	96.9
	No	116	3.1
Engaged in fishing (x_{6j})	*Yes*	421	11.4
	No	3155	85.5
Main source of income (x_{7j})	*Sale of crops*	1094	29.1
	Sale of livestock products	811	22.0
	Employment/labour	798	21.6
	Petty trading	774	21.0
	Other	235	6.4

It is obvious in Table 2.3 that most of the independent variables had uneven categories. Age of household head and ownership of livestock had sharp disparities. As this occurrence has even on analysis, it prompts for consideration of sampling weights, which is a subject of Chapter 4.

2.3. Derivation of the Dependent Variables

The analysis conducted explores three dependent/outcome variables, which are to be predicted by one or more salient factors determined from the set of observable assets and other independent (or explanatory) variables. The three dependent variables are: Food Consumption Score (FCS), Food Share of Total Household Expenditure and Experience of Consumption Coping with food crises. Methods for calculating each of these indices are presented below.

It should be noted that the study aim is to develop and present an appropriate, reliable and statistically robust alternative measure(s) for predicting food insecurity and to profile prevalence of food insecurity in South Sudan.

The rationale of the study, as illustrated in Figure 2.3, is grounded on the fact that the two constructs of intensity of food consumed (ascertained by food consumption score) and access to food (measured through level of expenditure on food) can be directly or indirectly determined by the level of resilience based on availability of livelihoods capitals and assets.

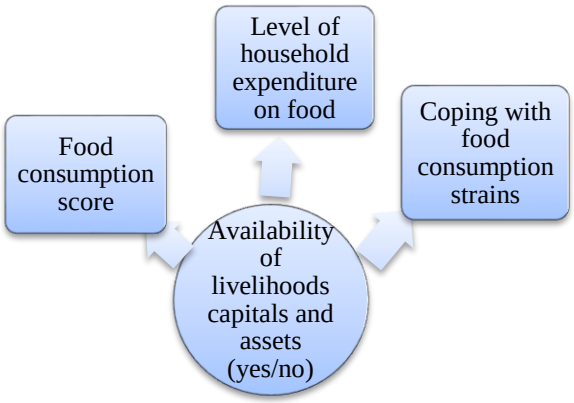

Figure 2.3: Dependence relationships between availability of household assets and three food security-related indicators (Postulation by the Authors).

Calculation of Food Consumption Score (FCS) takes into account three dimensions: type of diet consumed by a household in a week (or last 30 days), frequency of consumption and relative nutritional importance or weight (WFP VAM, 2008). FCS is an indicator of nutritional vulnerability and by extension a measure of the status of food insecurity in a household or a geographical area (Swindale & Bilinsky 2006). Mathematically, FCS can be calculated using the formula:

$$\theta_i = \sum_{j=1}^{J} w_j \, x_{ij}$$

where

θ_i = the food consumption score for the i^{th} household; $i = 1, \cdots, n$,

w_j = the nutritional weight or the number of times in a week the j^{th} food item was eaten; $j = 1, \cdots, J$,

x_{ij}= an indicator (dummy variable) of the j^{th} food item or diet consumed by the i^{th} household in a week; $\boldsymbol{x} \in 0,1$.

The nutritional weights of specific diets are shown in Table 2.4 below. WFP VAM lists nine food groups as shown in the second column. The protein rich and multi-nutrient food groups of meat or fish and milk are assigned maximum weight of 4 units. Pulses (beans, peas and nuts) follow second with a weight of 3 units, and the main staples (foods rich in carbohydrates – main source of energy) come third with nutritional weight of 2. Condiments and a pinch of salt and milk in tea are considered to have no nutritional importance and therefore a weight of zero. It is noteworthy that although intuition ruled in determining these nutritional weights, they are firmly grounded on scientific facts i.e. the value of certain diets over others in survival. For

example, a person can survive on milk or fish or eggs for many days and enjoy and full active and healthy life, than one living entirely on carbohydrate rich foods.

Table 2.4: Standard food groups and standard weights for calculation of the Food Consumption Score (World Food Programme 2008)

	Food consumption group	Food group	Weight (definitive)
1	Maize , maize porridge, rice, sorghum, millet pasta, bread and other cereals Cassava, potatoes and sweet potatoes, other tubers, plantains	Main staples	2
2	Beans, peas, groundnuts and cashew nuts	Pulses	3
3	Vegetables, leaves	Vegetables	1
4	Fruits	Fruit	1
5	Beef, goat, poultry, pork, eggs and fish	Meat and fish	4
6	Milk, yogurt and other diary	Milk	4
7	Sugar and sugar products, honey	Sugar	0.5
8	Oils, fats and butter	Oil	0.5
9	Spices, tea, coffee, salt, fish powder and small amounts of milk for tea.	Condiments	0

Calculating the food consumption scores for each household leads to a variable with values ranging between 0 and 112 units. The minimum value of 0 is arbitrary or theoretical as no household (with all members in it) can realistically stay for an entire week without eating any of the food items 1 to 8 (or possibly living on tea and other condiments for a week). The maximum score of 112 means that a household consumed each of the diets relating to the nine food groups seven days a week (or $\theta_i = 7(2+3+1+1+4+4+0.5+0.5+0)$.

After calculating the food consumption score for each household, thresholds for food consumption (or conventional food insecurity benchmarks) are obtained using the guide shown in Table 2.5.

Table 2.5: Profiling of food consumption behaviour based on the Food Consumption Score (World Food Programme 2008)

Food Consumption Score	Food security level
≤ 28	*poor*
28.1 - 42	*borderline*
42.1 – 112	*acceptable*

In mathematical notation the dependent variable *food consumption score* Y_{ij} is given by

$$Y_{ij} = \begin{cases} 1 & \text{if food security level is } 'poor' \\ 2 & \text{if food security level is } 'borderline' \\ 3 & \text{if food security level is } 'acceptable' \end{cases}$$

The three categories indicative of level of household food in/security is arbitrary and may differ according to local country situations and based on some good justification. It can be observed that the first score threshold of ≤ 28 (*poor*) amounts to about 25% of the maximal score value. This is suggestive of high risk involved when a household consumes food nutrient volume of 25% or less. Otherwise, the household members face the risk of vulnerability to hunger, and at the worst case scenario, morbidity of opportunistic diseases and even death.

Food Consumption Score is therefore a conceptually reasonable variable to be used for prediction by a household resilience measure. Technically speaking, it is logical to base inference of determination relationship on resilience measure *versus* a vulnerability indicator.

Food expenditure per capita is straightforward. The calculation of the indicator depends on recall of expenses on food items consumed, say in past one month (a shorter period is preferable in the case of low literacy communities), quantity of the item, and the type of items purchased. While it may be erratic to determine values of items consumed, it is perceivable to use standard list of prevailing prices of each item based on current market values. The total value of items

purchased (i.e. total expenditure per household) is calculated by multiplying the quantity of each item by the prevailing market price. The per capita expenditure is then determined by dividing the total household expenditure by the total number of household members. This is given mathematically as:

$$\psi_i = \frac{\sum_{j=1}^{J}(pq)_j x_{ij}}{h_i}$$

where

ψ_i= per capita expenditure for the i^{th} household; $i = 1, \cdots, n$,

$(pq)_j$= value (i.e. the product of quantity q by price p) of the j^{th} food item; $j = 1, \cdots, J$,

h_i= total number of members in the i^{th} household.

x_{ij}= a dummy or indicator variable of food item j purchased by household i; $x \in 0, 1$.

Coping Strategies Index (CSI) is a composite measure derived based on experiences of coping with incidences or food insecurity, specifically severe shortage of food or food price hikes. The index was first developed by Maxwell (1995) and from that time it became a major tool for measuring the incidence of food insecurity. Food insecurity surveillance and monitoring surveys often asked households whether they resorted to any coping strategy in the past 30 or fewer days. It is to be noted that the study specifically examined 'consumption coping'. The form of coping looked at household's options with regard to change from eating norms when confronted with severe shortage of food or non-affordability of food. The options could range from skipping of meals, to switching to less preferred foods, to going entire days without eating. For each response households were asked how often they had to adopt a certain coping

strategy. Maxwell (1995) validated a framework for scoring each coping strategy (Table 2.6), which was then used to calculate the Coping Strategies Index.

Table 2.6: Framework for calculating the Consumption Coping Strategies Index (Maxwell, 1995)

Variable	Severity Weight	Relative Frequency	Score
1. Dietary Change:			
a. Rely on less preferred /less expensive foods	2	xx	xx
2. Increase Short-Term Food Availability:			
b. Borrow food/rely on help from a friend or relative	4	xx	xx
c. Purchase food on credit	4	xx	xx
d. Gather wild food, hunt, or harvest immature crops	8	xx	xx
e. Consume seed stock held for next season	6	xx	xx
3. Decrease Numbers of People:			
f. Send children to eat with neighbours	4	xx	xx
g. Send household members to beg	8	xx	xx
4. Rationing Strategies:			
h. Limit portion size at mealtimes	2	xx	xx
i. Restrict consumption by adults in order for small children to eat	6	xx	xx
j. Feed working members of HH at the expense of non-working members	4	xx	xx
k. Ration the money you have and buy prepared food	2	xx	xx
l. Reduce number of meals eaten in a day	2	xx	xx
m. Skip entire days without eating	8	xx	xx

The multiple options for relative frequency are: everyday, 3-6 times a week, 1-2 times a week, once a week, or never, score as 7, 4.5, 1.5, 0.5, 0, respectively. A score for each type of coping strategy is calculated by multiplying its severity weight by the corresponding relative frequency score. The scores are then summed up to give the total score for the household. Finally, to ease interpretation of the coping indicator, *per cent*iles of total household coping strategy is obtained to transform it into an index. A classification scheme can then be devised such that the

households are distributed into a number of categories to determine their humanitarian phase, such as “food secure”, “borderline food secure”, “marginally food insecure”, “severely food insecure” and “extremely food insecure”. Note that a household that did not adopt any coping strategy in any single day of the month or in the lowest *per cent*ile of coping score, could, in relative terms, imply that there is “no threat” of food insecurity. Meanwhile, a household with a *per cent*ile of 60 and above could be facing the risk of being in a humanitarian crisis. Table 2.7 is an example of classification of households into *per cent*iles of coping scores obtained as described above and suggests classifications according to food in/security phases and cautionary actions to be taken.

Table 2.7: Food insecurity classification guide

CSI *Per cent*ile	Food Insecurity Phase	Warning Stage
0 – 25	Food secure	No threat
25 – 40	Borderline	Tolerable
30 – 40	Marginal	Watch
40 – 50	Moderate	Alert
50 – 60	Chronic	Alarm
>60	Severe	At High Risk

2.4. Asset-based measures of socioeconomic welfare

Development economists have measured socio-economic status at the national level using macroeconomic scales of income, consumption and expenditure. Income distribution from aggregate statistics, such as the Gini Coefficient and GDP per capita, was used. Measures based on income and expenditure introduced the so-called poverty line or threshold (Falkingham & Namzie 2001) to delineate the poor from the non-poor (or richer) sub-populations. On the household level, measures such as purchasing power parity (PPP) and the so-called

international poverty lines (IPL) are used in welfare analysis such as those by the World Bank. It is a common practice across many countries to measure IPL based on food and non-food expenditure components, with the food component derived based on specific energy requirement, using data from household expenditure surveys (Asian Development Bank 2008a).

The monetary (or money-metric) measures are referred to by the World Bank (2004) and Balen et al. (2010, p.2) as 'direct' measures of socioeconomic status (SES), defined in the domains of income and financial assets such as savings and pensions. Such SES measures are referred to as 'standard' (Vyas & Kumaranayake 2006) or 'conventional' (Booysen 2002). Moser and Felton (2007, p.1) justify the preference of income as a unit of welfare analysis in that income is 'a cardinal variable' that enables comparison between observations and that is 'straightforward to interpret and use in quantitative analysis'. Other analysts such as Ravallion (1992) strongly advocated the use of consumption-based measures of SES.

However, a number of arguments abound in the literature regarding the shortcomings of income-based poverty measures. For instance, Baullch and Hoddinott (2000) and Hulme and Shepherd (2003) argue that income-based measures of SES tend to indicate high levels of transitory poverty and underestimate chronic poverty. Gwatkin et al. (2000) argue that although an income measure used in assessment of economic inequalities in health is traditionally the preferred indicator of economic status, it lends itself to 'well-known difficulty'. They cite an example of survey informers being reluctant in disclosing their incomes as compared to giving information on social status such as religion, occupation or educational levels. They point out that this has led to the use of proxy of income in the form of social status variables.

Sahn and Stifel (2003) observe that, while developing countries have generally based measurement of socioeconomic standards on income, an aggregate of household's

consumption expenditures has largely remained the preferred metric measure of choice. This choice is said to have been dictated by the number of constraints encountered, such as seasonal variability in earnings and self-employment (Sahn & Stifel 2003).

On the expenditure side, Liverpool and Winter-Nelson (2010, p.3) observe that although expenditure-based measures indicate households' vulnerability to various shocks, they also 'complicate efforts to attribute poverty reduction in specific households to intervention'. They also observe that another complication is the tendency of some household expenditure to rise above or fall below the poverty line, along the time scale, regardless of policy interventions. According to them, this pattern is likely to occur, especially among farming communities in developing countries, where household income is a function of such influences as weather conditions, crop yields and commodity prices.

Money metric measures of poverty and socioeconomic welfare in developing countries are generally regarded as causes of measurement issues to economists and development analysts. Sahn and Stifel (2003) highlight and demonstrate a number of such problems. Topping the list of the limitations, is the issue of recall by household respondents of income earned or expenditures incurred on certain items in a recall period exceeding, say, two weeks. Recall lapses are usually prone to measurement errors, most of which are random (Sahn & Stifel 2003). A case in point is when the list of commodities on the recall sheet is long (Pradhan 2000). Scott and Amenuvegbe (1990) report that the longer the recall period the lower the reported consumption by households. Another major glitch is that of obtaining values for each of the items consumed. These values are normally commodity prices or, in rare instance, nominal interest rates and depreciation rates of semi-durable or durable assets (Sahn & Stifel 2003). Such data is difficult to obtain in developing countries. Added to this glitch, according to Sahn and Stifel (2003), is the near-impossibility of obtaining rental price equivalents,

especially in rural settings of Africa, where there is virtually no house rental market. Other pertinent problems highlighted in the literature concern the absence of price deflators and exchange rate distortions in developing economies (Sahn & Stifel 2003).

Countries characterised by protracted crises such as South Sudan make them to be perceived as manifesting peculiar socio-economic and food security characteristics; thus the need for baseline information. At the core of this study, therefore, is the need to identify baseline levels of resilience to food insecurity uncertainties in South Sudan.

Of recent there has been a marked shift in thinking with regard to focusing more attention to people's resilience strengthening than on vulnerability and relief (Pasteur 2011). This shift would require a shift to adopting measures for gauging how much resilience is required to withstand crisis. This is what makes this study relevant and of substance. The need to shift from concentrating measurement of food insecurity and malnutrition to measuring resilience of populations in situations of distress is fast becoming relevant and urgent, as experience over the last three decades has shown that food insecurity and what causes it keeps escalating. Otherwise, it is synonymous with concentrating on measuring the magnitude of ill health, while neglecting prognostic factors that make people become resistant to diseases and ill health and, thereby, informing authorities to allocate more resources to the areas that improve those positive influencing factors.

Against the dim light, as man-made and natural shocks or strains are gaining momentum, it is critical to concentrate efforts on measuring resilience, given its intrinsic value of cushioning against future vulnerability. In general, resilience enhancement is more a developmental strategy than the traditional humanitarian relief and rehabilitation. For more arguments along this line, see Barret and Maxwell (2005), Barrett and Heisey (2002) and Maxwell (1996).

It is evident that food aid organisations seem to shy away from determining resilience assessments and enhancement interventions, apparently on three grounds. First, resilience building requires a multi-dimensional and multi-sector approach. Improving resilience is mostly a function of long-term developmental strategies, rather than short-term actions, in order to bear impact. It is, therefore, seen to fall within the domain of long-term state development plans. Secondly, resilience enhancement measures and activities are seen to fall outside the fundamental mandate of humanitarian aid organisations. Thirdly, humanitarian aid organisations are more concerned with addressing and arresting the severe cases of food emergency such as famine, severe malnourishment, deaths (The Johns Hopkins and International Federation of Red Cross and Red Crescent Societies, 2004). This then makes measurement and monitoring of vulnerability more appealing than measuring resilience. Yet, according to Mousseau (2005, p.13), "food aid undermines local agricultural production", among several other effects.

The shift in focus on resilience has been necessitated by the need to control risk and prepare against the effects of emergencies. This shift away from earlier focus on vulnerability analysis (i.e., *ante-hoc* measurement) to more predictive and extrapolative (*post-hoc* measurement) analysis is essential for offering precautionary solutions.

Existing measurement of household food in/security and vulnerability has been approached through the use of a number of indicators. Measurement was done by analysing availability of resources, food and assets, examining harvests, food consumption, coping options and nutritional indicators. These diverse measurements make analysis complex, based on multiple sources of data, tends to give rather belated information for decision making and, most discouragingly, does not provide indications of the likelihood of future occurrence of food insecurity. In examining the spectrum of determinants of food insecurity, namely; *risk*,

resilience and *vulnerability*, and the dichotomy of their causes and effects (see Figure 3.1), the dilemma of existing measurement approaches can be seen in that they overly dwell on *risk* and *vulnerability*. In other words, the methods tend to be backward-looking or they yield indicators pointing at what has already occurred, rather than what might occur.

Interpretation of Figure 3.1 is somewhat easy. It shows that food security status is a function of the level of strength of vulnerability, which is in turn a function of the amount of resilience households exert to buffer against livelihood risks (hazards). The diagram explains that food security is attained when occurrence of risk is absorbed by strong resilience that lowers vulnerability. Conversely, a food insecurity outcome might occur when a household becomes highly vulnerable as a result of external hazards or when the household has weak resilience to those hazards.

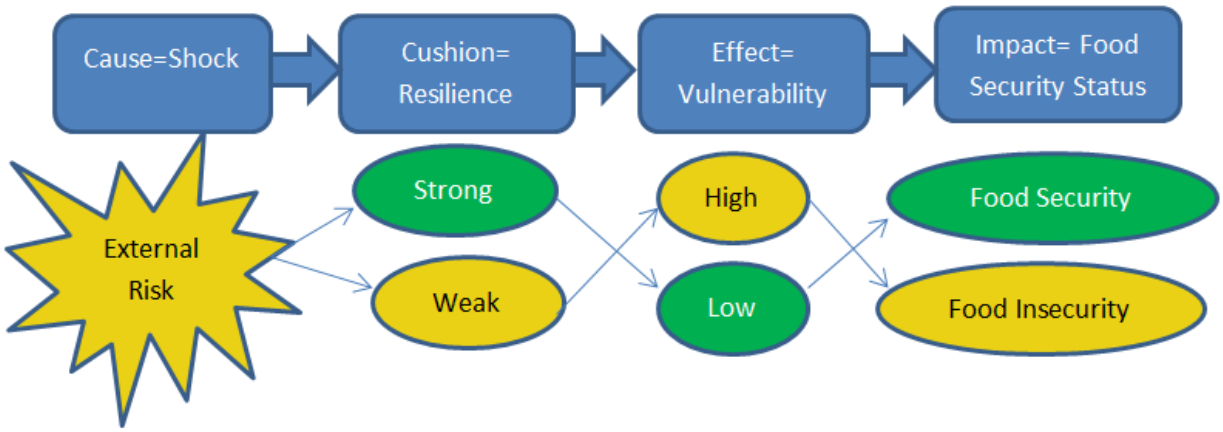

Figure 3.1: Conceptualisation of Food Insecurity Resilience (Author postulation)

Measurement aimed at determining risk from the household viewpoint will not yield desired results as occurrence of risk is a probability and therefore depends on a number of assumptions. Such measurement requires statistical modelling in order to determine the possibility of risk. Imperatively, this class of measures would only be conducted by engaging expert analysts rather than field officers and thus could be costly. Measuring vulnerability implies that the

household has already been affected. Customarily, household vulnerability has been determined through assessment of levels of household food insecurity. Resilience, on the other hand, is looked at from the perspective that it is a function of livelihoods capitals (Lokosang et al. 2014). There are five livelihood capitals that not only enable a household to sustain its livelihood, but also to withstand livelihood risks. These are: human, natural, financial, physical and social (Department for International Development 1999).

As better resourced households become more resilient than poorly resourced ones, the rationale to measuring how weak or how strong households will be is based on the available livelihood capitals. Thus, acting together or individually, these livelihood capitals directly influence livelihood outcomes as well as vulnerability.

Strong resilience boosted by livelihood factors is perceived to enable households to withstand what is likely to occur. It prevents or buffers the household from potential risk. Since strong or low resilience prevents or heightens vulnerability, there is need to find ways to determine which segments of a population have low or strong resilience. It, therefore, makes sense to use progressive (or forward-looking) analytics rather than deterministic approaches to investigating what has already occurred. In this sense, measuring resilience makes it possible to provide evidence for counteractive action or *a-priori* decision. Thus, the set of analysis explored in this research project could provide useful tools for food security and livelihood analysis to make relevant, and more so, assist in ongoing early warning and early preparedness interventions.

Since livelihood capitals can influence or determine livelihood outcomes such as sustainable health, socioeconomic wellbeing, vulnerability and food insecurity, measurement of how the strength of the relationship between the two sets of variables gains relevance. This study is,

therefore, motivated by the portrayal that inequalities in levels of livelihood capitals can be a proxy to potential food insecurity risks.

2.5. Conclusion

From the descriptive examination of data it was clear that South Sudan is entrapped in asset chronic poverty and food insecurity. Exploratory analysis revealed there were persistent manifestations of low resilience to economic and food insecurity shocks. The preliminary results give reason to explore more rigorous statistical methods in the subsequent chapters of this thesis.

CHAPTER 3

Indexing and Latent Variable Classification

3.1. Introduction

This chapter sets out to apply three multivariate techniques for identifying likely determinants of food insecurity risk and for generating a household food insecurity resilience index based on a set of household characteristics, livelihood capitals and endowments. The chapter draws its rationale from the notion that resilience to food insecurity is a property of wealth and thus its proxy. Analysis explored the statistical robustness and efficiency of the techniques in providing evidence for triggering alerts and action for curbing risk of food insecurity uncertainties. The chapter is structured into five sections. Section 3.2 describes the dataset and methods used. Sections 3.3, 3.4 and 3.5 present the methods of analysis featuring Principal Component Analysis (PCA), Multiple Correspondence Analysis (MCA) and Classification and Regression Trees (CART), respectively.

3.2. Sample and Data

The dataset explored is drawn from the National Baseline Household Survey (NBHS) described in Section 2.2.1. Table 3.1 presents some descriptive statistics of semi-durable and durable assets, housing conditions and characteristics in South Sudan in 2009. It also displays the scoring factors from the principal component analysis of the 33 variables.

Table 3.1: ***Per cent*** **distribution of ownership of assets** (National Bureau of Statistics 2010) **(n = 4968)**

Assets owned (X_i)	Relative Frequency (%)	Mean	Standard Deviation	Scoring Factor*
Semi-durable Assets				
Motor vehicle (x_1)	3.0	0.03	0.166	0.025
Motor cycle (x_2)	4.8	0.05	0.210	0.040
Bicycle (x_3)	29.0	0.29	0.454	0.106
Canoe/boat (x_4)	1.4	0.01	0.118	0.003
Animal transport (x_5)	2.0	0.02	0.139	-0.001
Television/sat. dishes (x_6)	6.4	0.06	0.241	0.051
Radio (x_7)	32.0	0.32	0.467	0.124
Phones (x_8)	25.7	0.26	0.437	0.101
Computer (x_9)	1.2	0.01	0.110	0.012
Refrigerator (x_{10})	1.7	0.02	0.123	0.017
Fan (x_{11})	2.2	0.02	0.145	0.021
Air conditioner (x_{12})	0.9	0.01	0.087	0.008
Sources of Income				
Crop farming (x_{13})	59.8	0.60	0.489	-0.045
Animal husbandry (x_{14})	5.2	0.05	0.221	-0.001
Wages and salaries (x_{15})	7.2	0.19	0.392	0.052
Business enterprise (x_{16})	5.8	0.06	0.233	0.003
Property income (x_{17})	1.1	0.01	0.104	0.002
Remittance (x_{18})	0.4	0.00	0.060	0.000
Pension (x_{19})	0.3	0.00	0.054	0.001
Aid (x_{20})	0.5	0.01	0.074	-0.002
Other source (x_{21})	7.2	0.07	0.262	-0.011
Housing characteristics				
Permanent dwelling (x_{22})	5.2	0.05	0.220	0.033
Semi-permanent dwelling (x_{23})	88.9	0.90	0.304	-0.029
Temporary dwelling (x_{24})	5.1	0.05	0.222	-0.004
Total number of rooms (x_{25})	---	2.51	1.518	1.512
Drinking water from pump/well (x_{26})	57.7	0.58	0.493	0.013
Drinking water from open source (x_{27})	36.3	0.37	0.482	-0.015
Drinking water from other source (x_{28})	5.1	0.05	0.223	0.002
Electricity for lighting (x_{29})	4.6	0.04	0.205	0.035
Cooking energy gas/electricity (x_{30})	0.4	0.00	0.065	0.004
Pit latrine (x_{31})	24.7	0.25	0.430	0.138
Flush toilet (x_{32})	1.1	0.01	0.100	0.007
No/other toilet (x_{33})	73.9	0.74	0.436	-0.145

*Scoring factors are composite variables which provide information about an individual's placement on the factor(s). The scoring factor coefficients are estimated using the Regression Method. They have a mean of zero and variance equals to the squared multiple correlations between the estimated factor scores and the true factor values. The scores may be correlated even when factors are orthogonal.

The interpretation of the information presented in the second column of Table 3.1 simply informs about the relative frequencies of household assets, endowments, conditions and

livelihood capitals in the sample. This is the type of result most surveys produce. It explains how certain assets are owned by more households than others at the time of data collection. For example, we learn that more households (32%) owned radios while fewer (5.2%) had livestock. It cannot be known for certain that owning a radio is an indicator of wealth and, therefore, that a household is resilient to food insecurity. In other words, the percentages cannot be a proxy for wealth or showing a consumption pattern. From intuition, although there were much fewer households that had livestock (animal husbandry), they could be relatively well off or enjoying much higher resilience to food insecurity shock or strain than those owning a radio only. This is based on the simple fact that a household with a stock of animals could readily sell or consume from it than one owning only a radio or bicycle and, therefore, "bounce back" economically. Percentages are calculated taking into account that assets have equal weights, which in itself presents arbitrariness and lacks statistical strength.

Considering the fact that the data came from a sample survey, it was important to carry out some validation. This involved quality checking the data for incomplete household identification numbers, variable values that were out of range and combinations of variable values that might have been entered in error. Data validation was carried out using IBM SPSS version 23 in which variables were specified that uniquely identified households, defined single variable rules for the valid variable ranges, and defined cross-variable rules to identify impossible combinations. The procedure then produced a report of the cases and variables determined to have problems. The errors were rectified or the problem case eliminated. After completing the check and cleaning the data, the procedure produced a report showing that the analysis variable passed the basic validation checks, as they were no empty cases detected.

The analysis explored in this chapter could best be validated by results of analysis with data from similar baseline survey data. However, due to a civil conflict that started at the end of 2013 and financial strains following an economic meltdown, this was not possible.

3.3. Principal Component Analysis

Based on deduction by Filmer and Pritchett (2001), Principal Component Analysis (PCA) is used to construct an asset index that proxies for wealth and long-run socio-economic status; thus resilience to food insecurity shocks or stresses. PCA is a mathematical approach that derives the weights for each asset based on certain latent variables known as "Principal Components".

PCA is described as a simple non-parametric method that reduces a complex dataset to a lower dimension of variables. PCA is defined as a linear combination of optimally weighted observed variables. The procedure simply aims at reducing variables to a small number of components that account for most of the variation in a set of observed variables. The concept of PCA is built on the assumption that some of the observed variables are correlated with one another (O'rourke et al. 2005).

The general form of the formula for computing the first principal component extracted from p variables is:

$$\mathrm{C_1} = \mathrm{b_{11}(X_1)} + \mathrm{b_{12}(X_2)} + \cdots + \mathrm{b_{1p}(X_p)},$$

where, $\mathrm{C_1}$is the subject's score on the first principal component, b_{1j} is the weight for observed variable j on the first component and X_j = the observed variable j. The strategy of PCA is to obtain total variation by standardising the observed variables. This is done by transforming each variable so that it has a mean of zero and a standard deviation of one. Then the variances

of the observed variables are summed up such that each observed variable contributes one unit of variance to the total variance in the dataset. This makes the total variance in a PCA to always equal the number of observed variables being analysed.

Principal Components can be derived in more than one way. The simplest method is by finding the projection which maximizes the variance. Conceptually, the aim is to look for the projection with the smallest average (by squaring the mean) distance between the original vectors and their projections onto the principal components. This is the equivalent to maximizing the variance. The overriding assumption is that the data have been "centred", so that every one of the factor has mean 0.

If we write the standardized data in a matrix **X**, where rows are objects and columns are factors, then $\mathbf{X^T X = nV}$, where $\mathbf{V}$ is the covariance matrix of the data. Two steps are essential in deriving the Principal Components:

First, to minimizing the component residuals we look for a one-dimensional projection. That is, we have **p**-dimensional factor vectors, and we aim to project them on to a line through the origin. We can specify the line by a unit vector along it, $\bar{\mathbf{w}}$, and then the projection of a data vector $\bar{\mathbf{x}}_\mathbf{i}$ on to the line is $\bar{\mathbf{x}}_\mathbf{i}.\bar{\mathbf{w}}\cdot$, which is a scalar.

This is the distance of the projection from the origin; the actual coordinate in p-dimensional space is$(\bar{\mathbf{x}}_\mathbf{i}.\bar{\mathbf{w}})\bar{\mathbf{w}}$. The mean of the projections will be zero, because the mean of the vectors $\bar{\mathbf{x}}_\mathbf{i}$ is zero.

$$\frac{1}{n}\sum_{i=1}^{n}(\bar{\mathbf{x}}_\mathbf{i}.\bar{\mathbf{w}})\bar{\mathbf{w}} = \left(\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_\mathbf{i}\right).\bar{\mathbf{w}}\right)\bar{\mathbf{w}} \qquad (3.1)$$

For any one vector, say, $\bar{x}_i$, it's

$$\|\bar{\mathbf{x}}_\mathbf{i} - (\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})\bar{\mathbf{w}}\|^2 = \|\bar{\mathbf{x}}_\mathbf{i}\|^2 - \mathbf{2}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i}) + \|\bar{\mathbf{w}}\|^2 \quad (3.2)$$

$$= \|\bar{\mathbf{x}}_\mathbf{i}\|^2 - \mathbf{2}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2 + \mathbf{1} \quad (3.3)$$

Adding all those residuals up across all the vectors:

$$\mathbf{RSS}(\bar{\mathbf{w}}) = (\textstyle\sum_{\mathbf{i=1}}\{\|\bar{\mathbf{x}}_\mathbf{i}\|^2 - \mathbf{2}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2 + \mathbf{1}\} \quad (3.4)$$

$$= (\mathbf{n} + \textstyle\sum_{\mathbf{i=1}}^{\mathbf{n}}\|\bar{\mathbf{x}}_\mathbf{i}\|^2) \; - \mathbf{2}\textstyle\sum_{\mathbf{i=1}}^{\mathbf{n}}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2 \quad (3.5)$$

The term in the big parenthesis does not depend on $\bar{\mathbf{w}}$, so it does not matter trying to minimize the residual sum of squares. To make the RSS small, the term subtracted from it must be made big. That is, we maximize

$$\sum_{\mathbf{i=1}}^{\mathbf{n}}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2$$

Similarly, since n does not depend on $\bar{\mathbf{w}}$, we aim to maximize

$$\tfrac{\mathbf{1}}{\mathbf{n}}\textstyle\sum_{\mathbf{i=1}}^{\mathbf{n}}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2,$$

which is the sample mean of $(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2$. The mean of the square is always equal to the square of the mean plus the variance:

$$\tfrac{\mathbf{1}}{\mathbf{n}}\textstyle\sum_{\mathbf{i=1}}^{\mathbf{n}}(\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i})^2 = \left(\tfrac{\mathbf{1}}{\mathbf{n}}\bar{\mathbf{x}}_\mathbf{i}.\bar{\mathbf{w}}\right)^2 + \mathbf{Var}[\bar{\mathbf{w}}.\bar{\mathbf{x}}_\mathbf{i}] \quad (3.6)$$

We can see that the mean of the projections is zero. Therefore, minimizing the residual sum of squares is the equivalent of maximizing the variance of the projections. It should be noticed that, in general, we do not want to project onto just one vector, rather to multiple components.

If those components are orthogonal and have the unit vectors $\bar{\mathbf{w}}_1, \bar{\mathbf{w}}_2 \cdots, \bar{\mathbf{w}}_k$, then the image of $\mathbf{x_i}$ is its projection into the space of these vectors,

$$\sum_{j=1}^{k} (\mathbf{x_i}. \bar{\mathbf{w}}_i)\, \bar{\mathbf{w}}_j$$

The mean of the projection on to each component is still zero.

The second step is to maximize the variance. If the n data vectors are stacked into an $\boldsymbol{n} \times \boldsymbol{p}$ matrix, i.e., **X**, then the projections are given by **Xw**, which is an $\boldsymbol{n} \times \mathbf{1}$ matrix. The variance is

$$\sigma^2_{\bar{\mathbf{w}}} = \frac{1}{n}\sum_i (\bar{\mathbf{x}}_i . \bar{\mathbf{w}})^2 \qquad (3.7)$$

$$= \frac{1}{n}(\mathbf{Xw})^T(\mathbf{Xw}) \qquad (3.8)$$

$$= \frac{1}{n}\mathbf{w}^T\mathbf{X}^T\mathbf{Xw} \qquad (3.9)$$

$$= \mathbf{w}^T \frac{\mathbf{w}^T\mathbf{Xw}}{n}\mathbf{w} \qquad (3.10)$$

$$= \mathbf{w}^T\mathbf{Vw} \qquad (3.11)$$

Now, to choose a unit vector $\bar{\mathbf{w}}$, we need to constrain the maximization. The constraint is that $\bar{\mathbf{w}}.\bar{\mathbf{w}} = \mathbf{1}$**,** or $\mathbf{w}^T\mathbf{w} = \mathbf{1}$. This necessitates constrained optimization. The first step is to maximize the function **f(w) (=$\mathbf{w}^T\mathbf{Vw}$**) given the equality constraint, **g(w) = c** , where $\mathbf{g}(\mathbf{w}) = \mathbf{w}^T\mathbf{w}$ and **c = 1**. Second step is to rearrange the constraint equation so that its right hand side is zero and **g(w)-c=0**. Next step is to add an extra variable, the Lagrange Multiplier **λ** to obtain our objective function **u(w,λ) =f(w)+λ{g(w)-c}**. We then differentiate with respect to both arguments and set the derivatives equal to zero.

$$\frac{\partial \mathbf{u}}{\partial \mathbf{w}} = \mathbf{0} = \frac{\partial \mathbf{f}}{\partial \mathbf{w}} + \boldsymbol{\lambda}\frac{\partial \mathbf{g}}{\partial \mathbf{w}} \tag{3.12}$$

$$\frac{\partial \mathbf{u}}{\partial \boldsymbol{\lambda}} = \mathbf{0} = \ \mathbf{g}(\mathbf{w}) - \mathbf{c} \tag{3.13}$$

It can be seen that the objective function is maximized with respect to $\boldsymbol{\lambda}$ to obtain the constraint equation, **g(w)=c**. Having satisfied the constraint, the new objective function equates to the old one. To derive our projection problem,

$$\mathbf{u} = \ \mathbf{w^T V w} - \boldsymbol{\lambda}(\mathbf{w^T w} - \mathbf{1}) \tag{3.14}$$

$$\frac{\partial \mathbf{u}}{\partial \mathbf{w}} = \ \mathbf{2Vw} - \mathbf{2}\boldsymbol{\lambda}\mathbf{w} = \mathbf{0} \tag{3.15}$$

$$\mathbf{Vw} = \ \boldsymbol{\lambda}\mathbf{w} \tag{3.16}$$

Thus, the desired vector **w** is an eigenvector of the covariance matrix **V** and the maximizing vector transform to the vector associated with the largest eigenvalue $\boldsymbol{\lambda}$.

V is a $\mathbf{p} \times \mathbf{p}$ matrix, so it will have p different eigenvectors. **V** is a covariance matrix, so it is symmetric, and in linear algebra terms, the eigenvectors must be orthogonal to one another. The second principal component is the direction with the most variance, which is orthogonal to the first principal component. Thus, the second principal component is the eigenvector of **V** corresponding to the second largest eigenvalue, and so on. Since it is orthogonal to the first eigenvector, their projections will be uncorrelated. In general, all principal components have projections which are correlated with each other. If k principal components are used, the weight matrix will be a $\mathbf{p} \times \mathbf{k}$ matrix **V**. The eigenvalues will give the share of the total variance described by each component.

The main motivation for using PCA in the analysis was that, apart from its being established as a good measure of socioeconomic wellbeing by researchers such Filmer and Pritchett (2001), it was ascertained to be helpful in constructing a summary measure (referred to here as Household Resilience Index or HRI in short), thus an efficient proxy for wealth index, which is based on consumption data, and which could well predict per capita consumption.

3.3.1. Results of the PCA Procedure

This section shows application of the PCA methods to the set of data from the South Sudan National Baseline Survey. Once again, the purpose was to generate an asset based index for measuring the amount of resilience exerted by households in the study population. The factor extraction method (PCA) was used to form uncorrelated linear combinations of the observed variables. The first extracted or principal component is the one with maximum variance and successive components explain progressively smaller portions of the variance and are all uncorrelated with each other.

We examined data on the 33 variables as listed in Table 3.1. The values of each variable were dichotomized (transformed into binary categories by collapsing the original categories of the variable) – except for the number of household members – to assign indicator values for each household. The SPSS Factor Analysis procedure is used to calculate ***z***-scores by standardizing the indicator variables. This then led to obtaining factor loadings and virtually the household index values. Finally, the first of the factors generated was then used as the wealth index. Principal Component Analysis (Table 3.2) used here, resulted in the first component extracted, although it explained only about 24% of the variability in the original 33 variables. As shown in Table 3.2, the first component carried far better weight (inertia) in the way of explaining variability than the subsequent extracted components. The first component reasonably

explained adequate amount of variance and was thus selected as our Household Food Insecurity Resilience Index (HRI).

Table 3.2: Variation explained by extracted Principal Components

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative
1	0.712	23.876	23.876
2	0.437	14.657	38.533
3	0.336	11.277	49.810
4	0.284	9.532	59.342
5	0.210	7.028	66.370
6	0.145	4.852	71.223
7	0.132	4.415	75.638
8	0.111	3.730	79.369
9	0.093	3.101	82.470

The asset index proxies for household wealth but it is also the natural Household Resilience Index (HRI), as affordability of certain assets, the value of some livelihood capitals as well as presence of certain household characteristics and endowments may enable the household to become resilient in face of food insecurity uncertainties and eventualities. The HRIs were grouped into quintiles to form five resilience categories: '*very weak'* (the household scores from 0 to the 20th *per cent*ile); '*weak*' (the household scores from the 21st to the 40th *per cent*ile), '*moderate*' (the household scores above 40 to the 60th *per cent*ile); '*high*' (the household scores above 60 to the 80th *per cent*ile and '*strong*' (household scoring from the 80th *per cent*ile and above).

A household with '*very weak*' resilience implies that it is very likely to face severe livelihood strains and resorting to extreme coping strategies as defined in Maxwell (1995; 2003). Such household has weak asset base at its disposal and during harder times characterised by food shortage or depleted resources as to afford food during scarcity, would have little or nothing to

dispose of so as to afford food. On the other hand, a household determined to be in '*strong*' resilience group, had good asset base to buffer it against food and humanitarian crisis.

As one of the aims of this study was to determine resilience profiles in South Sudan, this was done by cross-matching the resilience levels against states on one hand, and against residential setting (i.e. urban and rural), on the other. It is, however, to be noted that as the country was in a post-conflict stage, following a two-decade civil war, living conditions between rural and urban populations were basically similar. Separate analysis carried out on the same dataset showed that both populations fared equally in most comparisons involving livelihood conditions, such as dependence on firewood for cooking energy, reliance on unsafe drinking water sources, non-use of modern toilet facilities and living in houses constructed from rudimentary materials. A cross-tabulating of the HRI levels by state (Table 3.3) showed clear disparities between states – reflecting the past and present reality in South Sudan.

Table 3.3: State resilience profiles in terms of Household Resilience Index (South Sudan, 2009)

	Household Resilience Index Quintiles (%)				
State	Very Weak	Weak	Moderate	High	Strong
Upper Nile	16.5	19.2	**23.5**	19.9	20.9
Jonglei	22.6	**37.7**	20.0	14.5	5.2
Unity	23.3	23.1	22.5	18.3	12.8
Warap	**32.6**	23.9	21.3	14.4	7.8
Northern Bahr Al Ghazal	**25.8**	22.9	24.2	16.9	10.3
Western Bahr Al Ghazal	16.9	7.0	25.3	18.3	**32.5**
Lakes	**27.0**	12.4	**26.4**	20.9	13.2
Western Equatoria	6.3	2.9	8.4	**43.4**	**39.0**
Central Equatoria	10.2	10.0	14.2	**23.2**	**42.5**
Eastern Equatoria	**39.5**	**25.9**	12.1	9.8	12.7

The percentages shown in Table 3.3 were obtained by classifying the generated HRI (factor loadings of the first component) in five bins or quintiles using the SPSS RANK VARIABLES

command. The HRI quintiles are then cross matched against the variable state to obtain counts and percentages in each state.

Central Equatoria, Western Equatoria and Western Bahr Al Ghazal States were better off with over 30 per cent of their households indicative of 'strong' resilience to food insecurity shocks. In contrast, five states (Jonglei, Warap, Northern Bahr Al Ghazal, Lakes and Eastern Equatoria) had a generalized 'weak' resilience. One state, Upper Nile, had a generalized moderate resilience. These results were typical of known realities of the country at the time of the survey. The three states categorized as 'strong' in term of resilience to food insecurity were characterized by generally agrarian populations, who largely depend on agriculture as their source of income. They are also located in the 'Green Belt' agro-ecological zone according to categorisation of livelihood profiles in South Sudan. As the name suggests, conditions in the Green Belt zones favour agricultural production and sustained livelihoods as the area is located a few latitude degrees above the Equator, have rich porous and iron-stone soil and a mean annual rainfall of 1,800mm per year (National Bureau of Statistics 2010).

Another aspect that characterised states in South Sudan was their occupation by relatively urbane, stable and educated populations compared to the seven other states. Juba, the current Capital City of South Sudan, combined as the administrative Capital of Central Equatoria State. Wau, the second largest town, was the administrative Capital of Western Bahr Al-Ghazal State with more people who did not migrate or who were not displaced, as it had remained under control of government forces during the two decade civil war of the undivided Sudan. The states with generalized 'weak' resilience, on the contrary, had more rural and returning populations from internal displacement and exile. People were beginning to settle roughly three years into the return of peace in the country. These states were also relatively new and the population was predominantly comprised of pastoralists.

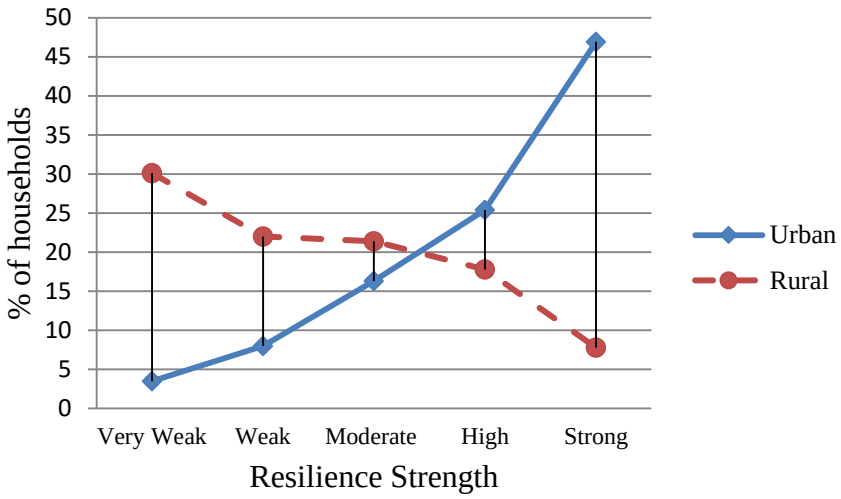

Figure 3.2: Levels of resilience by residential setting (South Sudan, 2009)

Classification of household resilience levels by rural or urban setting (Figure 3.2) showed clear disproportion between rural and urban households with regard to their resilience to food insecurity shocks and stresses. Whereas rural households in South Sudan were generally weak or moderate in their resilience levels, therefore facing more risk of vulnerability of food insecurity shocks, more urban households had generally 'stronger' resilience levels. This finding would have a bearing in planning for more rural and semi-rural development interventions.

The next step was to validate the household food insecurity resilience index (HRI) by comparing it with the consumption-based Household Wealth Index generated during initial survey analysis conducted by the National Statistics Bureau (NBS) of South Sudan (2010). Both the HRI and the HWI had a mean of 3.02 and 3.29, respectively (standard error of 0.009 for each). A test of association determined a strong relationship (Likelihood Ratio Test Chi-Square = 674.9 and DF=16 and *p*-value=0.000). The relationship between resilience and wealth can also be seen when the two variables were cross-tabulated as in Table 3.4.

Table 3.4: Household resilience levels by wealth index profiles (South Sudan, 2009)

Household Resilience Index Level	Wealth Index Quintiles (households)				
	Poorest	Poorer	Medium	Non-poor	Richer
Very Weak	23.1	24.4	19.9	17.4	15.2
Weak	25.6	20.0	21.7	16.7	16.0
Moderate	14.5	18.4	22.5	21.4	23.1
High	8.5	13.9	17.6	25.6	34.4
Strong	2.8	10.4	15.6	22.1	49.1

It is easy to note that 'poorer' households were associated with 'weaker' resilience to food insecurity while 'richer' households in terms of consumption expenditure were associated with 'stronger' resilience to food insecurity shocks. Whereas this result could be expected as 'natural' occurrence, it establishes the HRI as a good determinant of how households would fare if exposed to vulnerability.

Scale values of the HRI were cast in a linear regression model with the values of log-transformed per capita consumption (expenditure) in real terms. The rationale for this measure was to determine whether resilience could determine consumption. The distribution of the Log-transformed per capita consumption was closer to normal than per capita consumption.

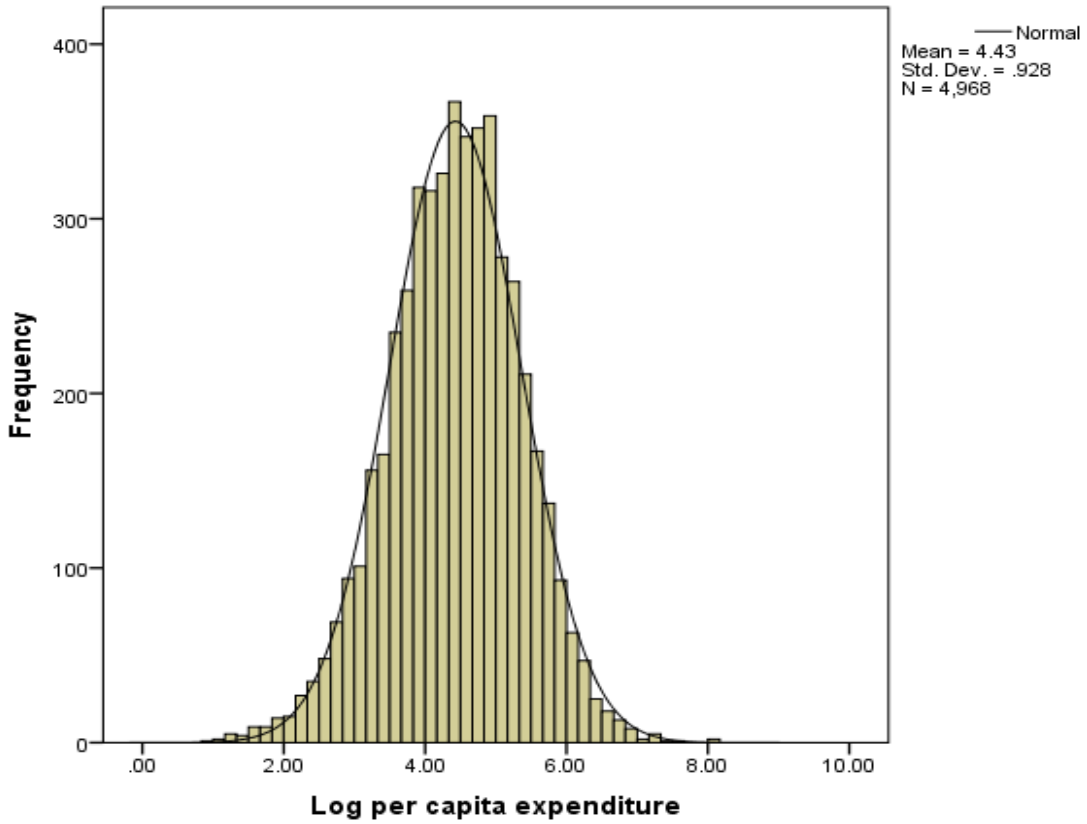

Figure 3.3: Distribution of the Log-transformed Per Capita Expenditure

A linear regression model assumes that there is a linear relationship between the dependent variable and each predictor. This relationship is described in the following formula.

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_p x_{ip} + \varepsilon_i \qquad (3.17)$$

where y_i is the value of the i^{th} case of the dependent scale variable, p is the number of predictors, b_j is the j^{th} coefficient, $j = 0, \cdots, p$, x_{ij} is the value of the i^{th} case of the j^{th} predictor and ε_i is the error in the observed value for the i^{th} case.

One way to validate the HRI is to establish whether it is a good predictor of per capita consumption. In order to do this, a suitable model for prediction of a scale dependent variable

by a scale predictor is a linear regression. Since there was only one predictor – the HRI – equation 4.17 translates to a simple linear equation,

$$y_{i=}b_0+b_1x_i+\varepsilon_i \tag{3.18}$$

with b_0 being the intercept or the model predicted value of the dependent variable when the value of the predictor is equal to *0*.

Table 3.5: HRI prediction of per capita consumption

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.373[a]	0.139	0.139	0.86131

a. Predictors: (constant), Resilience Index

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	594.650	1	594.650	801.569	0.000[b]
	Residual	3684.062	4966	0.742		
	Total	4278.712	4967			

a. Dependent Variable: log per capita expenditure; Predictors: (Constant), Resilience Index

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	Significance	
		B	Std. Error	Beta	t	Sig.
1	(constant)	4.302	0.013		331.013	0.000
Resilience Index		1.487	0.053	0.373	28.312	0.000

a. Dependent Variable: log per capita expenditure

The top part of the output in Table 3.5 shows that the regression model is poor in terms of the R-Square value. The regression analysis shows that the model explains only 14% of the variation in Log-transformed per capita consumption. The middle table gives the analysis of variance (ANOVA), which reports a significant F-statistic, indicating that using the linear regression model is better than inferences based on the estimated mean of the predictor variable. The bottom table shows the test of coefficients, which establishes the HRI as a very

strong predictor of the wealth measure generated using household expenditure on consumption of essential goods and services.

3.3.2. Discussion

The use of Principal Component Analysis in determining and profiling resilience levels was established to be sound both mathematically and statistically. Three important outcomes are generated from the application of the PCA technique: validity; reliability as a proxy measure and predictor of wealth; and determination and profiling of 'preparedness' of geographical entities against food unfavourable conditions causing food insecurity.

As regards its validity product, the Household Resilience Index, generated by reducing 33 variables into one component, which carries a substantive amount of the variance of those variables, was adequately representative of the weight of assets and housing characteristics. The large number of variables, especially those from related questions on availability of semi-durable assets, meant that variability spread out considerably, resulting in the first two extracted components accounting for a relatively low percentage of the variance (38.5 per cent). Moreover, most of the variables on semi-durable assets contained 'no' responses, as most of the populations did not have them. As South Sudan was barely two years old after two decades of civil war, the bulk of the population was in the process of settling down. Only a small segment of the population had assets such as motor vehicle, motor boat, television, air conditioners and refrigerators – these assets are typically associated with a settled population.

The relative frequencies in Table 4.1 clearly showed stark deprivation from assets associated with wealth of families such as motor vehicle, use of electricity, flush toilets, having air conditioners and using gas or electricity as source of cooking energy. At the time of the baseline survey it was clear that only a small proportion of households (0.9 to 3 per cent) had these types

of assets and what characterised their livelihood. Inclusion of these variables in the analysis is responsible for the low variability explained by the first extracted component, as shown in Table 3.2. Indications of the variance accounted for by the extracted components are very low for these variables. This occurrence is known as 'communalities' in PCA parlance. A solution to this problem could be to discard some of the variables known to have low frequencies from the analyses. However, this being a baseline study, it was seen worthwhile leaving the variables for future comparative analysis. A more pragmatic approach could be the use of non-classification techniques for generating the index.

Studies that used PCA to construct an asset-based index used a lower range of variables. Sahn and Stifel (2003) deployed 11 variables in their comparison of socioeconomic welfare in 12 developing countries. Even so, a considerable number of variables had low scoring coefficients or weights from the first extracted principal component, reflecting a large body of respondents reporting not having those assets or attributes. In their construction of an asset index for measuring asset accumulation in Ecuador, Moser and Felton (2007) conducted desegregated or structured analysis based on four livelihood capitals: physical (housing conditions and consumer durables); financial or productive capital (labour security, productive capital and transfer/rental income); human (mainly level of education attained); and social (house and community).

The second outcome of the analysis is that it has been established that the HRI is able to predict and associate with purchasing power or monetary wealth represented in per capita consumption. Table 3.4 clearly demonstrates the association of the HRI and the spending power in terms of consumption per capita. Stronger resilience manifests itself in the relatively wealthier households. Conversely, weaker resilience is a preserve of poorer households or those which spent less on food and other life necessities.

As regards its discriminating ability, the HRI is found to do well in profiling resilience to food insecurity adversaries by geographical or demographic characteristics. We profiled the ten states of South Sudan according to their resilience levels in 2009. We determined that the states of Warrap, Northern Bahr Al-Ghazal, Jonglei and Eastern Equatoria were characterised by weak to very weak resilience to food insecurity in as far as they had generally and commonly a lower asset base and poorer housing conditions than the rest of the states. Lakes State, Unity and Upper Nile State had what could be termed 'generalised moderate' resilience to food insecurity uncertainties. Both the 'worse' or 'moderate' states could be described in food security early warning jargons as 'alert' or 'watch' and, therefore, would need adequate preparatory measures for safeguarding against the eventuality of food insecurity causes, such as sudden market price increases, crop failure and low food commodity stock, supply road closure or others. On the other side of the scale, the states of Western Equatoria, Central Equatoria and Western Bahr Al-Ghazal, in that order, enjoyed relatively stronger resilience levels in 2009. As explained by their common advantage of favourable geographical and demographic conditions, these states would not be regarded as 'intervention areas' by food security mitigating and management organisations.

The rationale of opting for a measure of resilience is anchored on the fact that populations with low resilience become easily vulnerable to food insecurity calamities. Populations lacking in a combination of certain livelihood capitals, semi-durable and durable assets conceptually or naturally are low resilient to food insecurity eventualities and are, therefore, more vulnerable. Traditional measures of food security are largely based on vulnerability and more specifically on food consumption, micronutrient intake (e.g. calorie intake, dietary diversity and food consumption access) and anthropometrics in nutrition studies. Such studies are retrospective in that they examine data of events that have already occurred. A study for measuring resilience is, on the other hand, prospective, as it examines how the household or the area of study will

fare in the future when certain factors prevail. It was for this reason that this study got its motivation.

As food insecurity has proven to be an increasing problem of major concern in Africa, especially in settings where poverty is more rooted, there is need to explore pragmatic measures that prompt for appropriate and decisive action to prevent it from plunging a population into life threatening situations. In the case that certain population groups are affected by chronic or structured food insecurity, there is need for a measure that indicates how well prepared or how resilient the population will be. As explained in the preceding paragraph, the HRI explored provides a reasonable measure for ascertaining the level of resilience of households or the settings where they exist, such as states, counties or other localities. The HRI can be merited on six fairly good attributes.

The first attribute of the index is that it is a single summative measure. Being a composite indicator based on weights of several variables, it serves as a universal measure of livelihood attributes of a population group. Whereas previous studies using similar approaches explored the asset-based index as socio-economic welfare indicator, this study treats it as a measure of how resilient individual households, or geographically/demographically-grouped households, can potentially withstand the eventualities of food insecurity.

The second attribute of the index is that it has been established as an alternative to money metrics based on income or consumption expenditure data. Comparative analysis explored clearly demonstrates that the HRI can cater for the absence of the money metric-derived Wealth Index. Considerable amount of arguments have been presented that welfare measures based on monetary values of income or consumption variables present certain amount of biases attributable to recall, inaccuracies and others. Welfare and poverty researchers such as Gwatkin et al. (2000) argue that income measures lend themselves to practical difficulties such as

reluctance of informers to disclose how much income they have earned, lack of record keeping of money spent on consumption and many others. Liverpool and Winter-Nelson (2010, p.3) argue that consumption data can be affected by endogenous factors such as seasonality and weather conditions and therefore could not be a good measure of welfare.

The third advantage of the HRI is its ability to determine the probability of socio-economic conditions such as wealth, food consumption levels, among others. The index can inform vulnerability analysts to plan long-term interventions to limit adverse effects of conditions that threaten the livelihood and survival of a population. Furthermore, the index can determine or explain the state of socioeconomic deprivation and livelihood disparities among different population groups. Sahn and Stifel (2003) conclude that the index based on assets and livelihood endowments of households is a valid predictor of crucial manifestations of poverty such as child health and malnutrition. Filmer and Pritchett (2001) find the asset index to be as 'reliable' a predictor of school enrolment as a measure based on consumption.

As discussed in the foregoing section, the HRI has a fourth worth in profiling resilience by geographical or demographic setting. If the analysis were carried out immediately after the survey in 2009, it could act as an early warning on which states of South Sudan needed early preparedness against the eventuality of food insecurity shocks. The states which show very low or weak resilience can then be mapped out with red colour in order to invoke commitments and actions for early preparedness measures.

The fifth distinguishing characteristic of the HRI is that it is simple and easy to derive and interpret. Simplicity of the measure arises both in the raw data used in the analysis as well as the method used. Filmer and Scott (2008) observe that the data used to construct the index are simple to collect and frequently available. Moser and Felton (2007) describe the measure based

on PCA as ‘relatively easy to compute and understand’. Morris et al. (1999) put the PCA-based index in the category of simple measures that proxy for wealth indexes.

The sixth distinctive advantage of the index is durability. Since the index is based on semi-durable and durable assets, property owned (e.g. farmland, animal husbandry and other fixed assets) and households’ livelihood attributes (type of dwelling, sleeping rooms, source of lighting and cooking energy, etc.), it proves to be a medium to long term measure and, therefore, prompts for interventions with long-term goals and targets. It is important to note that the index is constructed using data that are always readily available in databanks of most national statistical agencies of developing countries. Datasets from national household budget surveys, demographic and health surveys and other socioeconomic status surveys are collected on a regular basis by statistical agencies. Survey questionnaires include items on different livelihood aspects as outlined in Table 3.1.

3.3.3. Conclusion

Drawing from the elaborated six advantages of the HRI, it is paramount to derive a conclusion that the HRI can withstand the test of being a reliable early warning measure for planning food security interventions, especially in chronically food insecure settings such as some livelihood zones in South Sudan.

Another important aspect to consider is that the index, based on a reasonably large sample size of 4968 households, which is representative of all ten states of South Sudan, has inherent statistical reliability, as its association with another livelihood measure – household wealth proxied by consumption data – has been determined to be strong.

It has also been established that, as recognised by other researchers who constructed a socioeconomic measure (index) based on assets, the Principal Component Analysis technique

provides a mathematically and statistically sound platform for constructing the HRI. One challenge that could be encountered in constructing the HRI is the range of asset and livelihood capital-based variables. In the case of South Sudan, the index could probably become more robust if fewer variables (say, less than 30) were used that typically reflected the reality at the time of the survey. We could not be certain whether assets that were largely owned by a small section of the population at the time, such as television, refrigerators, fans and motor vehicles, should be included. The reality at the time of data collection was that a substantial proportion of the population of South Sudan was still settling down and most of the people had no electricity or modernised assets. This could present a drawback findings.

Nevertheless, it is important to consider that the crux of the study is to present a procedure that might help in determining inequalities in a resilience of population groups to food insecurity uncertainties. This has been done as discussed above. The HRI has been established to a large degree as a prospective measure of potential risk and easy to determine using readily available data from periodical livelihood-related surveys. We, therefore, propose the adoption of the HRI for use in determining, mapping and profiling inequalities in resilience and potential vulnerability to food insecurity risk factors, as well as unveiling evidence for triggering early preparedness.

3.4. Multiple Correspondence Analysis

This Section examines the rigour and efficiency of Multiple Correspondence Analysis (MCA) in generating predictors of *FCS* – a proxy for food insecurity risk. The analysis builds on works by Booysen et al. (2005) who construct an asset-based index to estimate socioeconomic wellbeing (or poverty) in seven Sub-Saharan African countries, namely; Ghana, Kenya, Mali, Senegal, Tanzania, Zambia and Zimbabwe. They analyse and compare poverty trends in these countries. Booysen et al. (2005) clarify their use of MCA in the analysis of poverty trends

asserting that as PCA was primarily designed for analysis of continuous data, assuming normal distribution of indicator variables, MCA in contrast makes fewer assumptions about the underlying distribution of the variables. They observe that MCA is more suited to categorical (or discrete) variables.

Asselin (2002) construes that MCA is a formidable non-arbitrary tool for constructing a composite indicator based on categorical or qualitative indicators. The MCA-based composite indicator is generated by an optimisation process. Asselin (2002) defines the composite indicator of multiple qualitative poverty indicators as a set of categories for different population units.

The study used Multiple Correspondence Analysis (MCA) based on two grounds. First, in order to determine baseline predictors of household expenditure and, second, for constructing an index based on household characteristics and livelihood sources. MCA is used for classifying nominal variables and cases into a number of homogenous groups and determines the pattern of relationships amongst several categorical dependent variables. The method works to find optimal quantifications of data in the form of categories that are separated from each other as much as possible.

MCA is an extension of Correspondence Analysis (CA), which allows analysis of the pattern of relationships of several categorical dependent variables (Abdi & Valentine 2007). The method used in the analysis was considered suitable for the type of data examined and the purpose of the study, which aimed at constructing a statistically robust index for summarising socio-economic welfare of households and obtaining the resilience profiles of certain demographic characteristics.

The MCA methodology adopted by Booysen (2002) constructs the asset index based on the equation

$$P_i = R_{i1}W_1 + \cdots + R_{iJ}W_J, \tag{3.19}$$

where P_i is the i^{th} household's composite poverty indicator score, R_{ij} is the response of household i to category j and W_j is the MCA weight for the first dimension applied to category j.

By Asselin (2002) process for generating the MCA-based asset indicator follows two main steps: (1) Computing the profiles of each population unit relatively to the primary indicators; and (2) applying to these profiles the category weights, given by the normalized scores of these indicators on the first factorial axis produced from the indicators' MCA. This is expressed as:

$$C_i = \frac{\sum_{k=1}^{K}\sum_{j_k=1}^{J} w_{j_k}^{k} I_{i,j_k}^{k}}{K}, \tag{3.20}$$

where K is the number of categorical indicators, J_k is the number of categories for indicator k, $W_{j_k}^{k}$ is the weight (normalized first axis score) of category j_k and I_{i,j_k}^{k} is the binary variable 0/1, taking the value 1 when the unit i has the category j_k. Expression [3.20] is said to provide the solution that constitutes the inertia approach leading to the construction of the composite indicator.

Let n be the number of observations on p categorical variables. Assume that q_j different values for variable j. Next step is to define an indicator matric which is $\boldsymbol{n} \times \boldsymbol{q}_j$ matrix. Then $\boldsymbol{n} \times \boldsymbol{q}$ matrix $\mathbf{G}$ with $\boldsymbol{q}$ the sum of $\boldsymbol{q}_j$can be obtained by concatenating the $\mathbf{G}_j$'s (Greenacre 1984). By Benzécri (1992) MCA is defined as the application of weighted Principal Component to the indicator matrix $\mathbf{G}$. This matrix is further divided by its grand total $\boldsymbol{np}$ giving the

correspondence matrix $\mathbf{F} = \frac{1}{np}\mathbf{G}$ or $\mathbf{1_n^t F 1}_q = \mathbf{1}$, where $\mathbf{1}_i$, is $\boldsymbol{i} \times \mathbf{1}$ vector of ones. The vectors $\boldsymbol{r} = \mathbf{F1}_q$ and $\boldsymbol{c} = \mathbf{F}^t\mathbf{1}_n$ are respectively the row and column marginal vectors, which actually come from corresponding row and column masses. Suppose the diagonal matrices of the masses are defined as $\mathbf{D}_r = diag(\boldsymbol{r})$ and $\mathbf{D_c} = diag(\boldsymbol{c})$, for row and column, respectively. Note that the i^{th} element of $\boldsymbol{r}$ is $\boldsymbol{f_i} = \frac{\mathbf{1}}{\boldsymbol{n}}$ and the s^{th}element of $\boldsymbol{c}$ is $\boldsymbol{f_s} = \frac{\boldsymbol{n_s}}{\boldsymbol{np}}$ where n_s is the frequency of category $\boldsymbol{s}$ (Greenacre & Blasius 2006).

By its definition MCA is an application of PCA to the centred matrix $\mathbf{D}_r^{-1}(\mathbf{F} - \boldsymbol{rc^t})$ with distances between profiles given by the chi-squared statistic, which is defined by $\mathbf{D}_c^{-1}$. The $\boldsymbol{n}$ projected coordinate of the row profile on the principal axes are also the row principal coordinates. The $\boldsymbol{n} \times \boldsymbol{k}$ matrix $\mathbf{X}$ of the coordinates of row principal is defined by

$$\mathbf{X} = \mathbf{D}_r^{-1/2}\tilde{\mathbf{F}}\mathbf{V}_k \qquad (3.21)$$

where $\tilde{\mathbf{F}} = \mathbf{D}_r^{-\frac{1}{2}}(\mathbf{F} - rc^t)\mathbf{D}_c^{-1/2}$and $\mathbf{V}_k$ is the $\boldsymbol{q} \times \boldsymbol{k}$ matrix of Eigen vectors corresponding to the k largest eigen values $\boldsymbol{\lambda_1, \cdots, \lambda_k}$ of the matrix $\tilde{\mathbf{F}}^t\tilde{\mathbf{F}}$. The projected row profiles can be plotted in the different planes defined by these principal axes, which are also known as principal planes (Greenacre & Blasius 2006).

The categories for column profile can be described by the column profiles. The value can be calculated by dividing the columns of $\mathbf{F}$ by the marginal values of corresponding columns. When the rows are interchanged with columns all their associated entities can be used for the dual analysis of the column profiles. This is done by transposing the matrix $\mathbf{F}$ and then repeating all the steps. These are measures used in defining the principal axes of the centred profiles matrix $\mathbf{D}_c^{-\frac{1}{2}}(\mathbf{F} - \boldsymbol{rc^t})^{\boldsymbol{t}}$ are $\mathbf{D}^c$ and $\mathbf{D}_r^{-1}$.

The $\boldsymbol{n} \times \boldsymbol{k}$ matrix $\mathbf{Y}$ of columns principal coordinates is now defined by

$$\mathbf{Y} = \mathbf{D}_c^{-\frac{1}{2}} \tilde{\mathbf{F}}^{\boldsymbol{t}} \mathbf{U}_{\boldsymbol{k}}, \tag{3.22}$$

where $\mathbf{U}_k$ is the $\boldsymbol{n} \times \boldsymbol{k}$ matrix of eigen vectors corresponding to the $\boldsymbol{k}$ largest eigen values $\boldsymbol{\lambda_1, \cdots, \lambda_k}$ of the matrix $\tilde{\mathbf{F}}\tilde{\mathbf{F}}^{\boldsymbol{t}}$. These eigen values can be plotted for the purpose of visualisation and interpretation of the projected column profiles in the planes defined by principle axes, also called column principal planes (Johnson & Wichern 2007).

The absolute contribution of the variable $\boldsymbol{j}$ to the inertia of the column principal component $\boldsymbol{\alpha}$ in the $\boldsymbol{\alpha^{th}}$ column of $\mathbf{Y}$ is given by

$$\boldsymbol{c_{j\alpha}} = \sum_{\boldsymbol{s \in M_j}} \mathbf{S} \in \mathbf{M}_{\boldsymbol{j}} \boldsymbol{f_s y_{s\alpha}^2}$$

where $\mathbf{M}_j$ is the set of categories of variable $\boldsymbol{j}$. The relation between the absolute contribution $\mathbf{c}_{\boldsymbol{j\alpha}}$ and the correlation ratio between the variable $\boldsymbol{j}$ and the row standard component $\boldsymbol{\alpha}$ is given by

$$\boldsymbol{\eta_{j\alpha}^2} = \textstyle\sum_{\boldsymbol{s \in M_j}} \frac{\boldsymbol{n_s}}{\boldsymbol{n}} (\overline{\boldsymbol{x}}_{\boldsymbol{s\alpha}}^* - \mathbf{0})^{\mathbf{2}} = \boldsymbol{p} \times \boldsymbol{c_{j\alpha}} \tag{3.23}$$

Note that the PCA factor loadings are actually the correlations between the variables. Also note that the components or the correlation ratios are known as discrimination measures. These values can be interpreted in MCA as squared loadings.

Now, suppose $\mathbf{X}^* = \mathbf{X}^*\mathbf{T}$ and $\mathbf{Y} = \mathbf{YT}$, where $\mathbf{TT^t} = \mathbf{T^tT} = \mathbb{I}_{\mathbf{k}}$. Let $\mathbf{X}^*\mathbf{Y^t} = \mathbf{X}^*\mathbf{Y^t}$. Then, treat these relations show the lower rank approximation is not unique. However, the MCA solutions $\mathbf{X}^*$ and $\mathbf{Y}$ are not unique when conducted over orthogonal rotations. It is possible to explore this non-uniqueness so that the original solution can be improved by way of rotation. Rotation

of the column principal coordinates matrix **Y** to simple structure must be followed by the same rotation of the row standards coordinates matrix $\mathbf{X}^*$. The interpretation of the correlation ratios can be simplified for the matrices **Y** and $\mathbf{X}^*$ by rotation (Greenacre 2000).

For the method of rotation, the Varimax based function can be used. After rotation of **Y** and $\mathbf{X}^*$, relation (4.3) becomes

$$\tilde{\boldsymbol{\eta}}_{j\alpha}^2 = p \sum_{s \in M_j} f_{.s}\, \tilde{y}_{s\alpha}^2, \qquad (3.24)$$

where $\tilde{\boldsymbol{\eta}}_{j\alpha}^2$ is the correlation ration between the variable j and the α^{th} column of $\tilde{\mathbf{X}}^*$.

The graphical approach to represent the correspondence approach is the *biplot* representation. Therefore, *biplot* information is represented by $n \times p$ data matrix. As the name indicates, it refers to the two kinds of information contained in a data matrix. The information in the rows pertains to samples or sampling units and that in the columns pertains to variables. The scatter plot can represent the information on both the sampling units and the variables in a single diagram. This representation is useful to visualize the position of one sampling unit relative to another (Dray et al. 2003; Cao et al. 2001). In addition to this, it helps to visualize the relative importance of each of the two variables to the position of any variables. Matrix array can be constructed with several variables using scatter plots. The idea behind *biplots* is to add the information about the variables to the graph. Therefore, the construction of a *biplot* leads the sample principal components and the best two-dimensional approximation to the data matrix **X** approximates the j^{th} observation x_j in terms of the sample values of the first two principal components. Specifically,

$$\boldsymbol{x_j} = \bar{\mathbf{X}} + \hat{\boldsymbol{y}}_{j1}\hat{\boldsymbol{e}}_1 + \hat{\boldsymbol{y}}_{j2}\hat{\boldsymbol{e}}_2 \qquad (3.25)$$

where $\hat{e}_1$ and $\hat{e}_2$ are the first two eigenvectors of S and equivalent to $\mathbf{X_c'X_c} = (\mathbf{n-1})\mathbf{S}$ and $\mathbf{X_c}$ denotes the mean correlated data matrix with rows $\left(\mathbf{x_j} - \bar{\mathbf{X}}\right)'$.

The eigenvectors determine a plane and the coordinates of the $\boldsymbol{j^{th}}$ unit are the pair of values of the principal components $\left(\hat{\boldsymbol{y}}_{\boldsymbol{j1}}, \hat{\boldsymbol{y}}_{\boldsymbol{j2}}\right)$. The pair of eigenvectors has to be considered in order to include the information on the variables in the plot. These eigenvectors are coefficient vectors for the first two sample principal components. Thus, each row of the matrix positions ($\boldsymbol{E} = [\hat{\boldsymbol{e}}_{\boldsymbol{1}} - \hat{\boldsymbol{e}}_{\boldsymbol{2}}]$) a variable in the graph and the magnitudes of the coordinates of the variables show the weightings of the variables. The weightings represent each principal component of the variables. The plots of the variable with corresponding position are indicated by a vector. Singular value decomposition is the direct approach to obtain a *biplot*. Then, the singular decomposition expresses the $\boldsymbol{n \times p}$ mean correlated $\mathbf{X_c}$ as

$$\underset{(\boldsymbol{n}\times\boldsymbol{p})}{\mathbf{X}_{\boldsymbol{c}}} = \underset{(\boldsymbol{n}\times\boldsymbol{p})}{\mathbf{U}}\underset{(\boldsymbol{p}\times\boldsymbol{p})}{\boldsymbol{\Lambda}}\underset{(\boldsymbol{n}\times\boldsymbol{p})}{\mathbf{V}'},$$

where $\boldsymbol{\Lambda} = \boldsymbol{diag}(\boldsymbol{\lambda_1, \lambda_2, \cdots, \lambda_p})$ and $\mathbf{V} = \hat{\mathbf{E}} = \left[\hat{\boldsymbol{e}}_{\boldsymbol{1}}, \hat{\boldsymbol{e}}_{\boldsymbol{2}}, \cdots, \hat{\boldsymbol{e}}_{\boldsymbol{p}}\right]$ is an orthogonal matrix whose columns are the eigen vector of $\mathbf{X_c'X_c} = (\boldsymbol{n-1})\mathbf{S}$. The best rank two approximation to $\mathbf{X_c}$ is obtained replacing $\boldsymbol{\Lambda}$ by $\boldsymbol{\Lambda}^* = \boldsymbol{diag}(\boldsymbol{\lambda_1, \lambda_2, 0, \cdots, 0})$**.** This result is known as Eckart-Young theorem. The approximation is given as

$$\mathbf{X_c} = \mathbf{U\Lambda^*V'} = [\hat{\boldsymbol{y}}_{\boldsymbol{1}}, \hat{\boldsymbol{y}}_{\boldsymbol{2}}]\begin{bmatrix}\hat{\boldsymbol{e}}_{\boldsymbol{1}}' \\ \hat{\boldsymbol{e}}_{\boldsymbol{2}}'\end{bmatrix}, \qquad (3.26)$$

where $\hat{\boldsymbol{y}}_{\boldsymbol{1}}$ and $\hat{\boldsymbol{y}}_{\boldsymbol{2}}$ are the vector of values for the first and second principal components, respectively.

The *biplot* represents each row of the data matrix by the point located by the pair of values of the principal components. The $\boldsymbol{i^{th}}$ column of the data matrix is represented as an arrow from

the origin to the point with coordinates $(\hat{e}_{1i}, \hat{e}_{2i})$, the entries in the i^{th} column of the second matrix $[\hat{e}_{1i}, \hat{e}_{2i}]'$ approximations. Furthermore, the idea of a *biplot* extends to canonical correlation analysis, multidimensional scaling and even more complicated nonlinear techniques.

Other theoretical description, formulation of the methodology can be found in Asselin (2002, pp.10–13), Meulman (1996), Greenacre (1984), Benzécri (1992) and Tenenhaus and Young (1985). Appendix 1 of Asselin (2002) titled 'The Basic Principles of Correspondence Analysis and its Extensions to Multiple Correspondence Analysis', is particularly recommended.

The motivation for exploring the MCA procedure was to generate a single summary index based on homogeneous variables. It is to be underscored that Multiple Correspondence Analysis, also known as homogeneity analysis, is helpful in finding quantifications that are optimal in that variable categories are separated from each other as much as possible. The generated index is then used to profile population settings (states of South Sudan) by their levels of resilience in the period under study.

3.4.1. Results of the MCA Procedure

We apply the methods shown above to the data. Given that the MCA procedure does well with categorical data, the procedure performs graphical plots and produces statistics showing object scores, discrimination measures, correlations, among others. It is used to display the relationship between categories of variables.

A total of seven multiple nominal variables (Table 3.6), were selected from the original dataset to enable construction of the HRI. The rationale for selecting a limited number of variables is because the MCA procedure is based on quantification of nominal (categorical) data by assigning numerical values to the cases or objects and categories so that objects within the same

category are close together and objects in different categories are far apart. It is also to be noted that we wanted to select variables that are considered to modify or improve household livelihood and could indicate wellbeing of households. Therefore, considering the limitations of the procedure and the aim of the study, only those seven variables meet the criteria. It is to be noted that the seven variables in Table 3.6 are derived from Table 3.1, in particular the categories of *durable assets* (household characteristics) and *main source of income*. The key difference is that the 'Number of Levels' column of Table 3.6 are derived from collapsed variables that belong together. For example, what is now *Type of dwelling* in Table 3.6 was derived from *Permanent dwelling*, *Semi-permanent dwelling* and *Temporary dwelling* in Table 3.1.

Table 3.6: Variables and categories included

Variable	**Number of Levels***
Type of dwelling (x_1)	3
Number of bedrooms (x_2)	3
Main source of drinking water (x_3)	3
Main source of lighting energy (x_4)	4
Main source of cooking energy (x_5)	4
Toilet facility (x_6)	4
Main source of income (x_7)	9

* The different categories of a discrete or class variable,

Most of the variables in Table 3.6 were recoded for three main reasons: (a) in order to meet the requirements of MCA nominal variables and non-zero indicators, and (b) some of the codes or value labels assigned to variable categories had levels that were close in their meaning. For instance, the variable type of dwelling had value labels, "hut from mud", "hut from sticks", and "mud house with one floor". These types of houses were similar in South Sudan and this segregation could be confusing to the enumerator or interviewer, who could assign codes

haphazardly. Recording of responses could then differ from one interviewer to another. For the purposes of this study, these responses were lumped to one value called “semi-permanent”.

Most of the variables included were recoded in order to meet the requirements of the type of analysis pursued, that is, MCA nominal variables and non-zero indicators. Table 3.7 displays the seven variables included in the MCA for constructing a HRI by the marginal relative frequencies (percentages) of each variable category.

Table 3.7: Variables included in the MCA analysis by weights of each category

Variable	Categories (indicator)	Marginal Relative Frequency (%)
1. Gender of household head	Male (1)	72.7
	Female (2)	27.3
2. Cultivated crops last season	Yes (1)	81.0
	No (2)	19.0
3. Owned livestock last seasons	Yes (1)	96.9
	No (2)	3.1
4. Did fishing	Yes (1)	11.4
	No (2)	88.6
5. Main source of Income	Sale of farm crops (1)	29.1
	Sale of animal products (2)	22.0
	Employment/Labour (3)	21.6
	Petty/micro business (4)	21.0
	Other source (5)	6.4

It must be explained that what we call HRI is a special form of an asset index. The main difference is that this index is not limited to assets owned by a household, but is constructed

based on livelihood capitals, characteristics and amenities ascribed to a household using the MCA technique. The understanding is that resilience in food security sense is a function of all the aspects affecting the livelihood of people in a household. However, those livelihood aspects do not affect resilience uniformly. There are those that yield more effect than others. Conceptually, having a business enterprise as source of income can exert more resilience to food insecurity than, say, having a piece of farmland. Similarly, a household using flush toilet, piped water and brick and cemented house wall, should be economically better off than one living in a mud and straw hut, collecting drinking water from wells or boreholes and sharing a pit latrine with some neighbours. It is, therefore, the weight of each of the aspects describing the household, its head and occupants that need to be known. MCA helps determine the weights of these variables.

Prior to constructing the index, the problem of missing values had to be resolved. The SPSS (2013) Missing Value Analysis command was used to explore the amount and degree of missing values, which amount to less than 4 *per cent*. It was, therefore, decided to impute for the missing values to construct an index without any missing values. The SPSS (2013) Multiple Imputation Procedure was used to generate possible missing values and a complete dataset was created. This then led to a complete set of index values (scores) for each household in the sample.

Table 3.8 displays the seven extracted variables used for constructing the index by their respective first dimension weights. This index is constructed from multinomial variables, which are household attributes and sources of livelihood and wellbeing.

Table 3.8: Category weights of each variable from the first dimension of MCA

Variable	Category	Weight
1. Dwelling Type	Temporary	-0.018
	Semi-permanent	-0.118
	Permanent	2.275
2. Number of bedrooms	1 room	-0.356
	2 to 3 rooms	0.064
	4 rooms or more	1.052
3. Drinking water Source	Piped	1.8
	Borehole/pump	0.028
	Open or unprotected	-0.157
4. Main source of lighting	Electricity	2.887
	Paraffin or gas	0.693
	Other	-0.291
	No lighting	-0.312
5. Main cooking energy source	Firewood	-0.363
	Charcoal	1.59
	Gas or electricity	3.376
	Other source	0.045
6. Type of toilet facility	Pit latrine or bucket	1.126
	Flush	2.628
	Shared	0.736
	No toilet or bush	-0.372
7. Main source of income	Crop farming	-0.46
	Animal husbandry	-0.515
	Wages and salaries	1.24
	Business enterprise	0.729
	Property income	0.788
	Remittances	1.312
	Pension	1.674
	Aid	0.214
	Other source	0.068

Equation 3.1 (see Methods Section above) is deployed to calculate the HRI as a sum of the multiples of extracted variable category weights of Table 3.8 by each household responses to the respective category, or simply,

$$MCA_{P_i} = \sum_{i,j=1}^{J} R_{ij} W_j \qquad (3.27)$$

where MCA_{P_i} is the i^{th} household's composite poverty indicator score, R_{ij} is the response of household i to category j, and W_j is the MCA weight for the first dimension applied to category j. MCA was employed to calculate these weights using the *MULTIPLE CORRES* command in SPSS version 22 (2013). Fitting the above weights into formula 3.27 resulted in deriving a new variable, which we call HRI. The first dimension accounted for 36 *per cent* of the total variation (inertia).

The weights for each index component reported in Table 3.9 shows that components that reflect higher standards of living contribute positively to the resilience index, while components that reflect lower standards of living contribute negatively to the resilience index. For example, it is shown that permanent dwelling type, having two or more sleeping rooms, using a flush toilet, drinking from piped water, using gas or electricity for cooking or earning salaries and wages increase a HRI score, while living in temporary dwelling, having only one bedroom, having no toilet, using firewood for cooking, having access to lower quality sanitation and water supply or living mainly on crop farming or animal keeping, decreases HRI score.

After obtaining the resilience scores corresponding to each household in the sample, it was discretized by converting into quintiles and ranked as '*very weak*, 'weak', '*Moderate*', '*High*' and '*Strong*' for the first, second, third, fourth and fifth quintile, respectively. Geophysical profiles of household resilience to food insecurity shocks can be obtained by cross-tabulating them by states by other locational setting of the household. In the first resilience profiling instance, the mapping of resilience by state (Table 3.9) revealed some pattern.

Table 3.9: Resilience profiles by state (South Sudan, 2009)

	HRI Quintiles (*per cent* of households)				
State	Very weak	Weak	Moderate	High	Strong
Upper Nile	12.9	7.2	9.9	**36.8**	**33.2**
Jonglei	10.4	18.7	**34.3**	23.0	13.7

Unity	6.4	23.9	15.9	**29.9**	**23.9**
Warap	11.5	**44.3**	18.8	9.8	15.5
N. Bahr Al Ghazal	6.4	**34.7**	24.0	11.8	23.1
W. Bahr Al Ghazal	16.9	22.0	16.1	16.5	**28.4**
Lakes	5.7	**43.1**	20.9	18.1	12.2
Western Equatoria	**71.0**	9.8	6.5	7.1	5.6
Central Equatoria	28.0	14.4	10.2	**18.9**	**28.5**
Eastern Equatoria	12.5	27.3	18.6	26.9	14.6

It could be observed that Upper Nile State and, to some extent, Central Equatoria, Western Bahr Al Ghazal and Unity, had higher resilience than the rest of the ten states. On the contrary, Western Equatoria for some reason showed 'very weak' resilience to food insecurity shocks. Available information suggests that in 2009 large parts of Western Equatoria were affected by incursions of the Ugandan rebel Lord's Resistance Army (LRA), which occupied parts of the neighboring northern Democratic Republic of Congo and South Sudan.

The states of Warrap, Lakes and Northern Bahr Al Ghazal showed 'weaker' resilience. These states were commonly characterized by the main occupation of pastoralism and location in the Western Flood Plains agro ecological zone. Food security maps by National Bureau of Statistics' Food Security Technical Secretariat persistently paint these states as 'generalized food insecure', based on FAO's Integrated Humanitarian and Food Security Phase Classification (IPC) score card/tool (Food Security Analysis Unit (FSAU) 2006). The states of Jonglei and Eastern Equatoria seemed to have experienced 'generalized moderate' resilience to food insecurity levels during the time of data collection.

Profiling resilience by residential setting (Table 3.10) shows clear disparities between rural and urban areas. Whereas urban households had relatively 'higher' resilience, more households in rural areas had 'weaker' resilience.

Table 3.10: Resilience profiles by residential setting (South Sudan, 2009)

Residential Setting	HRI Quintiles (*per cent* of households)				
	Very weak	Weak	Moderate	High	Strong
Urban	19.0	8.8	8.0	**24.6**	**39.7**
Rural	**18.2**	**31.2**	21.5	17.9	11.2

In order to establish its robustness, it was seen necessary to compare the HRI with its counterpart the proxy wealth index, otherwise known as the wealth index scores. The HRI would assure some quality of rigour if it would be found to associate well with the proxy measure of wealth, which was the *per capita consumption in real terms*. Cross-tabulation and Chi-square test of association (not shown) of the categorical variable *wealth index quintiles* – measured through the proxy *per capita consumption* – and the categorical HRI yielded very highly significant Pearson's Chi-square and Likelihood Ratio values (16 degrees of freedom and 0.05 significance level; 2-sided test). A test of correlation involving the scale scores of the two variables yielded a significant correlation (Pearson's Rho=0.213; 2-tailed test). Therefore, this finding could imply that relative wealth of a household, proxied by per capita consumption expenditure, associated well with resilience of households, which was weighted based on certain household characteristics.

In order to establish the predicting ability of the index, its scale values were modelled by way of simple linear regression. It should be noted that the food insecurity resilience index has the intrinsic characteristic of assuring non-bias, as it was calculated based on uncorrelated extracted components. The index is also a summary measure of possible multiple variables.

The simple (bivariate) linear regression model is of the form

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where y_i is per capita consumption in real terms of household i, β_0 is the intercept or the model predicted value of the dependent variable when the value of the predictor is equal to 0, β_1 is coefficient of predictor variable (HRI), x_i is the value of the HRI corresponding to the i^{th} household and ε_i is the error term for the model; that is, the error in the observed value for the i^{th} household.

Linear regression was used to model how the index predicted per capita consumption. Fitting a model with the scale (continuous) values of the HRI scores did not result in significant F-statistic of the ANOVA test. However, the categorical or nominal values of the HRI quintiles yielded a highly significant F-statistic (p-value=0.004) as shown in Table 3.11 although with very low adjusted R-square value; meaning the model lacks good fit.

Table 3.11: Regression Analysis of relationship between HRI levels and Per capita consumption in real terms

a. Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.041[a]	0.002	0.001	148.21870

a. Predictors: (Constant), *Per cent*ile Group of HRI

b. Analysis of Variance (ANOVA)

Model	Sum of Squares	df	Mean Square	F	Sig.
Regression	180670.459	1	180670.459	8.224	0.004[b]
Residual	109096969.756	4966	21968.782		
Total	109277640.215	4967			

c. Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	113.979	4.935		23.094	0.000	104.304	123.655
	*Per cent*ile Group of HRI	4.284	1.494	0.041	2.868	0.004	1.355	7.213

a. Dependent Variable: Total per capita income in real terms

The test of the estimate of the regression coefficient of HRI Quintiles ($\hat{\beta}_1$) was also significant at the 0.05 level of significance (*t*-statistics *p*-value of 0.004), implying that the HRI was an important predictor of the proxy wealth index. Inspection of the predicted model revealed a few (10) outliners. Therefore, the model might improve if cases (households) with these outliers were excluded from the analysis and the adjusted R-square increased to indicate model good fit.

3.4.2. Discussion

Following in the footsteps of researchers such as Booysen et al. (2005) and Sahn and Stifel (2000), an index was constructed based on household attributes and sources of livelihood. Considering that a household can acquire resilience to food insecurity shocks, strains and stressors depending on the importance (or weight) of durable or semi-durable assets, use of certain wellbeing amenities shared by household members and livelihood characteristics, the

index characteristically manifested the attributes of a composite measure of resilience. The MCA procedure generated weights for each category of the selected variables. The first extracted dimension provided the weights or component loadings. The method did well in distinguishing components that positively contributed to the index from those that contributed negatively. The method was able to ascertain that a household living in permanent housing structures, with two or more rooms for sleeping, used safe source of drinking water, used more sanitary toilet facilities, used gas or electricity for cooking and lived on sources other than farmland and animal husbandry, had positive weights – quite what was expected.

The MCA technique was also able to isolate states that had better resilience in terms of livelihood capitals and shared facilities from those that had low resilience. It presented a baseline outlook in terms of status of resilience to food insecurity by states in South Sudan in 2009. The method could give food security analysts reason to investigate factors that characterised low resilience in some states and what made some of them to have favourable resilience profiles.

The method could thus help in developing appropriate food security preparedness interventions to allocate resources or in taking necessary actions for early warning or prioritising those states with low resilience in terms of resource allocation. Using the technique to attain resilience profiles by residential setting, led to the finding that household in rural communities had lower resilience than those in urban areas (Table 3.5). Consistent with common knowledge, rural households in South Sudan generally did not have the amenities characterising the lives of urban residents such as electricity, permanent dwelling structures and using sanitary facilities. The post-conflict nature of the country made this case typical.

The constructed resilience index (HRI) was cast in a regression model to determine its predicting capability and association with a variable it was expected to associate well with: the

index based on per capita consumption of households. The model did well in determining the significant F-test result as well as the test (t-test) of the estimated coefficient of the predictor variable HRI Quintiles. This indicates that HRI, on its own, contributes to prediction of per capita consumption. The regression model, therefore, validated the resilience index as a good measure of socioeconomic wellbeing.

Some inherent shortcomings were associated with the dataset from which the HRI was determined. Data was collected for purpose other than the one investigated in this analysis. The main aim of the survey was to determine baseline levels of poverty based on consumption variables. As is typical of nationwide household surveys, the National Baseline Survey questionnaire was very long and extensive, covering many living aspects, household and individual characteristics.

3.4.3. Conclusion

Analysis carried out revealed three key findings. First, the MCA technique ascertained the feasibility of constructing an index based on a set of commonly used household facilities and livelihood sources, as well as distinguishing those characteristics of the study elements that contributed positively to the index from those that contributed negatively.

The second key finding was that the MCA methodology was able to help in profiling resilience levels by the different residential settings of the sampled households. Therefore, the technique can be used in providing evidence for early warning, early preparedness and allocation of related resources. The technique could help justify why certain states or regions would qualify for relevant food insecurity risk-aversion resources than others. Overall, the index could guide both planning and program prioritisation.

The third key finding of the study was that with a good set of raw data the index could be a good measure of the status of resilience in the face of the often unavoidable uncertainties, especially as existing measures of food insecurity heavily dwell on assessing vulnerability based on retrospective events and incidences of past food consumption. In other words, existing indicators inform that food insecurity level has afflicted a certain individual or household, rather than how it is likely to occur based on current situation.

Nevertheless, the resilience index generated using MCA showed poor results due to the large number of variables. This limitation was manifested in that a large number of variables reduce the amount of variance represented by the extracted component. Therefore, there is need to re-apply the technique to data from a controlled study, or preferably to panel data or data from a longitudinal study.

3.5. Classification and Regression Tree Analysis

This Section explores the application of Classification and Regression Tree (CART) Analysis in identifying the most influential variables in a set of possible predictors of an outcome (or response) variable. The Classification and Regression Tree (CART) approach is a non-parametric technique used for selecting variables that are important for determining predicted values of an outcome (dependent) variable. The Tree procedure builds a tree one level at a time (i.e. recursively) and examines all predictors at each level for all possible splits. The procedure then chooses a predictor and splits at each node and level to minimize misclassification. The tree is then grown until no more splits can be made and is then pruned to be parsimonious while minimizing misclassification.

First developed by Breiman et al. (1984), the technique is capable of handling nominal, ordinal and scale variables. It classifies variables depending on whether the outcome variable is

continuous (scale) or discrete (categorical). For the former, CART produces regression trees. Otherwise, if the outcome variable is categorical, CART analysis takes the form of classification trees. According to Loh (2011) classification trees are designed for dependent variables with a finite number of unordered values and with prediction error measured in terms of misclassification cost. The outcome variable is the binary indicator of socioeconomic status (SES).

The classification tree is constructed based on three components: (1) a set of questions for deciding a split; (2) splitting rules and goodness-of-split criteria (for judging how good a split is); and (3) rules for assigning a class to each terminal node. For thorough and step-by-step description, derivation and rationalizing of the CART method see Yohannes and Webb (1999) and Breiman et al. (1984).

Weiser et al. (2009) observe that CART is particularly useful in cases where interactions are expected between multiple factors related to an outcome variable. The CART procedure is robust in segregating or identifying groups of variables. We use the IBM SPSS (2013) Decision Tree (CRT) procedure (CHAID option) in our analysis. The method is aimed at determining whether a household was 'poor' or 'non-poor' in its resilience strength.

As the outcome variable is categorical, we deploy the Classification Trees technique for identifying the set of predictors. The procedure uses probability priors in the class of Bayesian statistics. Prior probabilities play a central role in building the classification trees (Weiser et al. 2009). The procedure allows for three types of priors: *priors data, priors equal and priors mixed*. Illustrated using a notation, let

N = number of cases in the sample,

N_j = number of class j cases in the sample, and

π_j = prior probabilities of class *j* cases.

Priors data assumes that distribution of the classes of the dependent variable in the population is the same as the proportion of the classes in the sample. It is estimated as $\pi_j = N_j/N$. *Priors equal* assumes that each class of the dependent variable is equally likely to occur in the population. For example, if the dependent variable in the sample has two classes, then *prob(class 1) = prob(class 2) =* $^1/_2$. *Priors mixed* is an average of *priors equal* and *priors data* for any class at a node (Yohannes & Webb 1999).

In order to validate the index constructed using CHAID, the SPSS GENLIN procedure was employed. This procedure is based on the Generalised Linear Logistic Regression model (Nelder & Wedderburn 1972; McCullagh & J.A. Nelder 1989).

Like in the PCA and MCA, the main motivation for using the CART procedure is to explore its usefulness and versatility in generating a single composite index for determining risk of food insecurity. Research shows that the procedure had not been applied for identifying and selecting predictors of food insecurity or generating an index. It is for this reason that CART was used in selecting possible asset- and livelihood-based predictors of the proxy for food insecurity risks: socioeconomic status. Where the method proved to do fairly well in serving this purpose, it could lead to drawing a conclusion regarding its usefulness as a tool for early warning, disaster preparedness and resilience enhancement interventions (e.g. social protection of populations with food security-related vulnerabilities).

3.5.1 Results of the CART Procedure

The CART method and steps described above were applied to the set of data in Table 3.1 in which a tree-based classification model was fitted, which classified cases into groups or predicts values of a dependent (target) variable based on values of independent (predictor)

variables. The procedures then enabled identification of homogeneous groups with high or low food insecurity risk.

The dependent variable is the binary socioeconomic status (SES) indicator; calculated based on the value (or prices) a household spent on acquiring certain goods and services. It is used as a proxy for the risk to a food insecurity shock. A household scoring low on the socioeconomic metric is categorized as 'poor', which means it risks the potential of being food insecure. The very fact that it was not able to spend or had spent less on the semi-durable assets, taken together, could reflect its poor purchasing power; thus low access to food and for that matter, quality and nutritious food.

Nineteen variables representing household assets and livelihood amenities were entered into the model. The procedure automatically excluded any variable that did not make significant contribution to the final model. In the procedure, the SES category 'poor' is specified as the target category for the purpose of comparison in the analysis.

The CHAID model (see Figure 3.4) selected only six variables from the 19 independent variables as predictors of SES due to their statistical significance main cooking energy source, type of toilet facility, ownership of radio, ownership of a phone, main source of income and main source of energy for lighting. The rest of the 14 variables did not contribute significantly to the model, and were excluded from the model.

It is, therefore, with reasonable amount of classification, the six variables stand out as best predictors among the factors of household resilience included in the model. Of particular interest, using firewood or grass, having no toilet (i.e. using bush) and not owning any radio combine as the best predictor of 'poor' SES. In other words, households lacking in these assets are likely to have weak resilience to a food insecurity shock.

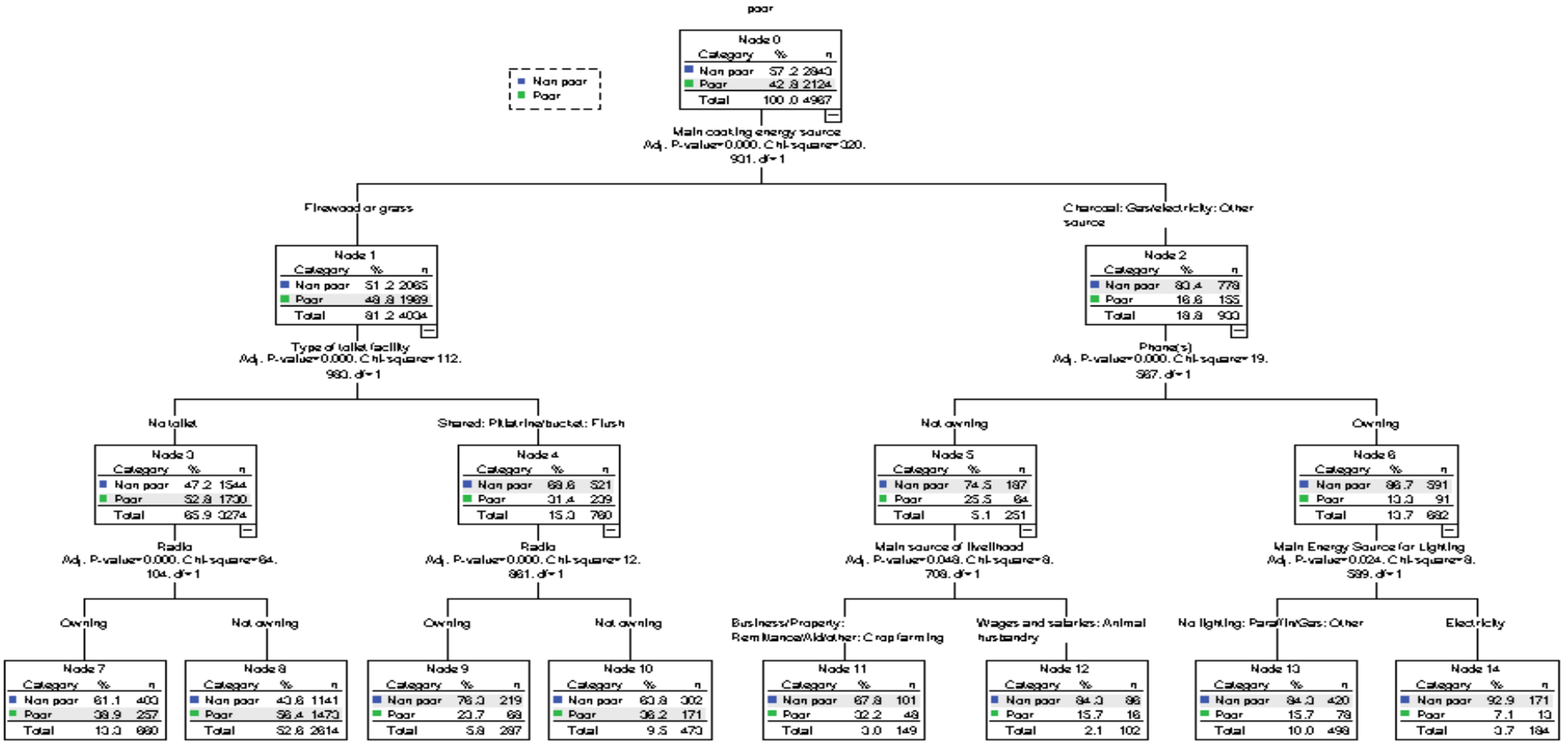

Figure 3.4: A Tree Diagram of the CHAID Model

In evaluating the model, the gains for nodes table (Table 3.12) is used which provides a summary of information about the terminal nodes in the model. It shows that one of the eight terminal nodes (i.e., the nodes at which the tree stops growing and represents the best classification predictions for the model) yields an index value greater than 100 *per cent*, meaning that there are more cases in the target category than the overall percentage in the target category. Node 8 is non-ownership radio, which is shown to be in the '*poor*' category.

Table 3.12: Gains for nodes

Node	Node		Gain		Response (%)	Index (%)
	N	*Per cent*	N	*Per cent*		
8	2614	52.6	1473	69.4	56.4	131.8
7	660	13.3	257	12.1	38.9	91.1
10	473	9.5	171	8.1	36.2	84.5
11	149	3.0	48	2.3	25.5	75.3
9	287	5.8	68	3.2	23.7	55.4
12	102	2.1	16	0.8	15.7	36.7
13	498	10.0	78	3.7	15.7	36.6
14	184	3.7	13	0.6	7.1	16.5

Both the gains chart and index chart (Figure 3.5) show that the model was fairly a good one and that the CHAID model provides information. For a good model, the gains chart will rise steeply toward 100 *per cent* and then level off. A model that provides no information will follow the diagonal reference line. Cumulative gains charts always start at 0 *per cent* and end at 100 *per cent* as you go from one end to the other. Likewise, for a good model, the index value is supposed to start far above 100 *per cent* and remain on a high plateau and then trail off sharply toward 100 *per cent*. For a model that provides no information, the line would be hovering around 100 *per cent* for the entire chart.

The index chart also indicates that the model was a good one. Cumulative index charts tend to start above 100 *per cent* and gradually descend until they reach 100 *per cent.*

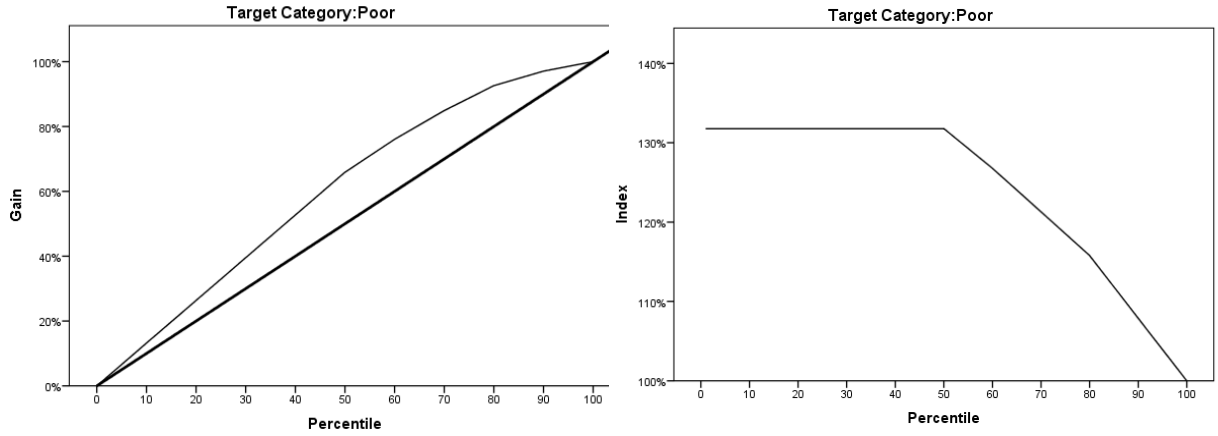

Figure 3.5: Gains and Index Charts (modified output of the CHAID procedure)

Proceeding with more evaluation of the model, we examined the risk of misclassifying households as ‘poor’ and ‘non-poor’. The risk and classification tables (Table 3.13) provide a quick evaluation of how well the model performs. It is shown that the risk estimate is 0.461, which indicates that the category ‘*poor*’ SES was wrongly predicted by the model for 46.1 *per cent* of the households, thus the "risk" of misclassifying a household is approximately 46.1 *per cent,* which is high. Consistent with this result, the classification table (part (a) of Table 3.13)) shows that the model correctly classifies approximately 89.5 *per cent* of the households as ‘*poor*’. The classification table does not invoke any concern, as it shows 89.5 *per cent* of the households in the ‘*poor*’ category to have been correctly classified. This improvement is after adjusting for the cost to outcomes. Thus, it leaves 10.5 *per cent* of the households misclassified. Obviously, this measure reduces the overall percentage of misclassification of households.

Table 3.13: Risk estimate and classification after assigning costs to outcomes

a. Risk

Estimate	Std. Error
0.461	0.007

b. Classification

	Predicted		
Observed	Non poor	Poor	*Per cent* Correct
Non poor	997	1846	35.1
Poor	223	1901	89.5
Overall Percentage	24.6	75.4	58.3

The CART model generated four new variables in the active dataset, namely; a) the terminal node number for each household; b) the predicted value of the dependent variable for each household; c) the predicted probability that the household belongs in the '*poor*' SES category; and d) the probability that the household belongs in the 'non-poor' SES category. Since the interest is to generate predicted values corresponding to the selected terminal nodes, they are taken as the desired HRI, which can be used in targeting households with resilience enhancement interventions. In this case, households with higher predicted probability values, say, above 50 *per cent*, have high probability of weaker resilience and thus need to be strengthened.

It is reasonable to see how the CHAID model compares with an alternative model. We fit a Binary Logistic Regression Model to the same data, which is most appropriate for modelling the event probability for a categorical response variable with two outcomes. It is to be recalled that the aim of the study is to select predictors of the probability of households encountering 'poor' food consumption – a proxy for socioeconomic status. In other words, the model is used to assess the risk of households that were likely to be affected by food insecurity shocks.

Therefore, it is necessary to identify the characteristics of households (assets and livelihood characteristics) that are indicative of households likely to face food insecurity shocks and stresses, if such eventuality would occur.

A generalized logistic regression model was fitted to the data with binary response (or outcome) variable and the 19 explanatory variables, it produced the output contained in Table 3.14 below. As the data came from a binomial distribution, the *Logit* link function was used for fitting the data. Included in Table 3.14 are the variables with significant effects to the probability of socioeconomic status. The rest of the variables were ignored under the Wald Type III test of significance.

Table 3.14: Test of model effects*

Source	Type III		
	Wald Chi-Square	**df**	**Sig.**
(Intercept)	38.629	1	0.000
Type of toilet (x_4)	50.807	3	0.000
Source of cooking energy (x_3)	39.753	3	0.000
Ownership of radio (x_8)	35.778	1	0.000
Ownership of animal transportation (x_7)	28.804	1	0.000
Source of income (x_5)	21.649	4	0.000
Number of bedrooms (x_1)	13.469	2	0.001
Ownership of bicycle (x_6)	7.509	1	0.006
Phone (x_9)	5.938	1	0.015
Source of lighting (x_2)	5.979	2	0.021

* Only variables with significant effects are shown

3.5.2 Discussion

The Classification and Regression Tree (CART) method was primarily explored for the following distinctive characteristics. First, the method was established to be robust enough in dealing with high dimensional data, as it is fundamentally non-parametric, than the parametric

regression techniques (Westreich et al. 2010). Second, it selected much fewer, but the most probable predictors than regression-based techniques. Results of the CART model showed that only three independent variables out of five selected variables combined as 'best' predictors of socioeconomic status. Third, results of CART were easier to interpret, explain and implement as compared to those of other regression techniques (Gordon 2013; Loh 2011).

The technique succeeded to identify the best overall predictor of the outcome. In the case study, using firewood for cooking associated with the likelihood of socioeconomic status, followed by not having a pit latrine and owning a phone. However, the proportions of populations in the 'poor' categories of these three variables are not significantly different compared to the 'non-poor' categories. This could be due to high misclassification errors.

Model evaluation shows that the CHAID model does fairly well in classifying the terminal nodes and in terms of information. This finding was strengthened by the Gains-for-nodes table and gain and index charts. However, the method was found to have some shortcomings related to misclassification errors and assigning costs to outcome as noted by Loh (2011) and Yohannes and Webb (1999). In fitting the CHAID model, missing data were treated as a non-observable variable for the purpose of classification.

It was found out that the GENLIN model identified the same three other possible predictor variables (use of energy for cooking, type of toilet facility used and ownership of radio) to be highly significantly associated with the outcome 'poor' SES, as did the CHAID procedure. This gives enough grounds that the CHAID model proved valid.

The method successfully used generated predictors based on the selected nodes of the CART method. This index might help in identifying households according to the probability of how they might withstand future risks to food insecurity.

3.5.3 Discussion

In summary, the study has revealed a number of interesting findings. Foremost, in its quest for appropriate, valid and simple methods for identifying predictors of food insecurity uncertainties, the Classification and Regression Trees (CART) analytic method seems to fare well. The technique distinguished itself in identifying an overall best predictor out of several possible predictors. The procedure also proved valid, as it compared well with the Generalized Binary Logistic Regression Model.

Secondly, review of substantial literature on security, revealed very limited application of the technique to data of similar type and characteristics. Thus, it could be worthwhile exploring how the method could be applied to establish the most statistically efficient predictors of potential food security shocks and strains in a population with weakened resilience to livelihood risks.

Thirdly, the CART method proves capable of generating predictor values per households based on the selected nodes of the regression tree model and on the proportion of households in each category of the dependent variable for the terminal node that contains each household. This generated predictor variable is taken as the desired 'Household Resilience Index' (HRI), which can be used for grouping households according to geographical areas and/or for targeting households identified to have a high resilience probability.

Fourth, data analysis has established the CART Analysis may provide a useful and simple tool in food insecurity early warning analysis and programming. It thus provides more grounds for recommending the method for use in analysis of data from similar source and purpose.

The method becomes of high relevance, especially as an emerging need for shifting attention to resilience analysis, rather than focus on vulnerability analysis, is gaining momentum. The

technique could be used in longitudinal and predictive analytic studies that help in providing evidence for developing resilience strengthening and poverty and eradication interventions.

Finally, a major drawback of the model was misclassification errors. This could be due to the data coming from large surveys. Moreover, the study was based on secondary data, which were collected for a different purpose. Thirdly, independent variables included in the model (assets and livelihood amenities), originally with multiple responses could suffer from within-case correlations – thus not truly independent. For instance, a household that could afford a bicycle or motorcycle could also afford to buy a radio, a mobile phone, etc. Therefore, in consideration of these limitations, it could be important to validate the technique further with data from more targeted, simpler and well executed surveys.

CHAPTER 4

Logistic Regression for Analysis of Binary Response Data

4.1. Introduction

In Chapter 3 three multivariate procedures were used for generating latent variables which were then used for classifying food insecurity resilience in accordance with geographical settings of households. This chapter introduces logistic based approaches for determining the level of resilience to food insecurity risk. It specifically aims to explore whether households mainly subsisting on agriculture and other forms of rural livelihoods determine coping with food insecurity risk, especially in a crisis. Two logistic modelling approaches were examined in this chapter: binary logistic comparing results of models with *logit, probit* and *complementary log-log* link functions and the survey logistic model. The chapter is structured into five sections. Section 4.2 highlights the importance of agriculture for largely rural populations and how the sector is regarded as playing an important role in shielding these populations against debilitating effects of shocks, with particular reference to protracted crisis. It also highlights arguments on how agricultural risks could in turn render households heavily dependent on the sector to be vulnerable. The section further introduces the concept of coping and its relevance. It also rationalises the choice of the methods used in the analysis. Sections 4.3 presents and discusses the sample selection and data, methods, results, discussions and summary. Section 4.4 presents and discusses the results of the binary logistic model. Section 4.5 presents, discusses and makes summary on the results of the survey logistic model.

4.2. Rural Livelihoods and Coping with Food Insecurity in Protracted Crisis

Agriculture is the main source of livelihoods for over 70 *per cent* of rural populations in Africa (EAO 2015). The sector employs over 62 *per cent* of the populations (Staatz & Dembélé 2007). It is essential for boosting food insecurity and nutrition, creating jobs and eradicating poverty (Committee on World Food Security 2015).

The Food and Agricultural Organization of the United Nations (FAO) (2002, p.3) determines that "in most countries with a high incidence of food insecurity, agriculture is the mainstay of the economy". This then places high importance for investing in agriculture as a key to alleviating poverty and cut down extreme food insecurity and undernutrition. The Committee on World Food Security (2015) underlines that "investing in agriculture and food systems is one of the most effective ways to reducing hunger and poverty" considering that the sector has an intrinsic advantage of multiplier effects to many economic sectors.

It is, therefore, imperative that agriculture can strengthen the resilience of a population and enable households to cope with food insecurity shocks such as conflicts and displacement. Drawing a lesson from Tajikistan, the 2013 State of Food Insecurity in the World underscores that structural changes in agriculture are necessary to creating resilience to shocks (IFAD et al. 2013). However, dependence on agriculture may also render households vulnerable to climate shocks. For instance, Pasteur (2011) cites a case of Zimbabwe in which rural poor households that were dependent on rain-fed agriculture became vulnerable when rains failed. She argues that financial capital or source of income could strengthen coping with such uncertainty. Nevertheless, it is imperative that if income from agriculture is saved, farming households can still withstand livelihoods shocks.

Populations characterized by rural livelihoods often depend on crop agriculture, livestock, fishing, marine resources and forest products for sustaining their food security. However, it is yet to be established whether factors characterizing the traditional livelihoods sector determine coping with transitory food insecurity risk when crisis strike or not. Robust statistical modelling techniques were not previously used to identify most important factors that improve coping in the eventuality of food insecurity shocks. Furthermore, as the search for a single composite measure of household resilience to food insecurity shocks is still on, there is need to explore the use of statistical methodology to construct and validate an index for assessing the risk of food insecurity based on factors determined to be important in the analysis.

Rural and poor households are often rendered highly vulnerable when disasters strike such as in the case of South Sudan from the end of 2013, when civil conflict flared. Pasteur (2011) observes that livelihoods of the poor, including smallholder farmers, artisans and fishermen, are hit the hardest, plunging them to more poverty and hunger-related fatalities. These vulnerable populations are then forced to extreme forms of coping strategies; a manifestation of high risk of food insecurity.

Coping with a food emergency is conceptually a function of household resilience to food insecurity shocks and stressors. Household resilience is based on their livelihood sources and characteristics. A predictor of the incidence of coping with food emergencies based on these livelihood sources and household characteristics presents a good basis for generating an index for monitoring emergencies.

The purpose of the study was to identify the set of variables that determine the risk of food insecurity based on factors that influence coping with food shortage in a crisis situation. The experience of coping with food emergency was chosen considering the fact that population in protracted crises or emergency settings such as that of South Sudan during the period of the

study, often times have to cope with the shock of severe food shortage. The study specifically aimed at exploring the extent to which typical factors such as household characteristics and sources of livelihood or income determine whether or not households in emergency had to adopt a coping strategy.

4.3. Sample and data

Data used were taken from the Food Security Monitoring Survey (FSMS), which was conducted in South Sudan in August 2014 during a raging conflict in which hundreds of thousands of households were displaced. Detailed description of the sample and collected data can be found in Chapter 2 Sub-section 2.2.2. or in World Food Programme (2014).

The target (or response) variable was the incidence of *adopting* or *not adopting a coping strategy* during food insecurity crisis or emergencies. The proportion of '*adopting a coping strategy*' to '*not adopting a coping strategy*' was 49 *per cent* to 51 *per cent*, respectively. These categories arose from a set of questions asked during the survey as to how a household managed to cope with situations of food shortage. In the survey (World Food Programme 2014), a question was floated to respondents as to whether or not in the past 30 days there were instances when they had to resort to any of the listed coping strategies due to shortage or lack of food or money to buy it and how often the household encountered the particular situation. Responses were then scored in accordance with a scale or framework as suggested in Maxwell (1995). Responses were weighted according to severity of the coping mechanism. Coping strategy such as consumption of less preferred food was ranked as 'less severe' and carried minimum weight or score. Extremely severe coping strategy such as skipping entire days without eating carried maximum score.

According to the scale for calculating the Coping Strategy Index, scoring categories for frequency of the coping incidence (up to seven days) range from 0 (never coped) and 7 (coped all the time). Meanwhile, weights of coping strategies ranged from 2 (less severe coping) to 8 (extreme coping). The index was generated by summing the products of the frequency of coping with and the weight of the coping strategy adopted. Denoted algebraically this is expressed as

$$CSI = \sum_{i=1}^{k} w_i z_i,$$

where z_i is the relative frequency score corresponding to coping strategy i, $i = 1, \dots, k$ and k is the number of coping strategies.

Households that had their coping strategy index amounting to zero must have obviously answered that they 'never' experienced any situation when they had to adopt a coping strategy. Households adopting any of the coping strategies were almost half of the sample. Maxwell and Caldwell (2008, p.10) distinguish between 'consumption coping strategy and 'livelihood coping strategy'.

Seven predictor (or explanatory) variables X_{ij}, where, $i = 1, \dots, 7$, $j = 1, \dots K$ and *K*= number of levels, were included in the Generalized Logistic Regression Model for the binary response variable '*coping with food security emergency*'. These were: a) three demographic variables (*age of household head*, *gender of household head* and *household size*); b) history of livelihoods activity prior to the crisis (*crop cultivation*, *livestock keeping* and *fishing*); and c) *main source of income* (during protracted food insecurity crisis). Mathematically, the response/outcome variable is given as

$$Y_j = \begin{cases} 0 \text{ if household did not adopt any coping strategy} \\ 1 \text{ if household adopted some coping strategy} \end{cases}$$

The rationale for selection of only a few variables is two-fold. First, each of these variables is seen to affect how households coped with food insecurity in during the crisis that hit South Sudan. Secondly, the sample survey where the data came from included a few variables, as it was a repeated monitoring survey (World Food Programme 2014). The survey concentrated on collecting data on variables for computing key food security and nutrition indicators, namely; coping strategies index, food consumption score, dietary diversity index and per cent share of total expenditure spent on food.

Table 4.1 shows the percentages of predictor (explanatory) variables included as possible predictors of coping with food security.

Table 4.1: Predictor variables included in the analysis

Variable Description (X_{ij})	**Category (*J*)**	***n***	***Per cent***
Gender of the household head (x_{1j})	Male	2683	72.7
	Female	1009	27.3
Age of the household head (x_{2j})	1=(< 17 yrs)	42	1.1
	2=(18-60 yrs)	3549	96.1
	3=(>60 yrs)	101	2.7
Size of household (x_3)	Scale	-	-
Cultivated crops past 3 months (x_{4j})	Yes	2990	81.0
	No	702	19.0
Owned livestock past 3 months (x_{5j})	Yes	3576	96.9
	No	116	3.1
Engaged in fishing past 3 months (x_{6j})	Yes	421	11.8
	No	3271	88.2
Main source of income (x_{7j})	Sale of agricultural crops	1074	29.1
	Sale of livestock products	811	22.0
	Employment/labour	798	21.6
	Petty trading	774	21.0
	Other	235	6.4

The six variables were selected based on a sound rationale. *Gender of household head* is important on the basis that a household headed by a female or male might fare differently in

situations of food crises. This argument could hold true especially when employment is the major source of income.

From theory *age of household head* would have considerable correlation with how a household coped, as the older the age of a person, he or she is supposed to have better experience in coping strategies. The *size of household* (or number of individuals in a household) conceptually has a bearing on some form of coping such as meal rationing. In situations of emergencies, households with more members, especially adults, could tend to employ more of its members to fetch food. Conceptually, households with more members could translate to more than one source of income or bigger food rations from food aid. However, care needs to be taken in interpretations based on *household size* as an indicator, as some respondents might tend to overstate the number of their members in order to receive bigger food rations.

The imperative of having *cultivated crops*, especially food crops in the past farming season could be a favourable factor to households during food crisis, although some might have lost or left behind everything when forced to flee their original habitats. *Owning livestock* prior to or during crisis might buffer households against facing food shortage. However, this might not have been the case, as experience in previous spells showed that pastoralist communities in mid-1990s were hit hard during droughts and armed conflict crisis. The survey asked whether any of the household members engaged in fishing in the last three months prior to the survey. For fishing communities along the River Nile and other main rivers in South Sudan (e.g., Jur River, Sobat River, Lol River, etc.), this variable is an important as livelihood determinant. The main source of income to a household obviously plays a key role in coping with food emergencies. Households dependent on sale of food crops could tend to cope better than others.

4.4. Generalized Logistic Regression

In this section we apply Generalized Logistic Regression Models (GLMs) featuring the Logistic Regression Model for a binary response (or Binary Logistic). GLMs are useful for exploring relationship between a set of predictors and categorical response variable. The models also allows for the dependent variable to have a non-normal distribution. Furthermore, GLMs provide for generating predictor variables of exposure to risk, such as that of food insecurity.

A two stage approach is adopted. First, the Binary Logistic is fitted to the data without considering the complex survey design. Data are analysed using the statistical package SAS version 9.3 (see Appendix 1). In the second stage the IBM SPSS Complex Samples Logistic Regression procedure was used to analyse the data accounting for the complex design. In this case, an overriding assumption was that cluster or random effects had effects on the estimates.

4.4.1. Binary Logistic Model

In applying this procedure, it is assumed that estimates of effects were not affected by random (or cluster) effects. In other words, the complex survey design had no effect in the estimates of fixed effects.

Since the aim of the study is to be able to identify characteristics that are indicative of household resilience in (or ability to cope with) protracted food crisis settings, where populations are likely to face further vulnerability and use those characteristics to identify households at high risk of the crisis, the Binary Logistic Regression Model (or Binary Logit) was the method of analysis of choice. The model is a member of the Generalized Linear Model (GLM) developed by Nelder and Wedderburn (1972) and further rationalized by McCullagh and Nelder (1989).

GLMs are in the domain of statistical models for analysing or determining the relationships between non-normal data (i.e., binary or yes/no responses, multinomial or more than two categories, ordinal categorical and others) and one or more explanatory (or independent) variables. The logistic regressions models for binary data, which include the *logit* model, *probit* model, and the *complementary log-log* are particularly relevant to the data and the type of analysis desired. The procedure was extended to binary data (i.e., data with dichotomous responses or 'yes/no' type data) as demonstrated in Agresti (2002). Appropriate link functions for these models are described below.

The *logit* model builds on the *Probit* function,

$$g(p) = \log(p/(1 - \mathrm{p}),$$

which is the inverse of the cumulative *logit* function, which is

$$F(x) = \frac{1}{1 + \exp(-x)} = \exp(x)\,/(1 + \exp(x))$$

Finally, the logit model is given by the expression

$$Logit\big(P(\mathrm{Y} = 1)\big) = \beta_0 + \beta_1 \mathrm{X}_{i1} + \sum_{k=1}^{3} \beta_{2k} \mathrm{X}_{2ki} + \beta_3 \mathrm{X}_3 + \cdots + \sum_{k=1}^{4} \beta_{7k} \mathrm{X}_{7ki} \qquad (4.1)$$

where $P(\mathrm{Y} = 1)$ is the probability that the i^{th} household will adopt a coping strategy, β_0 is the intercept of the logit model, β_j is the estimated coefficient for each effect j, X_{i1} is male head of the i^{th} household, X_{2ki} is age of head of the i^{th} household, x_3 is the size of the i^{th} household, and X_{7ki} is the main source of income in the i^{th} household. Note that the probability that a household did not adopt a coping strategy is given by $P(\mathrm{Y} = 0)$.

The *probit* (or *normit*) model builds on the *probit* function,

$$g(p) = \Phi^{-1}(p),$$

which is the inverse of the cumulative standard normal distribution, which is

$$F(x) = \Phi(x) = (2\pi)^{-1/2} \int_{-\infty}^{x} \exp(-\zeta^2/2)\, d\zeta\,.$$

Then *probit* model is given by

$$Probit\big(P(\mathrm{Y} = 1)\big) = \Phi^{-1}(P(\mathrm{Y_i} = 1)). \qquad (4.2)$$

The concept of the *Complementary log-log* (or *cloglog*) model derives from the function

$$g(p) = \log(-\log(1 - p)),$$

which is the inverse of the cumulative extreme value function given by

$$F(x) = 1 - \exp(-\exp(x)),$$

resulting in the cloglog model

$$Log(-\log(1 - P(\mathrm{Y} = 1))) =$$

$$\beta_0 + \beta_1 \mathrm{X}_{i1} + \sum_{k=1}^{3} \beta_{2k} \mathrm{X}_{2ki} + \beta_3 \mathrm{X}_3 + \cdots + \sum_{k=1}^{4} \beta_{7k} \mathrm{X}_{7ki}. \qquad (4.3)$$

The Stepwise approach of the Binary Logistic Regression was used to select a model based on 'fit' statistics; that is, *logit*, *probit* or *cloglog*. For each procedure the model with the intercept only (the "null" model) was fitted to the data, followed by one with more parameters ("fitted" model). The Likelihood Ratio (LR) test of significance was then used to test hypothesis. The LR test is given by $\chi^2_{LR} = -2\log(L_0/L_1)$, where L_0 is the maximized value for the "null" model

and L_1 is the maximized value for the "fitted" model. The null hypothesis is stated as $H_0: \beta_j = 0$ for all *j versus* $H_1: \beta_j \neq 0$ for at least one *j*.

The null hypothesis is rejected when at least one of the model parameters is different from zero. The Wald test of significance was also used to test the statistical significance of each of the parameters.

The stepwise approach begins by selecting the independent variable (or effect) that associated most with the outcome; thus demonstrating strong evidence of being the most important (or overall best) predictor of the outcome. In the second step, the effect with second best evidence and subsequent candidates, were chosen in that order. However, a notable drawback of a bivariate approach is that it tends to ignore the possibility of a variable with weak evidence of association with the outcome, becoming an important predictor.

The most common approach for determining unknown parameter estimates for the GLM, is the maximum likelihood (Olsson 2002). The log likelihood is given by

$$L(\beta) = \sum_i L_i = \sum_i \log f(y_i; \theta_i, \phi) = \sum_i \frac{(y_i\theta_i - b(\theta_i))}{a_i(\phi)} + \sum_i c(y_i, \phi). \qquad (4.4)$$

Parameter estimates are obtained by differentiating the log-likelihood function with respect to each β_j, equating the derivatives to zero, and then solving the system of the equations simultaneously for the β_j . Thus, the likelihood equations are then given by

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_j \frac{\partial L_i}{\partial \beta_j} = 0, \text{ for all } j = 0,1,2,\cdots,p. \qquad (4.5)$$

The maximum likelihood estimator of $\widehat{\beta}$ is derived using the Chain Rule and the Newton-Raphson method. For more thorough explanation of the theory of GLM and estimation of

parameters, see McCullagh and Nelder (1989) and Agresti (2002). For derivation of the procedure for binary data see Collett (2003). Hypothesis testing of the model assumptions, model prediction and estimates of standard errors is done using the maximum likelihood criterion. For more on these, see Olsson (2002) and Collett (2003).

It is evident from the foregoing narrative that the Binary Logistic Regression model is monotone depending on the sign of β. That is, $P(\mathrm{Y}=1)$ increases as β is positive and decreases otherwise.

In this analysis the unadjusted odds ratios (OR) were used. Unadjusted OR is a simple ratio of probabilities of outcome in two groups p_1, p_2. In this case, the odds of a *logit* model are calculated based on an exponential function of X. Thus

$$P(I_i = 1)/1 - P(\mathrm{Y}_i = 1) =$$

$$\beta_0 + \beta_1 \mathrm{X}_{i1} + \sum_{k=1}^{3} \beta_{2k} \mathrm{X}_{2ki} + \beta_3 \mathrm{X}_3 + \cdots + \sum_{k=1}^{4} \beta_{7k} \mathrm{X}_{7ki} \qquad (4.6)$$

is the odds of household *i* adopting a coping strategy during crisis situation. Interpretation of odds ratio estimates is meaningful. For every unit increase in X the odds increase by e^{β_j}. However, there are cases when one can include other confounding variables so as to control their influence on the dependent variable. This results in an OR that is adjusted for the influence of those confounding variables. Thus adjustment is carried out by controlling additional variables in the logistic regression model.

Evaluation of the fitted models was conducted using the AIC, SC and Deviance criteria. The Likelihood Ratio, Score and Wald tests of hypothesis were also used, comparing the fitted with the observed counts. Finally, the Hosmer-Lemeshow test (Hosmer & Lemeshow 2000) was employed to assess the model goodness of fit.

As stated earlier, one of the aims of the study is to generate a single measure, which for the purpose of this study we call 'the Household Resilience Index (HRI)', in order to be used as a tool for early warning and monitoring of the likelihood of vulnerability. HRI is simply generated based on the scores of predicted values of the linear predictor of the binary response variable (i.e. coping or no coping with food shortage, or 1 or 0 respectively). For more on how to determine the linear predictors, see Hosmer and Lemeshow (2000). The statistical applications SAS® (2011b) and IBM SPSS (2013) have functionalities for generating and saving predictor values. A further step taken was to compare the predicted values of the response categories against observed values to determine the validity of predictions.

Perusing the data in Table 4.1, one gets a clearer picture of poverty or food insecurity traps or the issue of low resilience in the sample population. A glaring finding is that although the study households were largely (96%) in the economically active age group of 18 to 60 years and male headed (73.5%), only about 22 *per cent* of them depended on employed labour. A significant number of the households had to sell their livestock and products to subsist; not a common practice for pastoralist communities in South Sudan where sale of livestock was a last resort (or kind of extreme coping strategy). Meanwhile, dependence on crop cultivation for sale would have been the most common form of survival, but only less than a third were engaged in the activity. In general, the data showed a sense that the sale of major sources of food in a population characterized by subsistence economy, could reveal substantial level of coping with a food insecurity emergency. It was, therefore, reasonable to explore how the sources of livelihood and resilience of this population predicted coping with food emergency in such setting. Determining the factors that contributed the most to the incidence of coping with food crises is of essence, but it is also prudent to know how the procedure is a potentially robust tool assessing the likelihood of future emergencies as well as for early warning systems.

In fitting the Binary Logistic Regression model it was assumed that the variables included in the model, i.e., assets and livelihood amenities, were independent and uncorrelated. This method also does not account for the survey design used for collecting the data. Rather it assumes that the data are from a simple random sampling. Furthermore, the procedure uses the Maximum likelihood (ML) estimation for unweighted observations and for constructing likelihood equations based on standard distributional assumptions to obtain the ML estimates of the model coefficients and the corresponding covariance matrix estimates.

Mathematically, the Binary Logistic Model is fitted as the transformation of π instead of fitting a model for π as in the Ordinary Linear Regression Model. In the transformation the odds of a "success" outcome is used, i.e.,

$$odds = \frac{P(success)}{P(failure)} = \frac{P(success)}{1-P(success)} = \frac{\pi}{1-\pi} \tag{4.7}$$

This means that the odds are defined as the probability of a "success" divided by the probability of a "failure". Conversion from probabilities to odds and back again is easy. Note that the odds can assume values between 0 to ∞. This being the case, the odds can be thought of as another scale for representing probabilities. Since division by zero is not permitted, the odds will be undefined when the probability of "failure" (i.e., $1 - \infty$) is 0. The logistic regression model for the odds is of the form

$$\frac{\pi}{1-\pi} = e^{(\beta_0+\beta_1 X_1+\cdots\beta_k X_k)} \tag{4.8}$$

It transpires that the range of values that the right-hand side can take is now between 0 and ∞ in the model, which is the same range as that of the left-hand side. Alternatively, the logistic regression model (4.8) can be expressed in terms of log odds of success, which is known as the *logit* form of the model

$$\log\left(\frac{\pi}{1-\pi}\right) = \text{logit}(\pi) = \beta_0 + \beta_1 \text{X}_1 + \cdots + \beta_k \text{X}_k \quad (4.9)$$

Notice that the logit is another transformation of the probability of success π. In fact, it is the (natural) logarithm of the odds of the success, which results in a linear model on the logit scale; thus the more common form of the logistic regression model. Rewriting the model in terms of the probability of a "success" outcome gives

$$\pi = \frac{e^{(\beta_0+\beta_1 \text{X}_1+\cdots+\beta_k \text{X}_k)}}{1+e^{(\beta_0+\beta_1 \text{X}_1+\cdots+\beta_k \text{X}_k)}} \quad (4.10)$$

Since the aim behind the analysis is to quantify the relationship between the probability of a "success" outcome, π, and the explanatory variables $\text{X}_1, \text{X}_2, \cdots, \text{X}_\text{k}$ based on some sample data, it is reasonable to assume that in the population there is a relationship between π and a single continuous explanatory variable X and that this relationship is of the form

$$\text{logit}(\pi) = \log\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 \text{X} \quad (4.11)$$

Using statistical software procedure, the model can be estimated as

$$\text{logit}(\hat{\pi}) = b_0 + b_1 \text{X}, \quad (4.12)$$

where b_0 and b_1 are the estimated regression coefficients. The estimation for logistic regression is commonly performed using the statistical method of maximum likelihood estimation.

4.4.2. Results of the Binary Logistic Model without Accounting for Random Effects

In modelling the data '*coping with food emergency*' (which takes the value 1) was treated as the response, while '*not coping*' (with 0 as its value) was taken as the reference category. This

means that the saved probabilities of the model estimate the chance that a given household takes the value 1; thus parameter estimates should be interpreted as relating to the likelihood of category 1 (*adopted a coping strategy*).The procedure generates goodness-of-fit statistics, which provides useful measures for comparing competing models. Recall that the data consist of household characteristics and means of agriculture-based livelihood factors, and whether or not a household coped with food insecurity risk.

The Binary Logistic Model tested each of the independent variables for association with the dependent variable (whether or not a household adopted a coping strategy), taking one variable at a time. Three tests of association were used: the Likelihood Ratio test, the Efficient Score test and Wald test. The results shown in Table 4.2 are for testing the null hypothesis that there is no difference between the levels of each effect in their associations with the dependent variable, that is, all slopes of the parameters ($\hat{\beta}_j's$) are equal to zero. The small *p*-values reject the hypothesis that all slopes are equal to zero.

Table 4.2: Testing Global Null Hypothesis $\boldsymbol{\beta = 0}$

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	151.2959	11	<0.0001
Score	143.7646	11	<0.0001
Wald	131.0118	11	<0.0001

Summary of the Stepwise selection procedure of the *cloglog* link function is displayed in Table 4.3. It is shown that five of the seven variables included in the model, were determined to be highly significant. Only *crop cultivation* and *fishing* were determined to be non-significant. Variables are ranked according to importance. The most important (or most influential) effect was selected in the first step, followed by the second most important, and so on. Accordingly,

Source of income was determined to be the most important variable followed by ownership of livestock; thus both effects can be considered as prime determinants (or highly influential factors) of coping with food insecurity risk.

Table 4.3: Summary of Stepwise Selection

Step	Effect Entered	DF	Number In	Score Chi-Square	Pr > ChiSq
1	Source of income	4	1	81.6455	<.0001
2	Owned livestock	1	2	37.7775	<.0001
3	Age of household head	2	3	13.6115	0.0011
4	Gender of household head	1	4	6.8419	0.0089
5	Household Size	1	5	6.8774	0.0087

Next, is to examine output of the Type 3 analysis of effects based on the Score Test (Table 4.4). *Crop cultivation* and *fishing* were determined to be non-statistically significant (*p*-values of 0.5202 and 0.4695, respectively), suggesting that there was no evidence that having cultivated crop, or not differentially affect coping with food insecurity, and no evidence that having fished or not affected coping differently. The two variables were thus not related to coping with food insecurity.

Table 4.4: Type 3 analysis of effects included in the model

Effect*	DF	Wald Chi-Square	Pr > ChiSq
Age of household head	2	12.5953	0.0018*
Gender of household head	1	8.0092	0.0047*
Household size	1	7.0687	0.0078*
Owned Livestock	1	25.6307	<0.0001*
Source of income	4	67.6762	<0.0001*

* Significant values (*p*-value<0.05)

Age and *gender* of household head were determined to be highly significantly associated with outcome of coping, indicating that there were significant differences between age groups and gender of household heads in how households coped with food insecurity risk during the conflict crisis. The same could be said of *household size* in terms of number of individuals living in them. Analysis also showed that *ownership of livestock* had highly significant relationship with coping outcome. There was sufficient evidence to indicate that *sources of income* had highly significant association with outcome of coping.

Serious caution, should, however, be taken when interpreting the result of the a*ge of household head* given the imbalance of its categories, as shown in Table 4.1, where 96 per cent of the households were in one group; 18 to 60 years! Similar sharp imbalances were noted in *ownership of livestock* and *fishing*. Such imbalances often lead to poor estimates and misleading results! The problem emanated right from the raw data and during data collection. The survey designers decided to use age categories to avert respondent bias due to low literacy revels, as some respondents might fail to know their dates of birth. It was then considered easier to determine whether a respondent was young adult (< 18 years), adult in the economically active age group (18 to 60 years), or over 60 years. However, this strategy in itself was problematic in that the middle age group was too large and naturally an overwhelming proportion of respondents occurred in it.

Results of tests of effects shown in Table 4.4 motivate exploring the magnitude of the relationships between each significant effect with the outcome of coping by examining parameter estimates (Table 4.5).

Table 4.5: Analysis of Maximum Likelihood Estimates (MLEs)

Parameter	DF	Estimate	Std. Error	Wald Chi-Square	Pr>ChiSq
Intercept		-0.8706	0.1560	31.1457	<0.0001
Gender: *male*	1	0.0807	0.0285	8.0092	0.0047
Age of household head: *< 18*	1	0.4394	0.1513	8.4297	0.0037
18-60	1	0.0456	0.0927	0.2423	0.6226
Household size	1	-0.0256	0.00963	7.0687	0.0078
Owned livestock: *yes*	1	0.5583	0.1103	25.6307	<0.0001
Source of income: *crops sale*	1	0.2307	0.0447	26.6666	<0.0001
Livestock products	1	-0.3070	0.0540	32.2626	<0.0001
Employment	1	0.2033	0.0483	17.7136	<0.0001
Petty trade	1	0.0184	0.0511	0.1288	0.7196

The result in Table 4.5 show that a household headed by a male aged less than 18 years, owned livestock and depended on sale of crops and livestock products and employment for income, associated significantly with non-adoption of coping strategy. The negative coefficient of the scale variable *household size* indicates that households with a smaller number of members did not adopt any coping strategy. These results should, however, not to be read too much into given the noted problem of imbalanced variable categories. It is to be noted that the larger variable categories were fixed as the reference categories.

Further analysis involved the use of goodness of fit test for assessing the overall fitness of the selected *cloglog* model. This is done by obtaining the Chi-square difference between the model

with intercept only (or "null" model) and the model containing one or more predictors (or "full" model). The Chi-Square test value of the Likelihood Ratio (i.e. X_{LR}^2) = 150.3997 with 9 degrees of freedom (df) and *p*-value<0.0001 indicates that a significant increase in the likelihood; thus a further indication of a good model. This is calculated by obtaining the difference between the Deviance (D) of the "null" model (M_0) and the "full" model (M_F), yielding $X_{LR}^2 = D_{M_0} - D_{M_F} \sim \chi_{(p-1)}^2$. The Deviance of the "null" model (D_{M_0}) was 4843.020 with 11 degrees of freedom (df), and for the "full" model (D_{M_F}) was 4993.419 with 2 degrees of freedom; thus giving the difference of 150.399 with 9 degrees of freedom.

Nevertheless, the Hosmer and Lemeshow (2000) test shows some evidence of lack of fit in the selected model (p=0.0433) at the 0.05 level of significance. This is expected, as the data are from a large sample survey with the elements of complexity and randomness unaccounted for. This gave reason to explore a model that accounted for random effects (Sub-section 4.4.3).

The next step was to examine the influence diagnostics of the selected model. The index plots of the Pearson Residuals and the Deviance Residuals (Figure 4.1) show cases that are poorly accounted for by the model. Fortunately, these are not so many considering that the sample size of this study is 3692. The index plot of the diagonal elements of the hat matrix gives the extreme point(s) in the design space. These are also not as many, which confirm earlier results suggesting good fitness of the *cloglog* model. Unfortunately, with so many cases numbering in thousands, it was not practical to delineate the cases of high influence.

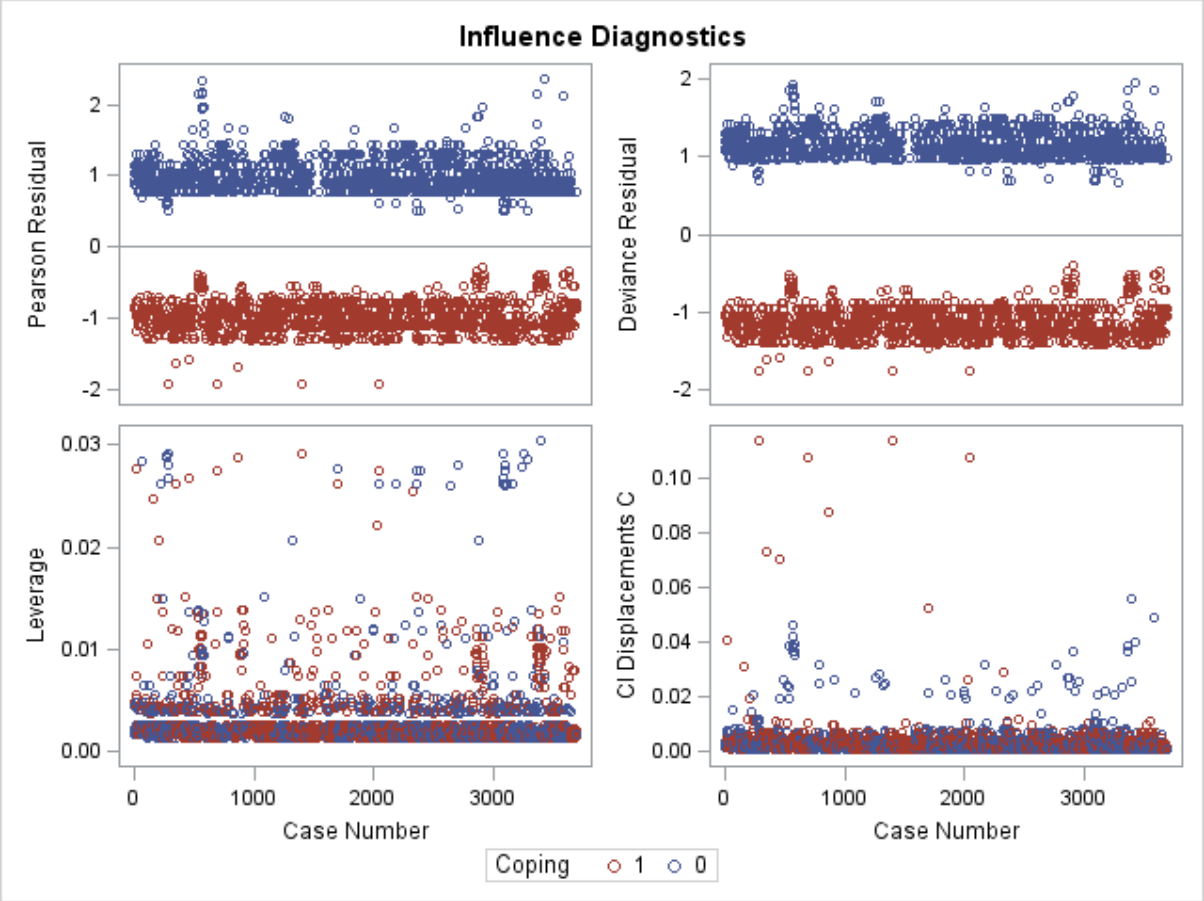

Figure 4.1: Plots of residuals, hat matrix, and CI displacement C values

Other sets of influence diagnostic plots against the predicted probabilities were also conducted (Figure 4.2). Furthermore, plots of several diagnostics were conducted against the leverage. For instance case number 3397 was identified to be of high leverage (i.e., greater than $\frac{2p}{n}$).

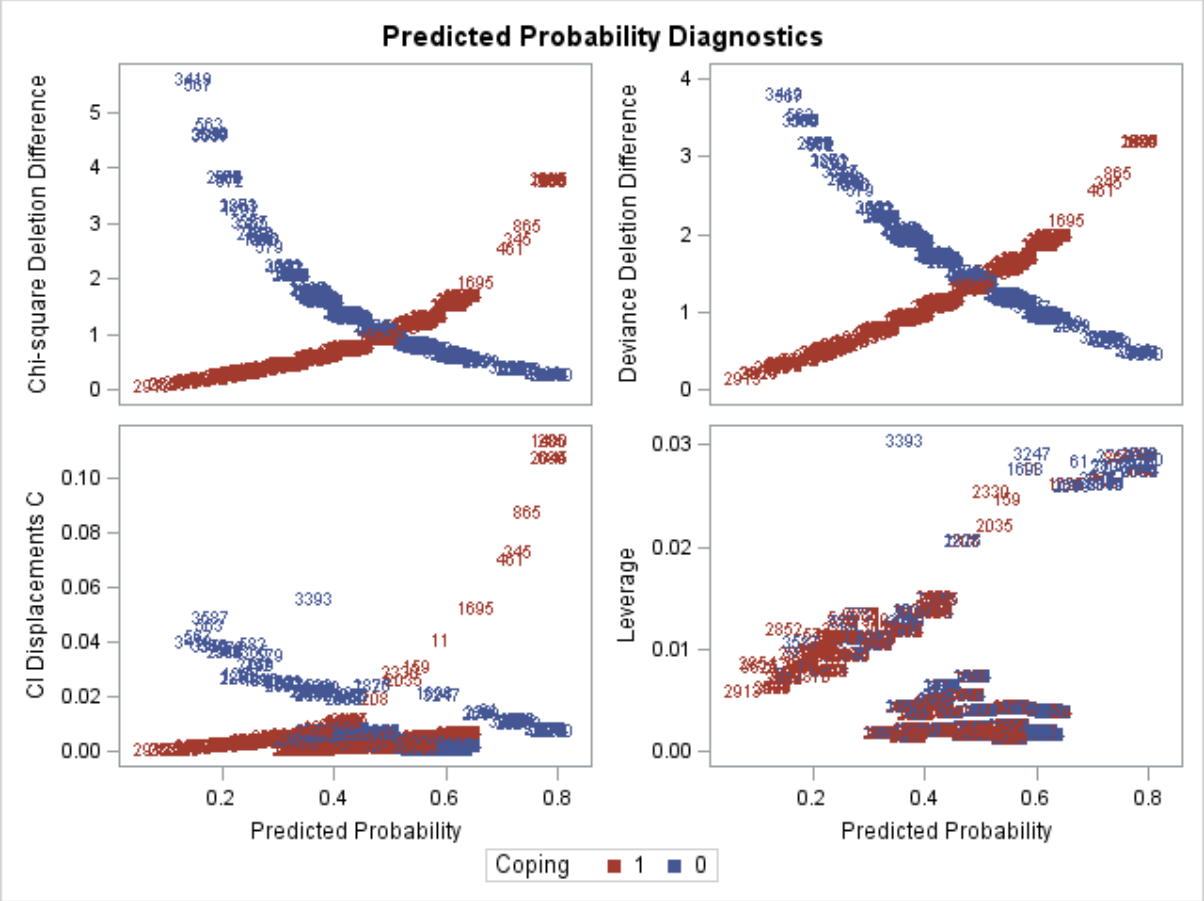

Figure 4.2: Diagnostics versus predicted probability

The influence and predicted probability diagnostics presented in Figure 4.1 and Figure 4.2 indicate that the fitted model was good. It shows that the binary logistic model adequately describes the data. It also shows that the selected variables were good predictors of coping with food insecurity risk in protracted crisis.

The last step in the analysis was to examine how correctly the model predicted whether a household had to adopt a coping strategy or not. The model generated predicted probabilities corresponding to each case (household) and for either level of the response variable (i.e., experience of coping). It is reasonable to consider the predicted probabilities of *Coping=0* (the reference category of the response variable) as a composite index since it is generated from

selected model. The new variable of 'predicted responses' can then be taken as the 'Household Resilience Index', as it is calculated based on predicted probabilities of factors that determine how resilient a household could be, or that a household was able to cope based on combination of favourable factors.

4.4.3. Results of the Binary Logistic Model Accounting for Random Effects

Since the data came from a multi-stage sampling design, it was seen rational to take into consideration the complex design that accounted for cluster effects. Analysis will then examine the estimates of the design effects. The SAS Version 9.3 Generalized Linear Mixed Model (GLMM) procedure with Gauss-Hermite Quadrature Likelihood Approximation was applied for data with the binary response '*coping*' or '*no coping*'. Analysis produced the results below. It is to be recalled that the main aim of the analysis is to see if the probability of a household coping with food insecurity risk during the period of crisis in South Sudan is related to each of the seven selected explanatory/independent variables, namely; *gender of household head*, *age of household head*, *household size*, *crop cultivation*, *livestock ownership*; *fishing* and *source of income*.

It is reasonable to start with examining the output of the covariance parameter estimates. As parameter estimation is based on maximum likelihood, it was reasonable to conduct a formal test of the hypothesis of no cluster variability. The variance of the cluster effect is estimated as 1.3790 with estimated asymptotic standard error of 0.2196. This indicates that there was significant between-cluster variability and thus sufficient evidence for rejecting the null hypothesis of no cluster variability.

Table 4.6: Solution for fixed effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	1.9795	0.5810	149	3.41	0.0008
Gender: *male*	-0.03245	0.08983	3441	-0.36	0.7180
Age: *<18*	-1.2668	0.4535	3441	-2.79	0.0052
Age: *18-60*	-0.5517	0.2472	3441	-2.23	0.0257
Household Size	0.01399	0.01600	3441	0.87	0.3821
Crop cultivation: *yes*	-0.2117	0.1102	3441	-1.92	0.0547
Livestock ownership: *yes*	-1.0553	0.4954	3441	-2.13	0.0332
Fishing: *yes*	-0.07913	0.1392	3441	-0.57	0.5699
Income: *sale of crops*	-0.4852	0.1832	3441	-2.65	0.0081
Livestock products	-0.05759	0.1895	3441	-0.30	0.7612
Employment	-0.3686	0.1871	3441	-1.97	0.0490
Petty trade	-0.1538	0.1873	3441	-0.82	0.4114

The 'solution for fixed effects' results (Table 4.6) shows that *age of household head*, *crop cultivation*, *livestock ownership* and *main source of income* had significant associations with adoption of coping strategy. There was no statistically significant evidence to suggest that *gender of household head*, *household size* and *fishing* had associations with adopting a coping strategy. This result is quite different from that of the binary model without consideration of random effects, where *gender* and *household size* were significant effects. Fishing has persistently remained non-significant in both approaches. However, as noted earlier, this occurrence could be to acutely disproportionate (unbalanced) categories.

The preceding result compares that of Type 3 tests of fixed effects (Table 4.7).

Table 4.7: Type 3 tests of fixed effects

Effect	Num DF	Den DF	F Value	Pr > F
Gender	1	3441	0.13	0.7180
Age	2	3441	4.29	0.0137
Household size	1	3441	0.76	0.3821
Crop cultivation	1	3441	3.69	0.0547
Livestock ownership	1	3441	4.54	0.0332
Fishing	1	3441	0.32	0.5699
Main source of income	4	3441	4.32	0.0017

The odds ratios generated by the procedure are further shown in Table 4.8. This table displays the odds ratios of *'coping'* with food insecurity at the factor levels of the categorical variables included in the model. In the first row, the reported values were the ratios of the odds of '*coping*' for a male headed household compared to the odds of '*coping*' for female headed household. Thus, the odds ratio of 1.304 in the first row of the table means that the odds of '*coping*' for a household headed by a male are 0.968 times compared to those of a household headed by a female. In this case, a female headed household did relatively worse in terms of coping with food insecurity risk compared to one headed by a male. Similarly, the odds of a household headed by a younger person (i.e., an economically active person), were better compared to those of one headed by a person aged over sixty years. Thus, a household headed by a person in the above sixty years category, was more at risk of food insecurity, implying that such household might not cope well during a crisis situation. In general, the odds of a household headed by a male aged 60 years or less, who cultivating crops, owned livestock, lived on sale of crops and employed, were better compared to those of households without these attributes.

Table 4.8: Odds Ratio estimates of fixed effects

Fixed Effect	Category	Odds Ratio	95% Confidence Interval	
			Lower	Upper
Sex of household head	*male vs female*	0.968	0.812	1.155
Age of household head	*<18 vs >60*	0.282	0.116	0.686
	18-60 vs >60	0.576	0.355	0.935
Household size	*Not applicable*	1.014	0.983	1.046
Cultivated crops	*yes vs no*	0.809	0.652	1.004
Owned livestock	*yes vs no*	0.348	0.132	0.919
Engaged in fishing	*yes vs no*	0.924	0.703	1.214
Main source of income	*Agric vs other*	0.616	0.430	0.882
	Livestock vs other	0.944	0.651	1.369
	Employment vs other	0.692	0.479	0.998
	Petty trade vs other	0.857	0.594	1.238

4.4.4. Discussion

The foregoing analysis was based on data collected at the peak of a man-made crisis that affected millions of South Sudan's population. At the time, there was widespread internal displacement reported. The World Food Programme's Vulnerability Assessment and Mapping (VAM) Update (2014) reported that "Conflict continued to uproot and displace households, preventing many from planting and forcing them to sell off assets and livestock for food". This forced 51 *per cent* of the households in the sampled population to employ some form of coping strategy. It was reported that levels of coping varied between states, with three highly affected states of Unity, Upper Nile and Jonglei have a proportion as high as 60 *per cent* of households that adopted some coping strategy. Diet coping was even worse (over 70%) in four of the ten states: Northern Bahr-el Ghazal, Unity, Upper Nile and Jonglei (World Food Programme 2014).

Data analysis applied the Binary Logistic Regression technique of the family of Generalized Logistic Regression Models. The simple Binary Logistic Regression Model was used under the underlying assumption that the data were not affected by random effects resulting from the complex sampling design used.

The study results to a greater extent confirmed what could be a common perception that agriculture-based rural livelihoods can enable a household to cope with crisis riddled with food insecurity chiefly characterized by widespread lack of food or means to access it. However, availability of regular source of income could even provide stronger resilience to crisis-affected households. It implies that lacking cash from any source could prove a high risk to displaced households or populations affected by crisis of the scale of South Sudan in 2014 and beyond.

Meanwhile, the results showed that families that had cultivated crops three months prior to data collection, but during the crisis, had to resort to at least one form of copying strategies. This finding confirms that of Gitz and Maybeck (2012) who opine that a household reliant on agricultural production become vulnerable to agricultural production shocks. In a conflict situation, farming households might be continually displaced to harvest their crops or it becomes too dangerous for them to return to their farms. Thus, conflict can be a form of production shocks.

The findings also point to that fishing was not an influential effect of coping with crisis in South Sudan. However, this result should be taken with a pinch of salt consider the lack of balance in the categories of this variable. Like crop cultivation, communities dependent on fishing for their livelihood are supposed to cope well in food insecurity crisis. It is obvious that the sample design did not take into consideration a proportionate number of households from fishing communities. The solution to this problem lies in either removing the variable from the

analysis, or including sampling weights in the analysis. The latter option is the subject of the method to follow.

Unlike crop cultivation and fishing, livestock was determined to as a significant effect of coping during the crisis in South Sudan. The results established that livestock keeping to a greater extent assures food security of populations; thus enabling households in possession of this type of asset to cope in crisis situation. However, this depends on the nature of the crisis. Crises emanating from climate change such as drought, which kills large herds of animals, could present a different type of impact on food insecurity. Households might actually resort to selling their herds at give-away prices to avoid losing everything. In South Sudan livestock keeping is regarded as form of livelihood assurance, and by extension food security.

The results showed that the best overall determinant of consumption coping in the crisis was source of income. Analysis involving estimates of parameters and odds ratio determined that income from sale of agricultural produce, sale of livestock products, wages and salaries associated with non-adoption of consumption coping strategy. This implies that resilience of households to food and humanitarian crisis could be boosted if households were empowered with more entrepreneurial skills in farming and agribusiness, livestock production and other forms of production. It further reinforces the findings by Slater et al (2015) and Tirivayi et al (2013) that income from cash transfer provided to farming households improve agricultural production and livelihoods.

Despite the reported results of analysis, the Binary Logistic Regression procedure does not account for complexity of the survey. A procedure such as the SAS (2011a) Survey Logistic might improve the result.

4.4.5. Conclusion

The study result had satisfied the two objectives of the study. First, the results established the likely predictors of or factors that affected how household coping with food insecurity in a crisis where food consumption is strained. Second, the results determined that, barring issues with the data and lack of fit, the new variable created from a set of generated predicted values, could serve the purpose of being a food resilience index. This composite index, if validated using more controlled sample, could provide an answer to the search for a single efficient measure of food resilience.

Findings of the two different approaches under different assumptions were examined. That is, analysis under the assumption that the data were not affected by complex design effects and another that took into consideration the complex design effects. It latter approach determined that there were between-cluster effects. The results of the Binary Logistic presented in Table 4.3 and those of the Complex Samples Design Logistic Regression give dissimilar significance results. Thus, it is always useful to take complex design effects into consideration. The Survey Logistic Regression procedure was introduced below for the purpose of comparing results with the foregoing analysis.

4.5. The Survey Logistic Regression

This Section examines the Survey Logistic Regression (SAS Institute 2011a) for exploring the relationship between factors typifying rural livelihoods and food consumption, which was used as a proxy for measuring food insecurity risk. The motivation for using the Survey Logistic Regression procedure is to determine predictors of food insecurity in a typical setting where resilience of population is weakened as a result of protracted crises.

4.5.1. An Overview of the Survey Logistic Model

Survey Logistic Regression model is a member of the Logistic Regression Models which is used to model data from a complex survey design. The model accounts for the complexity of survey design, i.e., it takes into account the effects of stratification and clustering used in the survey design.

The theory of both the survey logistic regression model and the ordinary logistic regression model are the same. The only difference is in the estimation of variances. If the data are from a simple random sampling then the survey logistic and the ordinary logistic give identical estimates. But if the data are from a complex survey design, then the estimates of the coeffcients and the standard errors will be different because of the effects of stratification and clustering.

In this section we discuss the effects of both sampling-survey design and weights on the data structure. As described in Chapter 2, the sampling technique used, is based on stratified random sampling. This was done in two stages. In the first stage, clusters of villages (on in the case of large towns, zones/townships) were selected at random from each county of the ten states. Clusters then formed strata made groups of villages (or in the case of urban centres, residential areas/localities). In the second stage, households were sampled from clusters. Let the response variable be denoted by Y_{ij}, where $i = 1,2,\cdots,m_j$ and $j = 1,2,\cdots,n$, which equals 1 if the i^{th} household adopting a coping strategy was located in the j^{th} cluster, and *0* otherwise. Note that, *j* is the stratum or cluster and *i* is the individual household. Note further that states and counties are not included in the notation because they were not randomly selected; they were constants. Let $\pi_{ij} = p(y_{ij} = 1)$ be the probability that the i^{th} household adopted coping strategy within j^{th} cluster (or stratum). Then the survey logistic model is given by

$$log(\pi_{ij}) = \boldsymbol{X}'_{\boldsymbol{ij}}\boldsymbol{\beta} \qquad 4.13$$

where the subscripts i and j are as defined above and $\boldsymbol{X_{ij}}$ is the row of the design matrix corresponding to the characteristics of the i^{th} household in the j^{th} cluster, and $\boldsymbol{\beta}$ is the vector of unknown parameters of the model. Hence, the log likelihood function is given by

$$l(\beta; y) = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \sum_{i=1}^{m_{hj}} \left[y_{ijh} \log\left(\frac{\pi_{ijh}}{1-\pi_{ijh}}\right)\right] - \log\left(\frac{1}{1-\pi_{ijh}}\right) \qquad 4.14$$

Parameter estimation of the Survey Logistic Regression model can be calculated using a number of methods. Heeringa et al. (2010) observe that the likelihood function for the simple random sampling of n observations on binary response variable y with possible values 0 and 1, is based on the binomial distribution.

$$l(\beta|x) = \prod_{i=1}^{n} \pi(x)^{y_i}[1-\pi(x)]^{1-y_i} \qquad 4.15$$

where π(x) is linked to the coefficients of the regression model and evaluated through the logistic cumulative distribution function(cdf)

$$\pi(x_i) = \frac{exp(x_i\beta)}{1+exp(x_i\beta)} \qquad 4.16$$

The parameters of the logistic regression model parameters and standard errors can be estimated using the method of maximum likelihood discussed in Sub-section 4.3.3. However, if the survey data is collected from a complex sample design, application of the maximum likelihood estimation (MLE) is no longer possible. This is because: 1) the probabilities of selection for the sample observations $i - 1,2,\cdots,n$ are no longer equal, due to the stratification and clustering of the survey design. Thus, sampling weights are required to estimate the finite population values of the parameters for the logistic regression model; and 2. the stratification and the clustering of the complex sample observation violates the assumption of independence of the observations that are crucial to the standard maximum likelihood approach used for

estimating the sampling variance of the model parameters as well as for choosing a reference distribution for the likelihood ratio test statistics (Heeringa et al. 2010). Generally, there are several methods used to estimate the covariance matrix (variance estimation) of the parameter estimates for data from complex survey designs. Among these methods are: the Jacknife method; the Pseudo-maximum likelihood method; the Taylor series (linearization) Method; the Balance Repeated Replication (BRR) Method; Fay's BRR Method; and the Hadamard Matrix. The variance can be easily estimated by these methods using SAS version 9.3) Survey Logistic procedure. The default Taylor series (Linearization) method was used because it estimates of between cluster variances. Taylor's method is the most commonly used method to estimate the covariance matrix of the regression coefficients for complex survey data. Generally, variance estimation can be estimated using the Taylor series (Linearization) method. The estimated covariance matrix of the model parameter $\hat{\beta}$ by the Taylor series method is given by

$$\hat{V}(\hat{\beta}) = \hat{Q}^{-1}\hat{G}\hat{Q}^{-1}, \qquad 4.17$$

where

$$\hat{Q} = \sum_{j=1}^{n}\sum_{i=1}^{m_j} w_{ji}\hat{D}_{ji}\left(diag(\hat{\pi}_{ji} - \hat{\pi}_{ji}\hat{\pi}'_{ji})^{-1}\hat{D}'_{ji}\right)$$

$$\hat{G} = \frac{n(1-f_j)}{n-1}\sum_{i=1}^{m_j}(e_j - \bar{e}_j)(e_j - \bar{e}_j)'$$

$$e_j = \sum_{i=1}^{m_j} w_{ji}\hat{D}_{ji}\left(diag(\hat{\pi}_{ji} - \hat{\pi}_{ji}\hat{\pi}'_{ji})^{-1}(y_{ji} - \hat{\pi}_{ji})\right)$$

$$\hat{e}_j = \frac{1}{n_j}\sum_{j=1}^{n_j} e_j$$

where,

$\boldsymbol{D_{ji}}$ is the matrix of the partial derivatives of the link function g, with respect to β

and $\widehat{D}_{ji}$ and the response probabilities $\hat{\pi}_{ji}$ evaluated at $\hat{\beta}$.

$j = 1,2,\cdots,n$ is the cluster or stratum index,

$i = 1,2,\cdots,m_j$ is the observation index within cluster j.

n is the total sample size.

$\boldsymbol{Y_{ji}}$ is a D-dimensional column vector whose elements are indicator variables for the first D categories of the variable Y. If the response of the j^{th} unit of the i^{th} cluster falls in category d, the d^{th}element of the vector is one, and the remaining elements of the vector are zero, where $d = 1,2,\ldots,D$.

w_{ji} is the sampling weight.

π_{ji} is the expected vector of the response variable.

f_j is the sampling rate for stratum j.

Finally, p is the number of covariates in the model.

For further discussions, see Lehtonen and Pahkinen (1995), and Hosmer and Lemeshow (2000) for further discussion on fitting logistic regression models to data from complex survey designs.

Inference and hypothesis tests for the survey logistic model can be calculated using the likelihood ratio, the score statistic, and the Wald statistic, to test the null hypothesis that assumes that all the explanatory effects in the model are zero. As was the case with the ordinary logistic regression model, the decision on whether to reject or not reject the null hypothesis is based on the chi-square test and the p-value. Thus, the null hypothesis is rejected at a *p*-value of less than 0.05, otherwise it is not rejected. The Wald chi-square statistic can also be used to test the significance of the model parameters (Heeringa et al. 2010). If the sample size is large, then the sampling distribution of the parameter estimators is approximately normal. The Wald chi-square statistic used for testing the significance and construction of the parameters confidence interval for the survey logit model, is given by

$$\hat{\beta}_j \pm z_1 - \frac{\alpha}{2}\sqrt{\sigma_j^2}$$

where $z_1 - \frac{\alpha}{2}$ the $100\left(1 - \frac{\alpha}{2}\right)^{\text{th}}$ percentile of the standard normal distribution, and *j* is the variance of $\hat{\beta}_j$ given by the diagonal elements of the variance covariance matrix of $\hat{\beta}$. It can be noted here that if the data is from a simple random sampling, then the logistic model and the survey logistic model are identical. However, if the data is from a complex sample design, then the survey logistic model uses the Pseudo maximum likelihood estimation or the Taylor linearization approach to estimate the variance estimates.

The study generally explores use of a method known to be appropriate in accounting for complex and large designs, which studies described in earlier chapters of this thesis did not

consider. The following sub-sections explain the important components of the analytical methods. Stratification and clustering of the complex sample design as done in some surveys, generally has an impact on the accuracy of both the model variance estimates and the test statistics. It is of essence examining whether or not the parameter estimates will change when the complexity of the survey design is taken into account by fitting a main effect model using survey logistic technique. A distinct measure in the procedure is that it accommodates sample weights.

As data for this study were drawn from a stratified survey sampling, the Survey Logistic Regression model was used to model data from a complex survey design (SAS Institute 2011b). The method accounts for the complexity of survey design, that is, it takes into account the effects of stratification, clustering used in the survey design and unequal assignment of sampling weights. The theory of both the survey logistic regression model and the ordinary logistic regression model are the same. The only difference is in the estimation of variances. If the data are drawn from a simple random sampling then the survey logistic and the ordinary logistic give identical estimates of the variances. But, if the data are from a complex survey design the estimates and their standard errors will be different due to the effects of stratification and clustering. The effects of both sampling-survey design and weights on the data structure are discussed in the next section of the chapter. Application of the method follows the works of Cox and Snell (1989), Walker and Duncan (1967), Morel (1989), Rao et al (1992a) and Roberts et al. (1987a). The procedure fits linear logistic regression models for survey data with discrete responses based on the maximum likelihood estimation. For more and exhaustive review of the statistical theory and mathematical formulation, please refer to Rao (1992a), Heeringa et al. (2010), Walker and Duncan (1967), Cox and Snell (1989) and McCullagh (1980).

The maximum likelihood estimation is carried out with either the Fisher scoring algorithm or the Newton-Raphson algorithm. One can specify starting values for the parameter estimates. The logit link function in the ordinal logistic regression models can be replaced by the probit function or the complementary log-log function.

After fitting the model odds ratio estimates, parameter estimates and variances of the regression parameters are computed by using either the Taylor series (linearization) method or replication (resampling) methods to estimate sampling errors of estimators based on complex sample designs. For more on the notation and mathematical derivation of these statistics, see Binder (2003), Särndal, Swensson, and Wretman (1992), Wolter (2007) and Rao, Wu, and Yue (1992b).

The model was evaluated using the Akaike's Information Criteria (AIC) and the Schwarz Criteria (SC). These criteria were used to impose penalties on the likelihood ratio statistic -2logL (Agresti 2004). Generally, the decision on either the AIC or the SC criterion is the best, depends on the objectives of the study and the more appealing model; thus if the interest is in the consistency of the approximation and the model fit, a model based on AIC is preferred (Burnham & Anderson 2002). However, if interest is in the order of the model, then a model based on SC is preferred. For further discussion on model selection refer to Buckland et al. (1997) and Burnham and Anderson (2002).

Results of goodness-of-fit tests were not presented. Due to the complex sampling designs, existing software were not yet developed or implemented for these tests based on the logistic regression. According to Archera et al (2006) available software usually take the form of simulation studies in which results of analysis were compared with ordinary goodness-of-fit statistics. For instance, the Hosmer and Lemeshow goodness-of-fit test statistic, the Pearson residual, and the deviance residual test, are not yet incorporated in the Survey Logistic

procedure. Thus, for the assessment of the goodness-of-fit of the model used, the Akaike Information Criterion (AIC), the Schwarz Criterion (SC), and the -2log likelihood statistic were used as approximations for the goodness-of-fit test.

4.5.2. Results of the Survey Logistic Procedure

Stratification and clustering of the complex sample design often done in a number of nationwide surveys has an impact on the accuracy of both the model variance estimates and the test statistics. In this section we examine whether or not the parameter estimates will change when the complexity of the survey design is taken into account, by refitting the main effect models of Section 4.4.

The Survey Logistic Model was fitted to the data with seven independent variables and the binary response coping with food insecurity during a period of crisis in South Sudan. The rationale for selecting the seven variables is given in Section 4.3 above. Analysis considered finite population correction with strata, clusters and survey weights included in modelling the data. The Fisher's Scoring optimisation technique was also employed. The response category *coping* = 0 (i.e., household did not adopt a coping strategy during the crisis) was taken as the reference category. This means that the probability modelled was *coping = 1* (i.e., household adopted a coping strategy). Note that in sampling, 69 counties were considered as strata and that the sampled households were selected from 150 clusters nested within strata.

In the first step results of type 3 analysis of effects based on the logit link function are presented in Table 4.9, which shows that four of the seven fixed effects were significant factors. Like in the Binary Logistic Regression model, the Survey Logistic Regression determined *fishing* as a non-significant factor. The method also showed *crop cultivation* and *livestock ownership* as not significant. This means that whether or not household fished, cultivated crops or kept

livestock three months before the survey, their coping levels remained the same. This could point to the chronic nature of the crises in South Sudan and generalised asset poverty.

Table 4.9: Type 3 analysis of effects for the Cumulative Logit Model

Effect (X_{ij}) *	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
Gender of household head (x_{2j})	1	10.6396	0.0011
Age of household head (x_{1j})	2	19.2414	<0.0001
Household size (x_{3j})	1	9.2707	0.0023
Cultivated crops (x_{4j})	1	0.7086	0.3999
Livestock (x_{5j})	1	2.0147	0.1558
Fishing (x_{6j})	1	0.0041	0.9487
Livelihood source (x_{7j})	4	36.5083	<0.0001

* Last category level is selected as reference

In the second step, we present the parameter estimates for the main effects model based on the maximum likelihood and the related Wald test of hypothesis (Table 4.10). The model fit statistics and prediction of model accuracy power. As in explained earlier, the negative coefficients of the parameters estimates indicate that the reference categories are associated with the probability of the response variable, which in our case is coping with food insecurity crisis. A positive coefficient suggests that the shown category associated with the modelled probability of the response. The positive coefficient of the scale variable household size means that a household with more than average number of members associated with coping. In this case, the more the number of people living in a household, the greater the probability of adopting a coping strategy during the crisis. Thus, the results showed that a household headed by a female aged 60 years and above, with above average number of people living in it, and depended on sale of crops, sale of livestock and employed labour, was associated with the probability of coping in the crisis.

Table 4.9: Analysis of maximum likelihood estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.0568	0.9731	4.4676	0.0345
Gender: *Male*	1	-0.2577	0.0790	10.6396	0.0011
Age: *<18*	1	-1.1267	0.3621	9.6848	0.0019
18-60	1	-0.7314	0.2001	13.3546	0.0003
Household Size	1	0.0482	0.0158	9.2707	0.0023
Cultivated crops: *yes*	1	0.0754	0.0896	0.7086	0.3999
Owned livestock: *yes*	1	-1.3632	0.9604	2.0147	0.1558
Fishing: *Yes*	1	0.00606	0.0942	0.0041	0.9487
Main Income: *Agriculture*	1	-0.3758	0.1686	4.9694	0.0258
Livestock	1	0.5301	0.1812	8.5569	0.0034
Employment	1	-0.4326	0.1790	5.8437	0.0156
Petty trade	1	-0.0603	0.1629	0.1369	0.7114

HH=Household; DF=Degrees of freedom; Pr=Probability; ChiSq=Chi-square

Table 4.10 below show the odds ratio estimates comparing reference categories of the variables included in the model. Only the results of significant effects are displayed. An odds ratio value less than one indicate that the reference category was better in the probability of adopting a coping strategy. Conversely, an odds ratio value above one means that the displayed category associated with coping compared to the reference category. For example, the odds of household headed by a male were 0.77 times those of female headed households in adopting some form of coping strategy.

Table 4.10: Odds Ratio estimates for significant effects

Effect	Point Estimate	95% Wald Confidence Limits	
Gender of household head: *male vs female*	0.773	0.662	0.902
Age of household head: *<18 vs >60 years*	0.324	0.159	0.659
18-60 vs >60 years	0.481	0.325	0.712
Household size	1.049	1.017	1.083
Main income Source: *sale of crops vs others*	0.687	0.494	0.956
Sale of livestock products vs others	1.699	1.191	2.424
Employment vs others	0.649	0.457	0.921
Petty trade vs others	0.942	0.684	1.296

After fitting the survey logistic model, we presented some diagnostic statistics tests using the Likelihood Ratio (LR) test, the Akaike Information Criteria (AIC) test, and the Score Criteria (SC) selection criteria, which are presented in Table 4.11. Although these are basically statistics used for model selection, they can also be used as approximations for comparing two or more competing models.

Table 4.11: Tests of Global Null Hypothesis $\beta = 0$

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	130821.584	11	<0.0001
Score	123643.467	11	<0.0001
Wald	206.8663	11	<0.0001

For the AIC, the model with the lowest value is generally selected, especially when the objective of the study is to check the consistency of the model. A model based on the SC is selected if the interest is on the order of the model (Burnham & Anderson 2002). These tests are also called global tests of the assumption $\beta = 0$. They show highly significant probabilities

(p <0.0001), indicating that the Binary Logit model might not adequately fit the data. This could be a natural result of limitations with the raw data as noted earlier. However, the concordant test of association of predicted *versus* observed responses shown in Table 4.12 indicates that over 60 *per cent* of the probability of 'coping' was predicted correctly.

Table 4.12: Association of predicted probabilities and observed responses

Percent Concordant	60.6	**Somers' D**	0.224
Percent Discordant	38.3	**Gamma**	0.226
Percent Tied	1.1	**Tau-a**	0.112
Pairs	3242376	**C**	0.612

Also shown in Table 4.12 is the concordance index (c), which is equivalent to the area under the curve (ROC curve). The concordance index checks the model predictive accuracy power. Interpretation of this index is the same as for the area under the ROC curve. That is to say that as the value of the concordance index approaches 1, the better the model predictive accuracy power. This finding indicates fairly good prediction accuracy by the Survey Logistic model.

4.5.3. Discussion

To perform estimation, the Survey Logistic procedure modifies the standard likelihood equations in order to cater for the case of weighted observations. Thus the method called pseudo-maximum likelihood estimation (PML) is used instead such that the clustering effects of the model are properly accounted for (Skinner et al. 1989; Roberts et al. 1987b; G. Morel 1989; Chambers & Skinner 2003).

This study sets to answer one fundamental question, "is there strong statistical evidence to suggest that the seven explanatory variables analysed in the model are predictors of household

food consumption based on the survey sample?" Using the sampling weights and the generalized logit link function, the survey logistic procedure determined all seven variables as predictors of coping with food insecurity. Comparing the model with generalised logit link function and one with cumulative function both methods yielded almost similar coefficient estimates of the Wald test.

As explained in Sub-section 4.5.2 above, the negative coefficients in Table 4.9 corresponding to three factors determined as significant effects, that is, age, gender and main source of income, however, indicate that the odds of these factors are in favour of the reference categories. Positive coefficient estimates, to the contrary, point to the odds being in favour of the corresponding category levels of the factors. It was clear that crop cultivation, owning livestock and fishing – three major possible sources of livelihoods in developing countries and rural populations, had no statistical evidence of significant associations with coping in food insecurity crisis. It could not be established that these sources of livelihood were potential causes of population resilience during the crisis.

Of concern though, is the finding that the global tests of the assumption $\beta = 0$ (the null hypothesis), which is highly significant and indicating the model was a poor fit to the data. Thus, arguments for an alternative analysis such as those advanced by Williams (2006) become relevant. He establishes a model (the Generalized Ordered Logit) that partially estimates the proportional odds assumption, which is quite often violated. He argues that this partial proportional odds model (or *gologit*) is "more parsimonious and interpretable than those estimated by non-ordinal method, such as multinomial logistic regression. The down side to this approach is that model convergence becomes an issue in presence of factors with missing data as is the case with survey data. This was the main constraint for not using the *gologit* technique.

4.5.4. Conclusion

Data analysis explored featuring the Survey Logistic Model led to the conclusion that the technique was appropriate for analysis of the type of data explored. Although with some limitation of accuracy, the method could be applied for analysis of similar data and for similar purposes. With more improved or controlled study, the results of such study may provide useful evidence for crises response and disaster recovery interventions targeting populations in distressful situations. Good data that have no or insignificantly few missing case, can guarantee the power of the model.

Furthermore, it is important to note that the best strategy to buffer against humanitarian disaster risks is how much a population is able to withstand their serious impact. Therefore, the results give cause for reflection on what and where to prioritize for improving resilience of populations.

Finally, the technique could help governments in targeting areas identified to be at higher risk of food insecurity shocks with a set of resilience building interventions such as rural development programmes, social protections, among many other possible options. It is recommended that similar food security surveys as the one on which we have based our analysis, should include collection of geographical information coordinates in order to enable spatial analysis.

CHAPTER 5

Generalized Linear Mixed Models for Analysis of Ordered Categorical Data

5.1. Introduction

In Chapter 4 analysis for determining a measure of resilience to food insecurity was based on univariate logistic regression. Random effects in the data were accounted for in the model using the Complex Samples Logistic Regression (IBM Corporation 2015) and Survey Logistic Model (SAS Institute 2011a). The survey logistic procedure has the property of accounting for complex survey designs. In this chapter, however, attempts are made to examine how mixed effects (random and fixed) influence the results. The chapter explores the use of Generalized Linear Mixed Models (GLMMs) in assessing the risk of food insecurity shocks in distressful livelihood situations. The GLMM procedure was fitted to identify a set of predictors of the likelihood of food insecurity risk – measured using food consumption outcomes.

The chapter is divided into six sections. Section 5.2 attempts to heighten the importance of measures aimed at determining predictors of food insecurity risk outcomes, as well as the need for a statistically established composite measure for assessing the likelihood of food insecurity, especially in protracted crisis situations. Section 5.3 gives highlights on the data used and described the outcome variable. Section 5.4 presents an overview of the GLMM method fitted for ordered categorical data with the outcome variable being *Food Consumption Score* to explore both the validity and strength of the model. Section 5.5 presents the results of the analysis. Section 5.6 entails discussions of the results and draws summary and conclusions.

5.2. Importance of Measuring and Assessing Food Insecurity Risk

Populations in conflict situations and protracted emergencies suffer the most from food insecurity vulnerabilities, as their resilience is tremendously weakened and their asset base is depleted, forcing them to resort to extreme coping mechanisms. Protracted crisis plunge populations into extreme poverty and chronic food insecurity. Food security resilience gets severely corroded, as households get entrenched in subsistence economy. Protracted crisis is described as situations in which crises are prolonged and recurrent. Among the manifestations of protracted crisis are disruption of livelihoods and food systems, which result in increasing rates in morbidity and mortality, as well as increased displacements (Committee on World Food Security 2015). In these situations large numbers of people or entire communities are displaced and affected by food and malnutrition, thus often require enormous amount of resources and relief interventions. This description typifies South Sudan, especially in the two years prior to this study.

Severe food insecurity causes anxiety, which in turn causes desperation, which in turn forces people to resort to extreme forms of coping strategies. In situations where firearms are rampant, like in South Sudan, extreme coping strategies might be in the form of banditry, armed robbery and rustling of cattle – a practice existing amongst pastoralist communities of South Sudan. It is on this grounds that recent recommendations for offering long term solutions for cutting hunger and malnourishment to a bare minimum, called for measures which result in increased food and agricultural productivity with social protection (African Union Commission 2013). This move is ostensibly the alternative to the first Millennium Development Goal, which spelled out the need to "cut extreme poverty and hunger by 2015" (United Nations 2015).

As the continent's population is predominantly dependent on agriculture, including its subsectors of livestock, fisheries, forestry and natural resources, it makes more sense that the

sector is enabled to boost social protection and *vice versa*. For this reason, recent consultations amongst Africa's agricultural development and food security stakeholders a study on strengthening the coherence between social protection and agriculture (Slater et al. 2015; FAO 2003). The paper by Slater et al. (2015) recommends support to programmes that include safety nets and a two-tract approach that combines promoting rural and agricultural growth as a measure to protect those who cannot produce food themselves.

Furthermore, within the context of the Comprehensive Africa Agriculture Development Programme (New Partnership for Africa's Development 2003), the Framework for Africa's Food Security under its objective "Increased economic opportunities for the vulnerable", recommends a set of medium and long-term options for improving resilience of the vulnerable. Such durable resilience enhancing developmental options augment and improve on the framework's other objectives of improved risk management, increased supply of affordable commodities and increased quality of diets among target groups (New Partnership for Africa's Development 2009). Indeed, the focus on durable and forward-looking options to build resilience of the vulnerable seems to feature prominently more than ever before in contemporary food security and nutrition frameworks.

As the recommendations and plans for integrating socioeconomic and rural development objectives with humanitarian efforts to mitigate vulnerability and strengthen resilience of vulnerable population are gaining momentum, the need for producing evidence for monitoring the state of resilience of populations in distressful food insecurity situations, equality gain interest. Current measures based on periodically conducted household surveys are still centred on determining vulnerability for the purpose of relief and rehabilitation, rather than for boosting resilience and prevent future vulnerabilities and the devastating after-shock effects. In other

words, there is need to establish measures for determining the probability of future risk, which resilience-based measures offer.

The purpose of the study is, therefore, to find statistically robust and efficient measures that identify the set of factors that determine the risk to food insecurity. Measures for determining resilience seem to provide the answer to this question.

5.3. Sample and Data

The data explored are from the Food Security Monitoring Survey (FSMS) as described in Chapter 2. The study models the relationship between a set of fixed effects mainly representing livelihood capitals and household characteristics and the outcome variable *Food Consumption Score* (*FCS*). *FCS*, which is in the form of ordered polytomous categories, was calculated as described in Lokosang et al. (2010, pp.108–109) and the World Food Programme, Vulnerability Analysis and Mapping Branch (2008). As defined in Section 2.3, let *FCS* be Y_{ij}, where $i = 1, \cdots, n$ and $j = 1, 2, 3,$ such that

$$Y_{ij} = \begin{cases} 1 & \text{if food security level is } 'poor' \\ 2 & \text{if food security level is } 'borderline' \\ 3 & \text{if food security level is } 'acceptable' \end{cases}$$

Since the aim of the study was to determine how the explanatory variables contributed to household food insecurity, the categories '*poor*' and '*borderline*' food consumption score were together taken to be the reference category; henceforth referred to as '*worse*' food consumption level. The *per cent* distribution of the response (or outcome) variable *food consumption score* is as shown in Table 5.1 below.

Table 5.1: Profile of Food Consumption Scores

Level	FCS Category	Frequency	%
1	Poor	430	11.6
2	Borderline	1053	28.5
3	Acceptable	2209	59.8

Also see Section 2.3 for explanation of the dependent variable *food consumption score*.

Seven independent variables were selected for analysis. These are attributes of the household head (*age* and *sex*), the household (*household size*), sources of livelihood (*crop cultivation*, *ownership of livestock* and *fishing*) and *main sources of income* (sale of crops, sale of livestock products, employment and other sources). In general, let X_{ij} be the effect of the i^{th} household on *FCS*, where $i = 1, \dots, n, j = 1, 2 \; or \; 3$ and n is the sample size.

The selection of limited number of variables was due to two factors. First, the data came from a secondary source that contained only variables included for the prime purpose of the survey – repeated food security monitoring during protracted humanitarian crisis. Second, the seven variables qualify as possible factors influencing the outcome variable – food consumption score.

Gender of household head is considered important in the sense that food access in the household may depend on the level of economic activity of the person providing the food. Common perception could assume that households headed by males could fare better given that males are dominant in formal employment and do harder jobs than females. However, this might not be the case in pastoralist communities, where men habitually spend most of the day looking after their stock of cattle, leaving women to toil with domestic workload, including working in home gardens and looking for food.

It is also expected that households might differ in the level of food consumption according to age of household head. Households headed by younger adults (such as in settings where conflicts or the HIV pandemic) are expected to have worse food consumption scores than households headed by more mature adults. The latter are undoubtedly supposed to be economically active, are more aware of their surroundings, and better educated. Similarly, household size may influence household food consumption in that the number of people in the house might either be an advantage (such as in the case where the household has more working adults) or a disadvantage. In the latter case, larger households might either be dominated by non-working dependants, while the economically active head earns less income to provide adequate food for the household, or the available food quickly does not last for long; thus creating food insecurity due to non-durability.

Whether a household cultivated crops, or owned livestock and practiced fishing in the past three months prior to data collection, are conceptually important determinants of the level of food consumption. A household is naturally expected to fare better when it is involved in some economic activity or has some livelihood asset to depend on. It is expected that a household depending on all three sources of livelihood have better food consumption and dietary diversity and thus improved nutrition status than those that depend on one source of daily diet.

Main source of income is conceptually related to food consumption. Households could differ in their levels of food consumption according to how they mainly derive their income. In a setting where food is mainly purchased from markets, such as in more urbane populations, a household which mainly derive its income from wages and salaries or business enterprise, could be at an advantage than those dependent on other sources. Similarly, in rural settings, where households largely depend on income from selling the product of their gardens or

livestock, such households might be in better food consumption status than those not depending on such source.

5.4. Generalized Linear Mixed Model for Ordered Categorical Data

Considering the structure of the data, the outcome variable and the purpose of the study, the Generalized Linear Mixed Models (GLMMs) procedure was thought to be appropriate. GLMMs are an extension of Generalized Linear Models (GLMs) in such a way that: a) the target is linearly related to the factors and covariates via a specified link function; b) the target can have a non-normal distribution; and c) the observations can be correlated (McCullagh & J. A. Nelder 1989; Agresti 2002) or have non-constant variability

GLMMs fit statistical models to data with correlations or non-constant variability and where the response is not necessarily normally distributed. Like linear mixed models, GLMMs assume normal (Gaussian) random effects. Conditional on these random effects, data can have any distribution in the exponential family (McCulloch & Searle 2001). Where random effects are absent, the GLMM fits generalized linear models. In incorporating random effects in the model, the GLIMMIX procedure allows for subject-specific (conditional) and population-averaged (marginal) inference.

There are a variety of GLMM applications in biological science, medicine, psychology, business and marketing, etc. GLMM are particularly useful for data with correlations among some or all observations. Such correlations arise as a result of repeated observations of the same sampling units, shared random effects, spatial proximity, multivariate observations, etc.

GLMMs are fitted to the data with three underlying assumptions:

a) For a model with random effects, the distribution of the data conditional on the random effects is known. This distribution is either a member of the exponential family of distributions or one of the supplementary distributions. In models without random effects, the unconditional or marginal distribution is assumed to be known for maximum likelihood estimation, or the first two moments are known in the case of quasi-likelihood estimation.
b) The conditional expected value of the data takes the form of a linear mixed model after a monotonic transformation is applied.
c) The problem of fitting the GLMM can be cast as a singly or doubly iterative optimization problem. The objective function for the optimization is a function of either the actual log likelihood, an approximation to the log likelihood, or the log likelihood of an approximated model.

For a model containing random effects, GLMM estimates the parameters by applying pseudo-likelihood techniques (Wolfinger & O'Connell 1993; Breslow & Clayton 1993). In a model without random effects (i.e., GLMs), the model parameters are estimated using the maximum likelihood, restricted maximum likelihood, or quasi-likelihood. Statistical inferences for the fixed effects and covariance parameters of the model are then performed once the parameters have been estimated. Tests of hypotheses for the fixed effects are based on Wald-type tests and the estimated variance-covariance matrix.

GLMMs are known to account for the complexity of survey designs and handle within-block correlations or similarity of responses (McCullagh & J. A. Nelder 1989; McCullagh & J.A. Nelder 1989). Other advantages GLMM lie in that, being linearization-based method, it includes a relatively simple form of the linearized model that can typically be fit based on only the mean and variance in the linearized form. The method suits well a model with correlated

errors, a large number of random effects, crossed random effects and multiple types of subjects. Furthermore, because the structure of the model fitted at each stage is a linear mixed model, Restricted Maximum Likelihood (REML) estimation is possible. The disadvantages of GLMM include the absence of a true objective function for the overall optimization process and thus might yield potentially biased estimates of the covariance parameters, especially for binary data. The objective function to be optimized after each linearization update is dependent on the current pseudo-data. Therefore, such process can fail at both levels of the double iteration scheme (Schabenberger 2004).

The formulation of GLMM is as follows:

Suppose $\mathbf{Y}$ represents the $(\mathbf{n} \times \mathbf{1})$ vector of observed data and $\boldsymbol{\gamma}$ is a $(\mathbf{r} - \mathbf{1})$ vector of random effects. In fitting a GLMM, it is assumed that

$$\mathbf{E}[\mathbf{Y}|\boldsymbol{\gamma}] = \mathbf{g}^{-\mathbf{1}}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma})$$

where $\mathbf{g}(.)$ is the differentiable monotone link function and $\mathbf{g}^{-\mathbf{1}}(.)$ is the inverse. The matrix $\mathbf{X}$ is an $(\mathbf{n} \times \mathbf{1})$ matrix of rank k and $\mathbf{Z}$ is an $(\mathbf{n} \times \mathbf{1})$ design matrix for the random effects, which are assumed to be normally distributed with mean $\mathbf{0}$ and variance matrix $\mathbf{G}$.

The GLMM contains a linear mixed model inside the inverse link function. This model component is referred to as the linear predictor,

$$\mathbf{E}[\mathbf{Y}|\boldsymbol{\gamma}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma}$$

The variance of the observations, conditional on the random effects, is

$$\mathbf{Var}(\mathbf{Y}|\boldsymbol{\gamma}) = \mathbf{A}^{\mathbf{1/2}}\mathbf{R}\mathbf{A}^{\mathbf{1/2}}$$

The matrix **A** is a diagonal matrix and contains the variance of the model. The variance function expresses the variance of a response as a function of the mean. The matrix **R** is a variance matrix.

If the conditional distribution of the data contains an additional scale parameter, it is either part of the variance functions or part of the **R** matrix. For a gamma distribution with mean $\boldsymbol{\mu}$ the variance function $a(\boldsymbol{\mu}) = \boldsymbol{\mu^2}$ and $\boldsymbol{Var}[\mathbf{Y}|\boldsymbol{\gamma}] = \boldsymbol{\mu^2}\emptyset$. If the model calls for G-side random effects only as described below, then $\mathbf{R} = \emptyset\mathbf{I}$ is modelled, where the **I** is the identity matrix. For a binary distribution with mean $\boldsymbol{\mu}$ the variance function $\boldsymbol{a}(\boldsymbol{\mu}) = \boldsymbol{\mu}(\mathbf{1} - \boldsymbol{\mu})$.

GLMM distinguishes two types of random effects: the "G-side" and the "R-side", depending on whether the parameters of the covariance structure for random components in the model are contained in **G** or in **R**. Similarly, the associated covariance structures of **G** and **R** are known as the G-side and R-side covariance structure, respectively. R-side effects are also called "residual" effects. Simply put, if a random effect is an element of $\boldsymbol{\gamma}$, it is a G-side effect; thus the fitted model is the G-side covariance structure. Otherwise, it is an R-side covariance structure model. If the model has no G-side effects it is termed a “marginal model”. Fitted models can have none, one, or more of each type of effect.

Note that in fitting the model, the **R** matrix is by default the scaled identity matrix, $\mathbf{R} = \boldsymbol{\phi I}$. The scale parameter $\boldsymbol{\phi}$ is set to one if the distribution does not have a scale parameter. This includes the binary, binomial, Poisson, and exponential distributions.

Unknown quantities subject to estimation are the fixed-effects parameter vector $\boldsymbol{\beta}$ and the covariance parameter vector $\boldsymbol{\theta}$ that comprises all unknowns in **G** and **R**. As the random effects $\boldsymbol{\gamma}$ are not estimated, they are not considered model parameters. The vector $\boldsymbol{\gamma}$ is a vector of random variables and their solutions are predictors of these random variables.

We used a two-step approach for modelling the data. In the first step, it was assumed that the data came from randomly selected counties and then randomly selected clusters and, therefore, there were random effects. It then necessitated fitting a GLMM model with random effects. Note that the sample data were from 150 randomly sampled clusters (groups of villages). It is possible that there were cluster effects in the data, considering that South Sudan has different livelihood or agro-ecological zones with different food systems. For instance, clusters in the River Nile livelihood zone could be characterised with fishing communities, and thus could influence the responses to the question "Was any household member engaged in fishing in the past 3 months?" Similarly, responses to the question "Does your household own any livestock, herds or farm animals?" could be influenced by whether or not a household is in a cluster located in clusters with predominantly pastoralist communities. Once the first fitted model showed no evidence of between-cluster variability, a second model was fitted with fixed effects only.

The adaptive Gauss-Hermite quadrature method was used, basically because it restricts the models for estimating parameters and also fulfils conditional independence assumptions and the processing of data by subject. The seven fixed effects were included in the models with multinomial distribution and cumulative logit link function in order to allow transforming the outcome variable into an approximately linear variable with normally distributed errors. This holds true as the outcome variable was nominal. The first model was fitted with fixed effects and random effects, while the second one did not include random effects to generate estimates of the fixed effects when there was no statistical evidence of between cluster variability. Odds ratios were requested. Finally, the standard variance components structure was specified for estimating the model. According to Ene et al. (2014) the variance components structure is the most simple variance-covariance structure.

In models with a *logit*, *generalized logit*, or *cumulative logit link*, estimates of odds ratios are obtained through the linear predictor of the dichotomous, ordinal or nominal outcome and an appropriate link function. For a model with a dichotomous outcome and *logit* link function,

$$\boldsymbol{\eta} = \mathbf{X}'\boldsymbol{\beta} + \mathbf{Z}'\boldsymbol{\gamma}$$

Suppose that $\boldsymbol{\eta_0}$ represents the linear predictor for a condition of interest. For example, in a simple logistic regression model with $= \alpha + \beta \mathrm{X}$, η_0 might correspond to the linear predictor at a particular value of the covariate, say, $\eta_0 = \alpha + \beta X_0$. The model probability is $\pi = 1/(1 + exp\{-\eta\})$ and the odds for $\eta = \eta_0$ are

$$\frac{\pi_0}{1-\pi_0} = \frac{1/(1+\exp\{-\eta_0\})}{\exp\{-\eta_0\}/(1+\exp\{-\eta_0\})} = \exp\{\eta_0\}$$

Because η_0 is a *logit*, it represents the log odds. The odds ratio $\psi(\eta_1, \eta_0)$ is defined as the ratio of odds for η_1 and η_0, $\psi(\eta_1, \eta_0 = \exp\{\eta_1 - \eta_0\}$

The odds ratio compares the odds of the outcome under the condition expressed by η_1 to the odds under the condition expressed by η_0. This ratio equals $\exp\{\beta(x_1 - x_0)\}$ in the preceding simple logistic regression example. The exponentiation of the estimate of β is thus an estimate of the odds ratio comparing conditions for which $x_1 - x_0 = 1$. If x and $x + 1$ represent standard and experimental conditions, for example, $\exp\{\beta\}$ compares the odds of the outcome under the experimental condition to the odds under the standard condition. For many other types of models, odds ratios can be expressed as simple functions of parameter estimates. To fit a logistic model with a single effect with *k* levels, the estimated linear predictor for level 1 of the effect, say A, is

$$\hat{\eta}_j = \hat{\beta} + \hat{\alpha}_j, \qquad j = 1,2,\cdots,k$$

Because the **X** matrix is singular in this type of model due to the presence of an overall intercept, the solution for the intercept estimates $\beta + \alpha_k$, and the solution for the j^{th} effect estimates $\alpha_j - \alpha_k$. Exponentiating the solutions for $\alpha_1, \alpha_2, \cdots, \alpha_{k-1}$ produces odds ratios comparing the odds for these levels against the k^{th} level of A.

The computations of odds ratios rely on general estimable functions which are based on least squares means. This enables obtaining odds ratio estimates in more complicated models that involve main effects and interactions, including interactions between continuous and classification variables.

5.5. Results

A Generalized Linear Mixed Model with cluster random effects was fitted to the data with random effects due to clusters. Two models were fitted based on the Maximum Likelihood with adaptive Gauss-Hermite quadrature and standard variance-covariance structure. The first model was fitted using the cumulative logit link function and the second with cumulative probit link function. The tests yielded estimated variances of cluster effects for each of the models as shown in Table 5.2.

Table 5.2: Covariance Parameter Estimates for Cumulative Logit Model and Cumulative Probit Model

Model	Covariance Parameter	Subject	Estimate	Standard Error
Cumulative Logit	Intercept	Cluster number	1.5444	0.2246
Cumulative probit	Intercept	Cluster number	0.5089	0.07262

Results of Table 5.2 showed that the Cumulative Logit model with an estimated covariance parameter (cluster effect) of 1.5444 and an estimated asymptotic standard error of 0.2246

showed significant cluster to cluster effects. For a model with the same maximum likelihood estimation and cumulative probit, the estimated variance of cluster effect is 0.5089 and estimated asymptotic standard error of 0.07262. Therefore, the model with cumulative link function was preferred to the one with probit link function. We then proceeded to examine results of tests of relationship between fixed effects and the *FCS*.

Table 5.3: Type 3 tests of fixed effects for the Cumulative Logit Model

Effect	Num DF	Den DF	F Value	Pr > F
Gender of household head	1	3440	39.36	<0.0001
Age of household head	2	3440	2.95	0.0527
Household size	1	3440	9.27	0.0024
Cultivated crops	1	3440	31.77	<0.0001
Owned livestock	1	3440	1.96	0.1615
Practiced fishing	1	3440	34.99	<0.0001
Main source of income	4	3440	4.92	0.0006

Den DF/Den DF = Numerator/denominator degrees of freedom

Table 5.3 presents the Type 3 tests of fixed effects based on the cumulative logit. The method determined how each of the seven independent variables associated with *FCS*. Coefficients of six of the seven fixed effects were determined to be very highly associated with food consumption, as they showed low *p*-values. The methods determined that *Ownership of livestock* had non-significant association with *FCS*. That is, it did not contribute significantly to the model under the 0.05 level of significance test of hypothesis (Fisher's test). In this test, the 'null' hypothesis states that the model with fixed effects is equal to the model without any fitted effects and that there is no difference between levels of the fixed effects, as they associated equally with an outcome variable. Generally, Table 5.3 shows that *gender* and *age of household head, size of household, crop cultivation, fishing* and *main source of income,* had significant relationships with *FCS* and thus important effects of *FCS*.

Since parameter estimation is based on maximum likelihood, it is possible to conduct a formal test of hypothesis of no cluster variability.

As shown in the intercept probabilities in Table 5.4, between-cluster variability gives a significant test. The category cut-offs for the cumulative probabilities are -0.9829 and 1.1298. This suggests that most of the variability in food consumption scores that is not explained by the fixed effects can be explained by cluster-to-cluster variation.

Table 5.4: Solution for fixed effects for the Gauss-Hermite Quadrature Likelihood Approximation method

Effect	Estimate	Standard Error	DF	T Value	Pr > |t|
Intercept: *poor FCS*	-0.9829	0.5373	149	-1.83	0.0694
Intercept: *borderline poor FCS*	1.1298	0.5381	149	2.10	0.0374
Sex of household head: *male*	-0.5261	0.08392	3529	-6.27	<0.0001
Age of household head: *<18*	0.08429	0.3874	3529	0.22	0.8278
Age of household head: *18-60*	-0.3898	0.2122	3529	-1.84	0.0663
Size of household	-0.1517	0.05650	3529	-2.68	0.0073
Cultivated crops: *yes*	-0.6119	0.1017	3529	-6.02	<0.0001
Owned livestock: *yes*	0.5991	0.4716	3529	1.27	0.2041
Engaged in fishing: *yes*	-0.8787	0.1463	3529	-6.01	<0.0001
Main income source: *agriculture*	-0.5278	0.1617	3529	-3.26	0.0011
Main income source: *livestock*	-0.7432	0.1701	3529	-4.37	<0.0001
Main income source: *employment*	-0.3746	0.1672	3529	-2.24	0.0251
Main income source: *petty trade*	-0.4833	0.1660	3529	-2.91	0.0036

Further analysis involved fitting a GLMM to the multinomial distributed data using cumulative logit model. Recall that in Section 5.3 the two categories of the outcome variable *FCS* '*borderline*' (or '*near-poor*') food consumption and '*poor*' food consumption, were together referred to as '*worse*' food consumption, in order to make meaningful interpretation of the

results. This is true since interest was centred on finding out which fixed effects were associated with food insecurity risk. Note that ordered categorical data were modelled using the cumulative logit link function. Hence, it was possible to compare this cumulated category with the '*acceptable*' food consumption category, which was constrained to be the reference category.

If particular interest is centred on knowing which particular levels of these fixed effects were associated with the likelihood of scoring '*worse*' food consumption levels, Table 5.4 provides the answer. It is worth noting that in this analysis the last level of each fixed effect was taken to be the reference level (i.e., assigned the value '0') for between-level comparison of the fixed effects. All fixed effects corresponding to significant values (i.e., *p*-value <0.05) showed negative coefficients of estimates, meaning that the reference categories of the corresponding fixed effects associated with the '*worse*' category of *food consumption score* (*FCS*). Conversely, had these fixed effects showed positive parameter estimates (i.e., the $\hat{\beta}$'s), it would have meant that their corresponding categories associated with the '*worse*' category of *FCS*.

Table 5.4 shows that except for livestock ownership, all fixed effects contrasts were significant. In general, it could be stated that a household headed by a female, aged above 60 years, with six or more members, did not cultivate crops, did not engage in fishing, and did not earn income from sale of crop, livestock products, employment and petty trade, associated with the '*worse*' category of food consumption score. Probably what could be interesting is that ownership of livestock showed non-significant association with the '*worse*' category of *FCS*. This means that owning or not owning livestock did not make affect coping with food insecurity. Livestock was owned by a large section of South Sudanese communities and in the sample by an overwhelming proportion of households in the sample (97%). However, it seemed that livestock was not kept for household food, but for cultural purposes such as marriage, payment

of bail or of ransom, and generally as a form of social insurance. Nevertheless, sale of livestock improved food consumption score. It is also possible that displacement caused affected households to leave their herds in their original places of domicile. It is to be recalled that the survey from which the data analysed was based, was conducted during conflict which, according to the World Food Programme (2014), caused significant population displacement.

In order to determine the relationships between each of the fixed effects in the model and the likelihood of being in the '*worse*' *FCS* category, it was important to examine the parameter estimates for each fixed effect in terms of odds ratios. Odds ratio values allow comparisons of between-level relationships in terms of associations of the fixed effects with the '*worse*' *FCS* levels. The calculated odds ratio values are shown in the last column of Table 5.5. An odds ratio value greater than unity (i.e., $\theta > 1$) indicates that the given level of a fixed effect associated more with the '*worse*' *FCS* score, compared to the reference category of that effect. The bigger the odds ratio value, the more significant is the relationship between levels of the fixed effect and the '*worse*' category of *FCS*. Odds ratio values greater than unity reflect positive coefficient estimates and mean that advantage in the comparison goes to the shown category of the fixed effect. For example, the odds of a household that *cultivated crops* having '*worse*' *FCS* were about half of those of a household that did not farm crops in the given period.

In contrast, an odds ratio value less than 1 indicates that the corresponding effect has less odds of being in the '*worse*' *FCS* category compared to the reference category. For example, the odds were fewer for a male headed household than a female headed household in facing the risk of food insecurity. The implication is that a female headed household associated with scoring '*worse*' *FCS* level. The odds of a household that cultivated crops being at risk of food insecurity were about 0.5 times (half) those of a household did not cultivate crops. This means that not cultivating crops exposed households to food insecurity risk.

Table 5.5: Odds ratio estimates for comparing between levels of fixed effects

Comparison	Estimate	DF	95% Confidence Limits	
Gender of household head: *male vs female*	0.588	3440	0.498	0.694
Age of Household head: *<18 vs >60*	1.375	3440	0.634	2.983
18-60 vs >60	0.721	3440	0.472	1.101
Size of household: *unit change from mean*	0.954	3440	0.925	0.983
Cultivated crops: *yes vs no*	0.558	3440	0.456	0.684
Ownership of livestock: *yes vs no*	1.992	3440	0.759	5.228
Involved in fishing: *yes vs no*	0.413	3440	0.308	0.554
Main source of income: *agriculture vs other*	0.59	3529	0.43	0.81
livestock vs other	0.476	3529	0.341	0.664
Employment vs other	0.688	3529	0.495	0.954
Petty trade vs other	0.617	3529	0.445	0.854

Furthermore, it was a good finding that income from agriculture, sale of livestock, wages and petty trade made households to improve on food consumption. Not having any main source of income in a crisis, would render populations to suffer from a generalised weakened resilience to food insecurity strains. Income improved the livelihoods of households dependent on other forms of traditional rural livelihoods such as subsistence crop farming, livestock keeping, fishing, and gathering of forest products. Crises such as conflicts, flooding, typhoons and droughts, which cause large sections of populations to migrate, can expose those highly dependent on traditional livelihoods to facing vulnerability to famine and extreme poverty.

Much as GLMMs are useful for determining the relationship between fixed effects and an outcome variable, establishing whether or not random effects have significant effect on the model, they also come with notable limitations of GLMMs, such as in the linear function that

have only a linear predictor in the systematic component and that their responses must be independent. Bolker et al. (2008, p.133) caution that optimum care should be exercised in using GLMMs, especially with regard to limitations in data. Schabenberger (2005) also report some shortcomings with regards to computation of predicted values. Some of the noteworthy shortcomings of fitting a GLMM lie with ordered categorical data as discussed in Lokosang et al. (2010, p.123), citing the World Food Programme (2008). Typical of these shortcomings lie in misclassification of predicted values of the outcome variable.

Since the prime aim of the study was to generate predicted values of food insecurity risk, it was seen prudent to examine the estimated predicted values. Bayes estimates of the cluster effects from Maximum Likelihood (ML) estimation by quadrature, along with prediction standard error bars, were determined. The Empirical Bayes Estimation criterion estimates the random effects (see Figure 5.1).

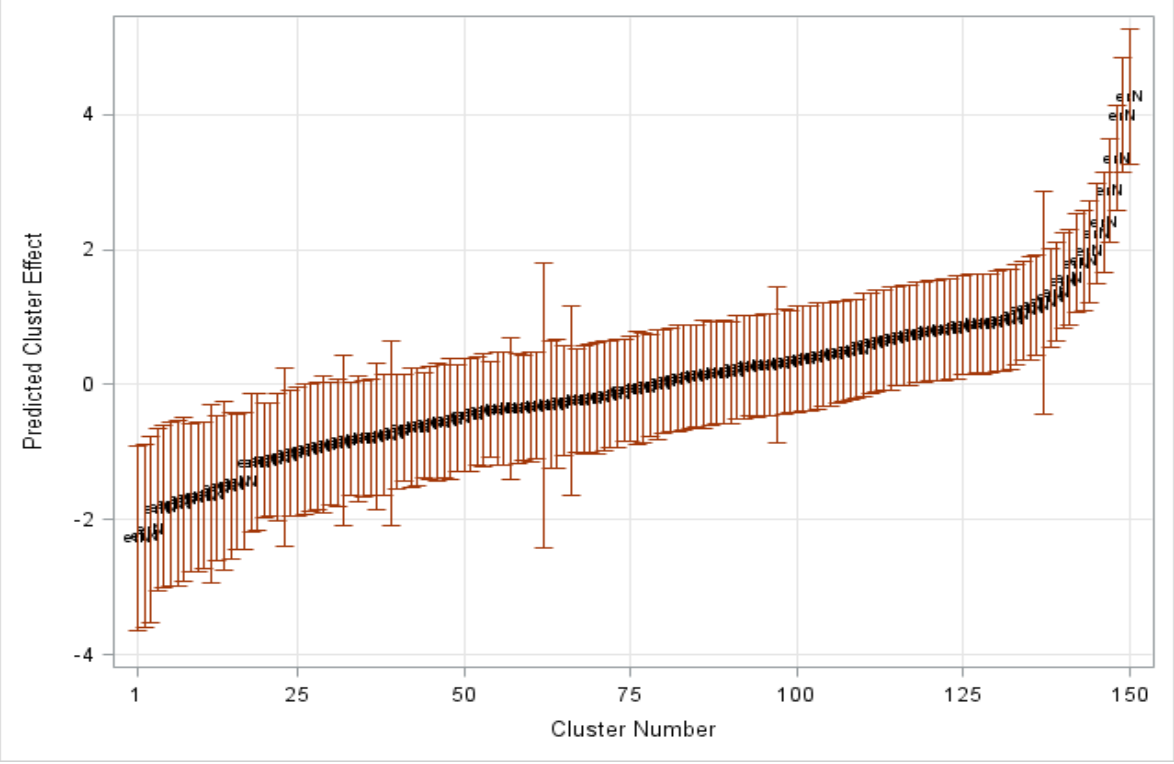

Figure 5.1 Predicted cluster effects and prediction standard errors

Figure 5.1 shows that there is significant evidence that the model with cluster random effects fits the data as well. Therefore, it gives reason to generate predicted values from the fitted model, which could then serve as an index for determining resilience to food insecurity risk.

5.6. Discussion

Analysis was conducted by fitting a model that tested for randomness and examined whether there were cluster effects. Random effects due to between-cluster effects were determined to be significant. Two GLMMs, one with cumulative logit link function and another with cumulative link function were fitted to the data, which succeeded in identifying significant effects after testing the relationship between the fixed effects and the outcome variable: *food consumption score*. The model was able to determine predictors of food insecurity risks, especially in crises situations. Both techniques identified six out of possible seven factors that could influence a 'worse' outcome of food insecurity. It further identified the states that could be affected more by any potential risk to food insecurity status. Similarly, it identified the range of factors that could adversely cause the population's food insecurity to worsen.

The next step was to examine how well the model predicted the risk of food insecurity. Analysis showed some issues with classification of the predicted categories, which are, however, typical for proportional odds models (World Food Programme 2008). Considering, however, that 98% of the responses in the '*acceptable*' *food consumption* category were correctly classified it called for either adjusting the outcome categories using appropriate statistical technique, or dichotomising the categories such that there are only two, say, '*poor*' and '*acceptable*' food consumption score categories and accordingly employ techniques for analysing binary data. For more information on the different strategies to improve predictions and model inspection methods, please refer to McCulloch and Searle (2001) and Collet (2003).

5.7. Conclusion

Data analysis explored in this chapter featuring the Generalized Linear Mixed Model, which accounted for between-cluster effects, revealed a number of mixed results. Although the method managed to identify significant predictors of *food consumption score*, it showed some misclassification issues of predicted values. To resolve this issue, three recommendations could be made. One, food security analysts may need to rethink the categorisation of the food security score. However, the outcome variable *food consumption score* could be a better measure than its comparators in that combines property of being both money metric and livelihood asset-based measure. Secondly, in future analysis with similar aim the use of modelling approaches that take complex survey designs into account might be useful. Thirdly, since shortcomings inherent in datasets from large household surveys are well noted in statistical modelling, it may be important to explore analysis with smaller, controlled, but representative sample sizes. Good data can guarantee the power of the model.

With improvements on data quality and study design, the model could be useful to food security and disaster early warning system analysis. It is hoped that the results of the study may provide evidence for future analysis. Results of the analysis explored confirmed that interventions for increasing income of households would improve food availability, access and consumption for households. Such interventions help in enhancing resilience of populations endowed with land, animal and other natural resources and favourable environment for agriculture, but subsisting in protracted crisis.

Finally, the analysis could not lead to recommending the approach for generating a resilience based index based on the set of identified important predictors. The highlighted issues of misclassification of the predicted cases, especially for the lower categories of the outcome/response variable must be resolved first.

CHAPTER 6

Joint Modelling of Coping with Food Insecurity and Food Consumption Expenditure

6.1. Introduction

In this chapter we explore joint modelling of coping strategies and food consumption expenditure responses using a dataset from South Sudan. A joint model attempts to explore how two mutually reinforcing effects are determined by a set of possible predictors of food insecurity. The chapter specifically examines and discusses the significance of the study and the use of Generalized Linear Mixed Model for determining predictors of the joint outcome of '*coping*' (i.e., household adopted a coping strategy) or '*no coping*' (i.e., household did not adopt any coping strategy) with food insecurity risks on one hand, and share of total household budget expenditure on food on the other. Section 6.2 explains the concepts of coping with food insecurity on one hand, and share of food consumption scores on the other, and discusses why the two variables are important indicators of food insecurity risk. The Section then presents the rationale behind the relationship between coping with food insecurity and share of food consumptions in total household expenditure. It further justified how this mutual relationship motivated joint modelling of the data.

Section 6.3 describes the sample and data used in the study. It also provides some literature on the two variables adoption of coping strategy and food share of consumption expenditure. Section 6.4 sheds some highlights on the Joint Model and its formulation. Section 6.5 presents the results of the Joint Model and some interpretations. Sections 6.6 discusses the findings and Section 6.7 is draws some conclusions based on the analysis.

6.2. The Relationship between Food Insecurity Coping and Food Expenditure

Often when data are clustered with different attributes observed on the same sampling unit, mixed outcomes may occur. Gardiner (2013) observe that such mixed outcomes can be in the form of continuous, count and categorical types (multinomial, ordered polytomous or binary). Food security outcomes can take all these forms. Monetary measures of consumption expenditure and income are often on continuous scale, while asset or index-based measures from household characteristics are largely ordered polytomous or binary. Nutritional status is often based on the continuous anthropometric measures of weight, height, age, arm circumference and z-scores. In this case, observed factors or events (random or fixed) may impact both outcomes. Multivariate or joint modelling is commonly applied in studies of repeated measures, longitudinal and spatial data. However, we examine the data with joint responses, given the nature of the study attributes.

Coping with food insecurity experiences and increased share of total household budget spent on food tend to lower household resilience to food insecurity risks. Certain household attributes and means of livelihood conceptually tend to influence the likelihood of the two food security outcomes. The higher the cost of food, the higher the household's expenditure on food, thus compromises acquisition of other equally important life sustaining essentials, such as health, education and developmental plans. This leads to adoption of coping strategies, some of which could have undesirable socioeconomic impacts. It is, therefore, important to understand the seriousness and the magnitude of determinants of the two outcomes, and how these determinants impact on food security status. Such knowledge could help in developing appropriate policy for preventable action against adverse events, such as lack of adequate investment in agriculture, animal production, fisheries and aquaculture. Where the effect is

quite dramatic, this might prompt policy makers and programmers to decide where and when to take action.

The Food and Agricultural Organisation of the United Nations (FAO) (2015, p.39) caution that major causes of food insecurity are likely to "persist for some time", forcing households to adopt short-term coping strategies and thus rendering livelihoods unsustainable. In Snel and Staring (2001, p.10), coping strategies are referred to as "all the strategically selected acts that individuals and households in a poor socioeconomic position use to restrict their expenses or earn some extra income, to enable them pay for the basic life necessities (food, clothing, shelter, security and health) and not fall too far below their society's level of welfare".

Share of total household expenditure on food is one of the measures of food in/security defined under the four dimensions of food security (availability, access, stability and utilization). It obviously belongs to the *access* component of food security. FAO officially catalogues this indicator as "Share (%) of food consumption expenditure in total consumption expenditure" in its list of indicators. It defines the indicator as "the monetary value of acquired food, purchased and non-purchased, including non-alcoholic and alcoholic beverages as well as food expenses away from home consumption in bars, restaurants, foodcourts, work canteens, street vendors, etc." The numerator excludes non-consumption expenditure, such as direct taxes, subscriptions and insurance premiums (Sibrián et al. 2008). A study of food consumption expenditure comparing Uganda, Vietnam and Peru concludes that rural households in Uganda use a larger share of their household budget for food consumption (Maltsoglou n.d.).

Conceptually, an increase in food consumption expenditure is not only a risk factor that entrenches resource-poor populations in poverty, and a vicious cycle of food insecurity-related vulnerability, but it also forces households in this category to adopt some form of coping strategy. Conversely, when households are barely coping with lack of food, it induces

propensity to spend on food. This implies that there is a reciprocal effect between the two variables (coping strategy and food consumption expenditure).

It is further worth noting that both co-indicators of structural food insecurity have dire implications on perpetual asset poverty. The more a population adopts coping strategies, the more they keep spending a lion share of their incomes on food, and the more they are drenched into more asset poverty. Maxwell (2003, p.8) determines that the indicator 'food share of household budget' has positive correlation (0.195) with the coping strategy index (CSI), which means that an increase in expenditure on food pushes a household to cope to some extent. This is consistent with a test of association we carried out in exploratory analysis of this study using logistic regression with log link function and the Wald Chi-square test of hypothesis. We found that adoption of coping strategy associated very highly with increase in share of food expenditure.

The reader might wonder as to whether food consumption expenditure applies invariably for rural population as it does for urban population. The concern could arise from the assumption that often times a substantial proportion of rural populations don't purchase food from markets, but rather produce their own food. The answer to such concern rests on two important considerations. First, although the variable "residential setting", or whether an interviewed household was rural or urban was not in the survey questionnaire, the sample data contain households from locations or clusters located in urban centres such as Malakal, Wau and Maridi; thus mostly depended on market purchase of food. Second, it is a known fact that South Sudan emerged from a protracted civil war with wide ranging displacement of citizens. This post-conflict status, worsened by a raging armed conflict during data collection, must have rendered a large proportion of the population to be unable to produce their own food.

The study aims to explore a joint model in the analysis. Most importantly, our aim is to assess the determinants of the combined risk to food insecurity, which could entrench vulnerable and chronically food insecure populations to get hopelessly exposed to a vicious cycle of asset poverty and exposure to associated risks of food insecurity shocks. Extensive literature search has found out that joint modelling techniques have not been explored before in the analysis of the mutually reinforcing outcomes of food insecurity risk. The technique has featured in biological science research, medicine and health other such as in Gardiner (2013). Thus, the study seeks to answer two related questions: i) Do enhancers of livelihoods in structural poverty and food insecurity settings determine the combined risk of entrenchment in coping with and spending highly on food? ii) Is joint modelling of the two outcomes a potentially good tool to be relied on in analysis of this nature?

6.3. Sample and Data

Data used are described in Section 2.2.2. It is worth noting that during the data collection, South Sudan, the study setting, was experiencing intense fighting between government and rebel troops, which continued until the time of writing, a large portion of the population was displaced and food insecurity was predicted to reach crises levels. The main purpose of the survey that produced the data was, therefore, to generate essential information for estimating vulnerability and reinforce planning for appropriate interventions. The purpose of the survey and that of this study are thus not the same.

The study response variables are a dichotomised *coping strategy index* (CSI) and ratio of *household expenditure on food*, which is a scale variable. The coping strategies index was first developed and established by Maxwell (1995) for "distinguishing and measuring short-term food insecurity at the household level", but also for monitoring food emergencies, early

warning and the impact of interventions (Maxwell & Caldwell 2008). For more detailed description of the derivation of the coping strategies index and its categories, see Section 4.2.

The proportion (ratio or percentage) of household expenditure on food is a good indicator of poverty as well as potential food insecurity vulnerability and risk. The measure is obtained by asking questions on the amount of money spent on a range of food and non-food items and services in past 30 days (31 items in total). The total consumption expenditure is then calculated and the percentage of amount spent on food is obtained. This indicator is thus by far straight forward.

The four livelihood-based effects asked during the survey were *crop cultivation* in preceding farming season, ownership of *livestock, fishing* by any household member and *main source of income* being. Apart from *main source of income,* which had four levels, all the other effects had two levels. In consideration of arguments by FAO (2015, p.39) that "Gender and age are two powerful determinants of the impact of protracted crises on individuals", three covariates *gender of household head, age of household head* and *size of household,* were also included in the explanatory variables.

6.4. The Joint Model (or Multi-equation model)

As argued in Section 6.2, spending more of the household income on food (measured as a continuous outcome Y_1) leads to adopting of a coping strategy (a binary outcome Y_2) during food scarcity. In order to establish the inter-dependence of the two outcomes a test of correlation was carried out before proceeding with joint modelling.

As discussed in Chapter 4 and 5, seven variables were justified to influence food insecurity outcomes, especially in humanitarian crisis, namely; *gender* and *age of household head, household size, crop cultivation, livestock ownership, fishing* and *main source of income*. A

challenging aspect of the analyses is the presence of exogenous factors affecting these dual outcome variables. Hence, we need to specify a joint model for Y given the exogenous factors z. Each outcome was modelled separately using an appropriate generalized linear model by structuring the mean $E(\mathrm{Y}_k|z_k)$ and variance $Var(\mathrm{Y}_k|z_k)$, where $k = 1,2$. The covariates z_1, z_2 do not necessarily need to be the same, although in practice some overlaps do occur. To simplify interpretation and identification, some variables may be excluded from a model for one outcome, which are included in a model for another outcome. An alternative approach to joint modelling is Copula Regression that has received some attention (Kolev & Paiva 2009).

Linking the two outcomes is done through a shared random effect ζ in $E(\mathrm{Y}_k|\mathrm{z}, \zeta)$ or by structuring the covariance matrix $Var(\mathrm{Y}|z)$ to ensure potential correlations are included in the model. As our interest is centred on the influence of food expenditure percentage Y_1 on coping strategy Y_2, gives the joint distribution

$$\int(\mathrm{Y}_1, \mathrm{Y}_2|\mathrm{z}) = f(\mathrm{Y}_1|\mathrm{z})\, f(\mathrm{Y}_2|\mathrm{Y}_1, \mathrm{z}) \qquad (6.1)$$

where $f(.|.)$ is a conditional distribution. Thus Y_1 is endogenous in a second term model.

The approach for modelling the data was to use SAS GLIMMIX procedure (IBM Corporation 2015), for modelling the two responses conditional on random effects.

Data analysis proceeded in two stages. In the first stage, each outcome was analysed separately based on the univariate logistic regression model (GLM) fitting the selected explanatory variables. In the second stage both outcomes were analysed jointly using GLIMMIX. This procedure was used to take care of random cluster variations as the data came from randomly selected clusters of households. In the first stage the estimates of the model parameters were obtained by maximum likelihood. For formulation of the maximum likelihood of a model with

continuous or Gaussian distribution and for a binary response, see McCullagh and Nelder (1989), Collet (2003) and McCulloch and Searle (2001).

In modelling two responses, suppose that the outcome variable are Y_1=*Food expenditure* – a continuous outcome, and Y_2=*Coping* – a binary outcome. The two events are somehow correlated. High household food expenditure and being compelled to cope with risk of food insecurity increase vulnerability and at the same time reduce resilience in food insecurity crises. Putting it differently, when a household is obliged to adopt a coping strategy it equally forced to spend more of its income or savings on food. Coping with food insecurity crisis is like a co-morbidity impacting on high expenditure on food and thus increases vulnerability and adversely lowers resilience to food insecurity and livelihood shocks. Therefore, one adverse event Y_1 occurs jointly with another adverse event Y_2.

If interest is centred on the effect of household food expenditure Y_1 on coping with food insecurity, a joint distribution might be worth considering such that $f(Y_1, Y_2|z) = f(Y_1|z)f(Y_2|Y_1, z)$, where the generic notation $f(.|.)$ denotes a conditional distribution. This makes Y_1 to be potentially endogenous in a model of the second term.

Consequently, a generalized linear model $g_k(Y_{ik}|z_i) = z'_{ik}\beta_k$, where i is a household, $k = 1,2$ and g_k is a link function for the outcome k is then fitted to the data. The joint model may be fitted with the different covariates from the data. The models for the two equations are in the form

$$\mathbf{Y}^*_{i1} = \boldsymbol{z}'_{i1}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_{i1} \text{ and } \mathbf{Y}^*_{i2} = \boldsymbol{z}'_{i2}\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}_{i2} \quad (6.2)$$

where $\mathbf{Y}_{i1} = [\mathbf{Y}^*_{i1} > \mathbf{0}]$ are the observable indicators, The covariates $z_i = (z_{i1}, z_{i2})$ are exogenous, which means $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}, \boldsymbol{\varepsilon}_{i2}) \sim \boldsymbol{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{1} & \boldsymbol{\rho_{12}\sigma_2} \\ \boldsymbol{\rho_{12}\sigma_2} & \boldsymbol{\sigma_2^2} \end{bmatrix}. \tag{6.3}$$

This model assumes $\boldsymbol{\Sigma} = \boldsymbol{diag}(\mathbf{1}, \boldsymbol{\sigma_2^2})$ which is the same as assuming the two responses separately, with exception of degrees of freedom. The covariance, expectations and variance of the parameter estimates are respectively

$$\boldsymbol{Cov}(\mathbf{Y_{i1}}, \mathbf{Y_{i2}}|\boldsymbol{z_i}) = \boldsymbol{\rho_{12}\sigma_2\phi}(\boldsymbol{z'_{i1}\beta_1}) \tag{6.4}$$

$$\boldsymbol{E}(\mathbf{Y_{i1}}|\boldsymbol{z_i}) = \boldsymbol{\Phi}(\boldsymbol{z'_{i1}\beta_1}) \tag{6.5}$$

$$\boldsymbol{Var}(\mathbf{Y_{i1}}|\boldsymbol{z_i}) = \boldsymbol{\Phi}(\boldsymbol{z'_{i1}\beta_1})\big(\mathbf{1} - \boldsymbol{\Phi}(\boldsymbol{z'_{i1}\beta_1})\big), \tag{6.6}$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\Phi}$ denote the density and cumulative distribution of the standard normal distribution. The procedure then structures the variance matrix of $\mathbf{Y}_i = (\mathbf{Y}_{i1}, \mathbf{Y}_{i2})$ as

$$\boldsymbol{Var}(\mathbf{Y_i}|\boldsymbol{z_i}) = \mathbf{A_i^{1/2}R}_i\mathbf{A_i^{1/2}} \tag{6.7}$$

where $\mathbf{R}_i$ is a user specified 2×2 covariance structure and $\mathbf{A}_i$ is the diagonal matrix of the variance of $(\mathbf{Y_{i1}}, \mathbf{Y_{i2}})$.

The error matrix of Equation derives estimates for $\boldsymbol{\sigma_2}$ and $\boldsymbol{\rho_{12}}$ based on the residual pseudo-likelihood.

Further discussion of the GLMMs can be found in Littel et al. (2006), Breslow and Clayton (1993) and Wolfinger and O'Connell (1993), who derive extensions of the generalized linear models (GLMs) to the GLMM.

6.5. Results

Before starting analysis using joint modelling, it was necessary to establish the correlation between the two response variables. Both Kendall's tau and Spearman's rho tests of correlations showed significant (p-value > 0.005 at the 0.01 significance level, two-tailed test) correlations between share of *food consumption expenditure* (a scale variable) and *coping strategy index* (also a scale variable). Since interest was to determine whether or not the variable *share of food expenditure* and *coping* were correlated, both variables needed to be measure on scale. Coping strategy index was initially scale variable as it was generated based on weights.

With a positive correlation of 0.152 (Pearson's rho), the result is consistent with that of Maxwell et al. (2003, p.8). This finding offers optimism to proceed with exploring joint modelling of the two correlated outcomes.

The univariate logistic regression model was fitted to the data with the first response variable *share of food expenditure* using Generalized Linear Model Procedure with a logit link function. From the analysis shown in Table 6.1, all the seven fixed effects except *gender* of household head were determined to have significant contributions to the model. Specifically, a household whose head was aged between 18 to 60 (the economically active group), had above average members, cultivated crops and owned livestock in the previous farming season, engaged in fishing and earned income from sale of livestock products and petty trade, showed significant associations with increase in food expenditure. Table 6.1 also shows that gender of household head and main source of income from farming and employment did not have significant association with increase in consumption expenditure. This clearly meant that both female- or male-headed households did not differ significantly in relation to increase in consumption expenditure. Similarly, there was no statistical evidence suggesting that whether a household

earned income from sale of agricultural harvest or not, or from some form of employment, it would still fare equally in terms of spending on food.

Table 6.1: Maximum likelihood parameter estimates from the *expenditure* model

Parameter*	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr>ChiSq**
Intercept	4.3777	0.0488	8052.96	<0.0001
Gender: *male*	0.0113	0.0115	0.96	0.3280
Age: < 17	0.0692	0.0604	1.31	0.2524
Age: *18-60*	0.0821	0.0335	6.01	0.0142
Household Size	-0.0097	0.002	23.29	<0.0001
Cultivated crops: *yes*	0.0351	0.0135	6.71	0.0096
Owned livestock: *yes*	-0.1829	0.0249	54.18	<0.0001
Engaged in fishing: *yes*	0.0515	0.0154	11.16	0.0008
Income Source: *Agriculture*	0.0196	0.0241	0.66	0.4166
Livestock products	0.123	0.0242	25.85	<0.0001
Employment	0.0365	0.0246	2.2	0.1383
Petty trade	0.0625	0.0244	6.54	0.0105

* Values corresponding to reference categories are set to zero and are not shown.

Overall, the result resonated with expectations as regards increased spending on food, especially for a population in crises. In practice, however, this finding does not tell much, since, for instance, a household can sell its harvest to earn income, which could be spent on buying food when food prices are high. Yet, the results still show generalised high poverty levels, where earnings and usual agriculture-based sources of livelihood cannot offset demands for or improve access to adequate food. The results generally manifest entrenched food insecurity threat. With generalized high poverty levels, expenditure on food tends to prevail. Therefore, it is worthwhile examining a model with the other co-response.

We continued to fit a GLMM model for the data with a '*coping*' response, but this time allowing for random effects due to clusters. The reference category of the response was set at 0 (i.e., household did not adopt coping strategy). Note that results of Table 6.2 with significant

effects showed negative estimates of coefficients for all seven fixed effects. This means that the odds of the reference levels associated better with the probability of adopting a coping strategy. Table 6.2 shows four of the seven effects included in the model as significantly associated with adoption of a coping strategy. These are *age of household head*, *crop cultivation*, *livestock ownership* and *main source of income*. The model determined *gender*, *household size* and *fishing* to have no significance difference in adopting a coping strategy.

Table 6.2: Estimates of effect parameters from the 'coping' model

Parameter*	**Estimate**	**Standard Error**	**DF**	**t Value**	**Pr > |t|**
Intercept	1.9795	0.5810	149	3.41	0.0008
Gender: *male*	-0.03245	0.08983	3441	-0.36	0.7180
Age: < 17	-1.2668	0.4535	3441	-2.79	0.0052
Age: *18-60*	-0.5517	0.2472	3441	-2.23	0.0257
Household Size	0.01399	0.01600	3441	0.87	0.3821
Cultivated crops: *yes*	-0.2117	0.1102	3441	-1.92	0.0547
Owned livestock: *yes*	-1.0553	0.4954	3441	-2.13	0.0332
Engaged in fishing: *yes*	-0.07913	0.1392	3441	-0.57	0.5699
Income Source: *Agriculture*	-0.4852	0.1832	3441	-2.65	0.0081
Livestock products	-0.05759	0.1895	3441	-0.30	0.7612
Employment	-0.3686	0.1871	3441	-1.97	0.0490
Petty trade	-0.1538	0.1873	3441	-0.82	0.4114

* Values corresponding to reference categories are set to zero and are not shown.

Note the negative coefficients of estimates of the fixed effects, which are interpreted as in earlier models (Chapter 4 and Chapter 5). In general, there was statistical evidence that a household headed by a person aged above 60 years, did not cultivate crops, had not owned livestock and did not depended on income from a source other than employment, of association with adoption of a coping strategy for its livelihood. Meanwhile, there was no sufficient evidence to suggest that gender of household head, household size, fishing and main income from sale of livestock products and petty trade associated with coping.

Note also that in terms of significance of the fixed effects, results of the '*coping*' model did not differ considerably with those of the '*expenditure*' model. However, there are also some differences. Most of the estimates of model in Table 6.1 are positive whereas almost all the significant effects in '*coping*' model (Table 6.2). The two procedures modelled two separate outcomes.

The foregoing results of analysis of fitting the two outcome variables separately (Table 6.1 and 6.2), give a sense that a household with certain attributes such as being headed by an older person, having a larger household size could risks being food insecure as a result resorting to adopting a coping strategy and at the same time tending to spend more on food than other life essentials. Both types of food insecurity risks seem to be associative from the point of view of common factors (age and size of household size). It would, therefore, be of interest to explore how the possible association could if they are jointly modelled. This motivated the idea of fitting a joint model, in order to see whether a joint distribution of the two outcome variables have significant associate with some or all of the fixed effects. It is also to be recalled that the two outcome variables were determined to be correlated.

We fitted a GLMM for both responses jointly and with the explanatory variables and latent random effect, which accounts for the association between coping and expenditure. This means that we fitted the model with random cluster intercept, i.e., a variance matrix blocked by *cluster*. The model is fitted using the maximum likelihood with adaptive Gauss-Hermite quadrature, given that it restricts the models for estimating parameters and also fulfils conditional independence assumptions and the processing of data by subject (Lange 1999). The choice enables linearization of the non-linear random effects (i.e., the *cluster* variable).

As shown in Table 6.3, the joint model was by far improved. It showed all seven effects as highly significant as opposed to the univariate models. That is, all fixed effects had associations

with the joint outcome of '*food expenditure*' and '*coping*'. Unlike in previous models, fishing was significant after fitting the Joint Model.

Table 6.3: Type III tests of the explanatory variables from the Joint Model

Effect	Num DF*	Den DF*	F Value	Pr>F
Intercept	2	6998	770.17	<0.0001
Gender	2	6998	0.96	0.3824
Age	4	6998	3.96	0.0033
Household size	2	6998	14.08	<0.0001
Cultivated crops	2	6998	5.46	0.0043
Owned livestock	2	6998	19.47	<0.0001
Engaged in fishing	2	6998	5.42	0.0044
Main source of income	8	6998	8.97	<0.0001

* Numerator and denominator degrees of freedom; Pr is short for 'probability'.

The joint model also generated the estimates of coefficient of the explanatory variables (Table 6.4). Gender of household head stayed non-significant for the both outcomes, as in the univariate models. The joint model determined fishing to be significant in the normal response. There were also other changes in the results of the joint model that showed significant relationships between some fixed effects and either of the food insecurity outcomes. An example is main source of income from employment and from petty trade which became significant, after they were shown as non-significant in the separate models. A significant probability indicates that there is sufficient statistical evidence to suggest that the corresponding variable influences both household coping and increased expenditure on food.

The coefficient estimates from the '*coping*' distribution in the joint model were close to those of the univariate '*coping*' model, and it was also dominated by negative signs, meaning that the odds of association with coping strategies were worse for the reference levels. These reference categories were age of household head above 60 years, household size of above seven persons, household that did not cultivate crops in the previous season, and household that did

not depend on agriculture for income and food, and did not earn salaries and wages. These findings are not far from expectation.

For the expenditure responses using log link function, ownership of livestock showed significant relationship. The result showed that a household headed by a person aged 18 to 60 years, had the size of four to six members, cultivated crops, owned livestock and mainly earned income from sale of livestock and petty trade, had significant associations with increase in food expenditure. Cattle are sold in case of extreme coping strategy. This finding led us to further investigate as to whether keeping livestock is associated with extreme coping strategies, where selling cattle is a last resort. Separate analysis based on descriptive and non-parametric tests confirm that there is non-significant association between livestock keeping and adoption of severe coping strategy. In fact, only 0.6 *per cent* of households which kept cattle reported they adopted extreme coping strategies.

As noted above, some values that were significant in either of the bivariate independent models no longer seemed significant in the joint model. Although these differences did not cause major changes in significance values as to upset the test results, the disparities could arouse some concerns. Meanwhile, the factor *gender of household head* remained non-significant in all the three models; even experimental ones with results not shown. For this reason, we extended the joint analysis with a view of removing such variables from the analysis.

Table 6.4: Estimates of the explanatory variable coefficients under the Joint Model

Effect*	Distribution	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	Binary	1.9188	0.5996	6998	3.20	0.0014
	Normal	80.2495	3.6664	6998	21.89	<.0001
Gender: *male*	Binary	-0.03716	0.09081	6998	-0.41	0.6824
	Normal	1.1331	0.8573	6998	1.32	0.1863
Age: *<18*	Binary	1.2990	0.4573	6998	-2.84	0.0045
	Normal	5.0093	4.3001	6998	1.16	0.2441
18-60	Binary	-0.5843	0.2496	6998	-2.34	0.0192

Effect*		Distribution	Estimate	Standard Error	DF	t Value	Pr > |t|
		Normal	5.9723	2.3113	6998	2.58	0.0098
Household size		Binary	0.01301	0.01619	6998	0.80	0.4214
		Normal	-0.7763	0.1484	6998	-5.23	<.0001
Crops cultivation: *yes*		Binary	-0.2406	0.1115	6998	-2.16	0.0310
		Normal	2.4171	0.9834	6998	2.46	0.0140
Owned livestock: *yes*		Binary	-0.9075	0.5163	6998	-1.76	0.0789
		Normal	-13.6979	2.2145	6998	-6.19	<.0001
Fishing: *yes*		Binary	-0.1207	0.1413	6998	-0.85	0.3931
		Normal	3.7548	1.1910	6998	3.15	0.0016
Main source of income	*Agriculture*	Binary	-0.4685	0.1856	6998	-2.52	0.0116
		Normal	1.3368	1.7019	6998	0.79	0.4322
	Livestock	Binary	-0.09925	0.1922	6998	-0.52	0.6056
		Normal	8.7699	1.7426	6998	5.03	<.0001
	Employment	Binary	-0.3716	0.1897	6998	-1.96	0.0501
		Normal	2.9137	1.7426	6998	1.67	0.0946
	Petty trade	Binary	-0.1629	0.1899	6998	-0.86	0.3911
		Normal	4.7002	1.7420	6998	2.70	0.0070

* Reference categories are not shown.

In inspecting the covariance parameter estimates of the joint model (Table 6.5), we found that the estimate of the variance of the random effect cluster intercept was 1.6543 with a corresponding standard error estimate of 0.2704. This indicates that there could be significant within-cluster variation in the intercepts. This variation was accounted for in the inference.

Table 6.5: Covariance Parameter Estimates

Covariance Parameter	Subject	Estimate	Standard Error
Intercept	Cluster Number	1.6543	0.2704
Residual		502.35	12.1140

The joint model was then refitted for modelling correlations directly such that instead of a shared G-side random effect, an R-side covariance structure is used to model the correlations of a marginal model, which models covariation on the data scale. Specification of the standard variance component (*vc*) in the new model causes different clusters (pools of villages in one geographical location) to be independent, while single clusters followed this model. The *vc* structure was such that for each effect a distinct variance component was assigned, which is

also known as a G-side covariance structure. This enables the R-side variance structure to only add the effects of over dispersion. The *vc* covariance structure is of the form

$$\begin{bmatrix} \boldsymbol{\sigma_B^2} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\sigma_B^2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\sigma_B^2} \end{bmatrix}$$

With this specification of the covariance structure, some changes in the estimates of the joint model occurred (Table 6.6). The most notable change is the removal of the variable *gender of household head* from the analysis. The fixed effect *fishing* even became highly significant.

Table 6.6: Type III tests of effects of selected factors with unstructured covariance structure under the joint model*

Effect	Num DF	Den DF	F Value	Pr > F
Intercept	2	148	1565.24	<.0001
Age of household head	4	170	3.34	0.0117
Size of household	2	6999	24.36	<.0001
Cultivated crops	2	264	7.01	0.0011
Owned livestock	2	6	46.73	0.0002
Engaged in fishing	2	171	11.14	<.0001
Main source of income	8	873	17.15	<.0001

* After removal of gender of household head

An interesting observation (Table 6.7) is that no fixed effect had significant *p*-values in both co-responses in the joint model that includes G-side and R-side correlation. Meanwhile, in the model with only G-side correlations, three variables (with values shown in boxes) were significant in both responses. It is also interesting to note that both models showed very highly significant relationships (*p*-value < 0.05) between five fixed effects (*household size*, *crop cultivation*, *fishing* and *livestock ownership* and *income from sale of livestock products* and

petty trade) and the response higher food expenditure. This means households with these attributes risked being food insecure due to their high spending on food.

Table 6.7: Solution for fixed effects of the model with covariance structure

		Joint Model with G-side Correlations only			Joint Model (G-Side and R-side Correlations)		
Effect	**Dist***	**DF**	**t Value**	**Pr > |t|**	**DF**	**t Value**	**Pr > |t|**
Intercept	Binary	7000	3.18	0.0015	148	0.37	0.7108
	Normal	7000	21.93	<0.0001	148	31.62	<0.0001
Age: *<18*	Binary	7000	-2.84	0.0045	166	-0.20	0.8407
	Normal	7000	1.36	0.1731	166	1.62	0.1064
Age: *18-60*	Binary	7000	-2.33	0.0199	166	0.19	0.8457
	Normal	7000	2.80	0.0051	166	3.63	0.0004
Household size	Binary	7000	0.77	0.4388	6999	0.15	0.8804
	Normal	7000	-5.15	<0.0001	6999	-6.98	<0.0001
Crops cultivation	Binary	7000	-2.18	0.0295	264	-0.05	0.9568
	Normal	7000	2.53	0.0114	264	3.74	0.0002
Owned livestock	Binary	7000	-1.77	0.0760	6	-0.35	0.7363
	Normal	7000	-6.27	<0.0001	6	-9.66	<0.0001
Fishing	Binary	7000	-0.87	0.3842	171	0.04	0.9666
	Normal	7000	3.20	0.0014	171	4.72	<0.0001
Income: *agricul.*	Binary	7000	-2.57	0.0101	873	-0.23	0.8172
	Normal	7000	1.27	0.2032	873	1.29	0.1981
Livestock	Binary	7000	-0.54	0.5926	873	0.07	0.9423
	Normal	7000	5.49	<0.0001	873	7.74	<0.0001
Employment	Binary	7000	-2.00	0.0451	873	-0.21	0.8307
	Normal	7000	2.10	0.0361	873	2.40	0.0167
Petty trade	Binary	7000	-0.88	0.3815	873	-0.11	0.9164
	Normal	7000	3.11	0.0019	873	3.89	0.0001

It is to be noted that the degrees of freedom multiplied as a result of fitting a joint model. His is mainly because of each of the fitted effects was taken to have interacted with a latent variable representing a joint distribution of the two outcome variables considered in the analysis.

The foregoing finding led to selection of the model with R-side covariance structure only to the data and generating linear predictors and residuals. This is of course done after removing the variable *gender* of household head. Figure 6.1 shows a plot of the residuals against clusters. Clearly most of the errors of the model are clustered around 0, which shows good amount of

prediction of the joint response variable (‘*response’*). The linear predictor can, therefore, be used as an index for determining the likelihood of food insecurity risk represented in the joint outcomes of food consumption expenditure and incidence of coping with food insecurity strains.

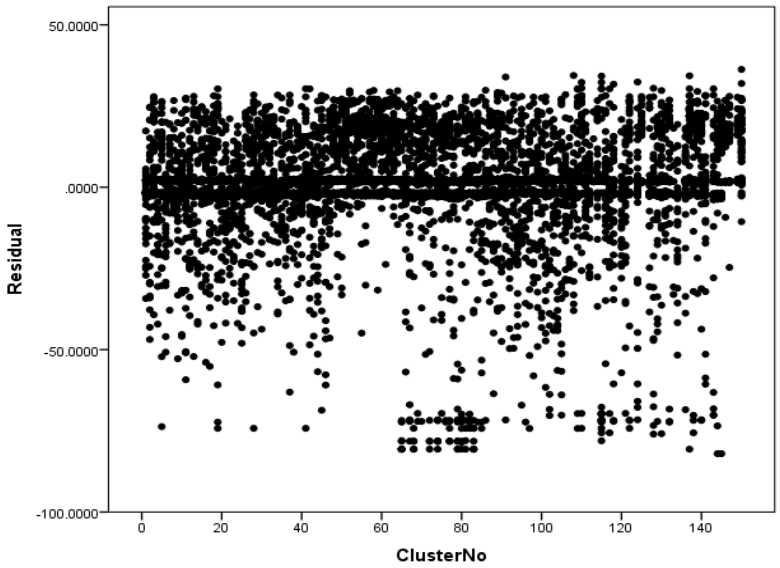

Figure 6.1: Plot of residuals from the Joint Model

6.6. Discussion

Seven explanatory variables were examined in analysis involving univariate joint models for both coping during food insecurity and expenditure on food experiences. This is in order to establish how indicators of a typical agro-pastoralist economy determines the likelihood of food insecurity risk, as characterised by the combined outcomes of coping with shortage of and spending on food. In all three models explored, one of the seven assumed factors, gender of household head, persistently showed non-significant relationship with adoption of coping strategy and/or increased expenditure on food. There were noticeable disparities in the significance of the relationship of the different levels of fixed effects and either response. There

were also some noticeable inconsistencies in the way explanatory variables showed significant relationships with either outcome of food insecurity (increase in the share of expenditure on food and coping strategy experiences. This could be due to the generalized vulnerability and poverty where coping was rampant and the order of the day in many of the communities. It could also be due to misperceptions by respondents of the meaning of coping strategies.

The three-stage analysis also determined that three of the main variables investigated (crop cultivation, engagement in fishing and main source of income), in addition to two co-factors (age of household head and size of household) to have significant relationships to with both coping strategy and increased food consumption expenditure. This finding could be interpreted in that presence of these livelihood constructs provided adequate statistical grounds as to predict the probability of the two outcomes being practised by a household in the sample. In other words, if conditions stayed the same, there was risk of the population becoming vulnerable to food insecurity.

Of special importance, the results of the analysis showing that the joint model effectively modified the results of the separate, univariate modelling of the outcomes studied. The joint model specifically caused values of test of hypothesis that had non-significant or low-significance probabilities, to be highly significant. A case in point was crop cultivation and fisheries. These findings were consistent with those of Gardiner (2013). An attempt to improve the model further by adding R-side random effects did not result in removing any of the fitted effects. A fitted model with the variable *gender* of household head was refitted to the data, which led to generating predicted probabilities. These predicted probabilities were plotted using a scatter plot to check for goodness-of-fit of the model.

6.7. Conclusion

Joint modelling has not been used before in analysing food security outcomes, especially those explored. The method has mainly prevailed in longitudinal health and medical studies, as well as in survival analysis. The study has unsealed three important findings. First, coping strategies and food expenditure in chronic food insecure situations combine a function of agriculture-based livelihoods and household characteristics. Second, the methods have demonstrated that agriculture and fishing and income from sale of livestock combine with age of household head and size of household size as predictors of either or both of the two outcomes of chronically food insecure and poor populations. Three, based on earlier works by Filmer and Pritchett (2001), Moser and Felton (2007) and Lokosang et al. (2014) in which they established that household assets provide good basis for generating an index of socioeconomic status and food security of a household, the joint modelling might have the capability of generating an index for determining household resilience to food insecurity risks, especially in settings characterised by structural food insecurity. Once efficiency of the method is established (see Figure 6.1) predicted values of the joint response variable can be taken for the index that predicts the probability of food insecurity exemplified by how a particular cluster of households cope with food insecurity uncertainty and increased expenditure on food.

CHAPTER 7

Discussion and Conclusion

This research project attempts to extend the application of statistics in food security measurement and informatics. Despite the problem of food insecurity, malnutrition and strained livelihoods gaining importance due to its chronic and endemic nature, especially in Africa, statistical applications such as the ones explored in this project, have not been adequately utilised. Food insecurity measurement is gaining focus, but flexible statistics modelling techniques are seldom applied. In addition, latest discussions centre on some pertinent constructs of food security, namely; vulnerability, resilience and risk. This heightens the need to find more efficient measurement approaches.

The approaches for measuring the incidence, prevalence or probability of food insecurity occurrence are neither harmonized nor streamlined, as different institutions continue to use different approaches. By the break of the Third Millennium, substantial body of literature has drawn attention to the need to develop standard measures for monitoring food insecurity and malnutrition, especially in protracted crisis settings. Ever since food insecurity and recurrent famines imposed themselves as major global problems, numerous national surveys such as the demographic and health surveys, food security monitoring surveys and annual livelihood assessments, have been carried out, but mainly to generate information on the status of food in/security or for determining the need for relief interventions.

Despite the plethora of measures and approaches for determining the risk of food insecurity, there is limited research on the use of robust statistical techniques such as the ones, especially for determining the likelihood of occurrence of vulnerability. In addition, there seems to be no attempt at developing a single measure for assessing the risk of food insecurity. Instead, most

food insecurity and malnutrition assessment focused on retrospective measurement of vulnerability and status of food insecurity.

It is against this backdrop that this study sets focus to examine potential statistical approaches for determining potential risk of food insecurity with the aim of averting risk rather than determining it after it has occurred. The study borrowed significantly from research work in which asset-based measures were generated for determining socioeconomic and poverty outcomes.

Six different statistical procedures were explored that took the form of two streams of approaches. The first stream examined nonparametric, classification-based factor analysis, namely; principal component analysis, multiple correspondence analysis and classification and regression tree analysis. This class of analysis helped in identifying important predictors of food insecurity outcomes. Then based on factor loadings of these predictors, an index was generated named "Household Food Insecurity Resilience Index" (or HRI, in short) to represent a single latent or composite predictor variable of food insecurity outcome.

In the second stream of analytical methods, four regression based models were explored, namely; Binary Logistic model and three models featuring Generalized Linear Mixed Models to identify most important predictors of food insecurity outcomes and determining an index variable generated from the predicted values of the outcome variables. Important conclusions were generated from this type of analysis. The study found each method explored to have peculiar strengths as well as limitations.

Principal component analysis succeeded to generate an asset-based household food insecurity resilience index, which was determined to have six important attributes. Three of these attributes are noteworthy. Prime among these is that the index presents a single summative

measure of the amount of resilience a household exerts to withstand food insecurity shocks. The index is composite in the sense that it is computed based on weights of several variables, which are assets possessed by a household or are attributes characterising means of its livelihood and subsistence. The analysis followed in the footsteps of researchers who developed an asset-based index to determine predictors of socioeconomic welfare.

Another important attribute of the index is that it is established as good alternative for traditional measures of food insecurity, which are based on income or consumption, which are intrinsically prone to response or enumeration biases or other factors such as seasonality and weather conditions. Furthermore, the food insecurity resilience index generated using PCA was determined as capable to predict the probability of a socioeconomic and food insecurity outcome. It was also found to be capable of explaining the state of socioeconomic and livelihood deprivation and the likelihood of household vulnerability to food insecurity shocks; mainly due to poor resilience amongst different population groups. This finding is similar to that by (Moser & Felton 2007). Nevertheless, PCA is not without limitations. A noteworthy shortcoming of the method is that dimension reduction can only be achieved if the original variables were correlated. If the original variables were uncorrelated, PCA does nothing, except for ordering them according to their variance. Another limitation is that PCA is based only on the mean vector and the covariance matrix of the data.

Multiple Correspondence Analysis (MCA) led to generating a food insecurity resilience index based on weights or component loadings of the first extracted dimension. The technique succeeded to segregate the components that positively contributed to the index from those that contributed negatively. In addition, the technique was established to help in profiling or classifying population settings according to their amount of resilience. Furthermore, the MCA-based resilience index was determined to associate well with socioeconomic, typically per

capita consumption. However, due to the nature of data such model lacked a good fit. This problem is suspected to have emanated from sampling, data collection and the large number of variables, which could reduce the amount of variability represented by the extracted components.

Classification and Regression Trees (CART) analysis distinguished itself in being able to classify single overall best predictor of the food insecurity outcome out of potential candidates. CART analysis was also found to have good predictive ability and thus useful for supporting early warning and preparedness interventions. However, the technique is not without major drawbacks! These drawbacks mainly rest on presence of misclassification errors, which is a typical case of large surveys and use of secondary data collected for different purposes other than that of the study.

Univariate analysis applying univariate binary logistic regression confirmed that agriculture-dominated constructs of rural livelihoods boost coping with food insecurity in protracted crisis. Income from sale of farm crops and livestock activity was determined to even better enhance resilience of populations faced with food crisis. This information reinforces evidence for advocates of social protection for improving rural livelihoods through microcredits to rural farming communities. Such form of support rather than food aid promise mid-term to long-term solutions, since it supplements field farming activities, especially in situations of massive displacement. Of special importance was the finding that a latent variable could be generated from the predicted variables that showed significant associations with food insecurity outcomes and coping with food insecurity. Such variable could stand as a single measure of the amount of resilience a household sample exerts, and also to serve as single composite predictor of food insecurity outcome in crisis setting.

Considering issues with complexity of survey designs and multistage sampling, analysis explored the application of Survey Logistic to the data. This procedure uses the Pseudo-Maximum Likelihood estimation criterion and the generalized logit link function to incorporate sampling weights. Contrary to other regression modelling procedures, the method determined all seven variables included in the model as significant predictors of food consumption score. Analysis based on tests of hypothesis, Wald test and odds ratios indicated strong associations between each of the seven livelihood factors and food consumption score. Worthy of noting is the finding that throughout the analysis agriculture and livestock keeping were associated with food insecurity outcomes; which is consistent with other research in investigating factors of livelihood. However, model diagnosis revealed some fitness issues such as violation of the proportional odds assumption.

Like in health and other social outcomes, food security could have associated outcomes. Analysis investigated the possibility of joint outcomes that combine to pose risk of food insecurity, namely; incidence of coping with food insecurity and proportion of food consumption out of total household expenditure. It was determined that resorting to coping associated with food expenditure, thus it was important to examine a joint model to determine the relationship between independent predictors and the joint responses.

Analysis using the Joint Modelling determined that gender of household head and ownership of livestock showed non-significant relationship with the joint food insecurity outcome – adopting a coping strategy and increased food expenditure. This could indicate generalized vulnerability with rampant coping or due to response bias. The rest of the factors included in the model, i.e., *crop cultivation*, *fishing* and *main source of income* associated significantly with both adoption of coping strategy and increased food consumption expenditure. This led to the conclusion that these livelihoods aspects could be taken as predictors of the probability

of coping and increased food expenditure. Of essence was the finding that the Joint Model modified the results of the univariate models, in the form of significance probabilities based on tests of hypothesis.

From the preceding findings and conclusions it can be concluded that for the factor discriminant analysis, the Classification and Regression Trees combine both the rigour of selecting fewer and overall best predictor of food insecurity outcome. The simplicity of the procedure is also ascertained. However, if tested with fewer well selected variables, Principal Component Analysis does well in computing single index for assessing an outcome of food insecurity risk based on the first principal component (a latent variable), as well as for classifying food insecurity by population settings. For data models based on regression models, given that the data come from complex survey designs, the Survey Logistic Modelling analysis is recommended on the principles that the method factors in this aspect based on survey weights. Secondly, the procedure succeeded in confirming all the factors identified to be potential determinants of food insecurity outcomes as significant influential predictors.

However, each of the statistical procedures examined could have its own strengths and merits as well as limitations depending on the type of data, methods of collection and sampling. Oftentimes data from long questionnaires tend to be poor due to issues such as respondent and interviewer fatigue, monotony of questions and even misunderstanding of questions. Another drawback of the study is that it was based on only two dataset, which were from one-off periods rather than from longitudinal data collection. It might be worthwhile examining the techniques with similar cohorts of populations in order to account for inter-temporal variability. This is basically because resilience could be a function of durability over time.

A solution to overcome some of the limitations of the study cited above would be to use more controlled, simple and straight forward questionnaire designs that fit specific research

questions can result in more accurate, complete and cleaner data in terms of internal consistency and other elements. Indeed, some of the procedures such as multivariate analysis could be better used for longitudinal and panel analysis. If the issues limiting the analytical power of the methods could be resolved, some of them seem to have the capability of serving as useful tools in early warning and preparedness analysis.

In conclusion, this study being exploratory was able to discuss and establish evidence of the strength of each of the statistical procedures used in the analysis, as well as its limitations. However, what remains of high importance is what future avenues for research could be derived at the end of the exercise so that the findings in this work provide useful guide toward breakthroughs in the measurement and prediction of food insecurity resilience status and the likelihood of vulnerability, especially of populations in stressful food insecurity settings. The most striking observation from the findings that carefully selected livelihood enhancing factors such as sources of income and durable assets, rather than semi-durable assets seem to tend to yield strong resilience, as they were shown to associate strongly with food security outcomes, had better estimates factor loadings or coefficients.

It is, therefore, recommended that future studies aimed at probing resilience or assessing the risk of food insecurity in a population, explore the inclusion of these variables. Moreover, it is recommended that future research involving survey datasets should explore the survey logistics model since it accounts for complexity of survey designs and does well in identifying important predictors of the food insecurity outcomes. Meanwhile, it has transpired that analytical approaches of the multivariate types should be based on longitudinal studies of the same subjects (households). However, the limitations of each study based on the data collected should be noted and considered before selecting the appropriate method of analysis. Furthermore, it is recommended that future studies that aim at resilience and food insecurity

risk analysis may consider including of geographic coordinates when collecting data. This will facilitate spatial analysis and mapping of areas by level of resilience and by level of food insecurity risk.

REFERENCES

Abdi, H. & Valentine, D., 2007. Multiple Correspondence Analysis. *Encyclopedia of Measurement and Statistics*.

Adongo, J. & Deen-Swarray, M., 2006. *Poverty Alleviation in Rural Namibia through Improved Access to Financial Services*, Windhoek, Namibia.

African Union Commission, 2013. High-Level Declaration on Renewed Partnership for a Unified Approach to End Hunger in Africa by 2025. Available at: http://pages.au.int/endhunger/events/declaration-high-level-meeting.

Agresti, A., 2004. *Analysis of Ordinal Categorical Data*, New York: Wiley.

Agresti, A., 2002. *Categorical Data Analysis* 2nd ed., New York: John Wiley and Sons.

Akaike, H., 1987. Factor analysis and AIC. *Psychometrika*, 52, pp.317–332.

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki, eds. *Proceedings of the 2nd International Symposium on Information Theory*. Budapest: Akademiai Kiado, pp. 267–281.

Alinovi, L., Mane, E. & Romano, D., 2010. Measuring Household Resilience to food insecurity: Application to Palestinian Households. In R. Benedetti et al., eds. *Agricultural Survey Methods*. Chichester, UK: John Wiley & Sons, Ltd.

Archera, K.J., Lemeshow, S. & Hosmer, D.W., 2006. Goodness-of-fit tests for logistic regression models when data are collected using a complex sampling design. *Computational Statistics and Data Analysis*, 51(2007), pp.4450–4464.

Asian Development Bank, 2008a. *Research Study on Poverty Specific Purchasing Power Parities for Selected Countries in Asia and the Pacific*, Manila: Asian Development Bank.

Asian Development Bank, 2008b. *Research Study on Poverty-Specific Purchasing Power Parities for Selected Countries in Asia and the Pacific*, Manila: Asian Development Bank.

Asselin, L.-M., 2002. *Multidimensional poverty: Composite indicator of multidimensional poverty*, Lévis, Québec: Institut de Mathématique Gauss.

Asselin, L.-M. & Vu Tuan, A., 2008. Multidimensional Poverty and Multiple Correspondence Analysis.

Balen, J. et al., 2010. Comparison of two approaches for measuring household wealth via an asset-based index in rural and peri-urban settings of Hunan Province, China. *Emerging Themes in Epidemiology*, 7(7). Available at: http://www.ete-online.com/content/7/1/7.

Barret, C.B. & Heisey, K.C., 2002. How effectively does multilateral food aid respond to

fluctuating needs? *Food Policy*, 27(5-6), pp.477–491. Available at: http://www.sciencedirect.com/science/article/pii/S0306919202000507.

Barret, C.B. & Maxwell, D.G., 2005. *Food Aid After Fifty Years: Recasting Its Role (Priorities for Development Economics)*, New York: Routeledge.

Baullch, B. & Hoddinott, J., 2000. Economic mobility and poverty dynamics in developing countries. *Journal of Development Studies*, 36(6), pp.1–24.

Benzécri, J.P., 1992. *Correspondence analysis handbook*, New York: Marcel Dekker.

Beswick, S., 2001. "We are bought like clothes": The war over polygyny and levirate marriage in South Sudan. *Northeast African studies*, 8(2), pp.35–62.

Binder, D.A., 2003. On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, pp.279–292.

Bishop, S., Catley, A. & Hassan, H.S., 2008. Livestock and livelihoods in protracted crisis: The case of southern Somalia. In L. Alinovi, G. Hemrich, & L. Russo, eds. *Beyond Relief: Food Security in Protracted Crises*. FAO and Rugby: Practical Action Publishing, pp. 127–153.

Bolker, B.M. et al., 2008. No Title. *Trends in Ecology and Evolution*, 23(3), pp.127–135.

Booysen, F. et al., 2005. Using an Asset Index to Assess Trends in Poverty in Seven Sub-Saharan African countries. In *Conference on Multidimensional Poverty*.

Booysen, F., 2002. Using Demographic and Health Surveys to measure poverty: An application to South Africa. *Journal for Studies in Economics and Econometrics*, 26(3), pp.53–70.

Breiman, L. et al., 1984. *Classification and regression trees*, Monterey, Calif., U.S.A.: Wadsworth, Inc.

Breslow, N.E. & Clayton, D.G., 1993. Approximation Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 8(421), pp.9–25. Available at: http://www.public.iastate.edu/~alicia/stat544/Breslow and Clayton 1993.pdf.

Buckland, S.T., Burnham, K.P. & Augustin, N.H., 1997. Model Selection: An Integral Part of Inference. *Biometrics*, 53(2), pp.603–618.

Burnham, K.P. & Anderson, D.R., 2002. *Model selection and Multimodel Inference: A practical Information-Theoretic Approach.* Second Edi., New York: Spinger-Verlag Inc.

Cao, Y., Larsen, D.P. & Thorne, R.S.J., 2001. Rare species in multivariate analysis s for bioassessment: some considerations. *Journal of the North American Benthological Society*, 20, pp.144–153.

Chambers, R.L. & Skinner, C.J., 2003. *Analysis of Survey Data*, Chichester, UK: Wiley.

Clausen, S.E., 1998. *Applied Correspondence Analysis: An Introduction*, Thousand Oaks, CA:

Sage.

Collett, D., 2003. *Modelling Binary Data* 2nd ed., Boca Raton, FL: Chapman and Hall.

Committee on World Food Security, 2015. *Framework for Action for Food Security and Nutrition in Protracted Crises*, Rome.

Cox, D.R. & Snell, E.J., 1989. *The Analysis of Binary Data* 2nd ed., London: Chapman and Hall.

Deation, A., 1997. *The analysis of household surveys – a microeconomic approach to development policy*, Baltimore & London: Johns Hopkins University Press.

Department for International Development, 1999. Sustainable Livelihoods Guidance Sheets. Available at: http://www.efls.ca/webresources/DFID_Sustainable_ livelihoods_guidance_sheet.pdf [Accessed January 1, 2015].

Dray, S., Chessel, D. & Thioulouse, J., 2003. Co-Inertia Analysis and the Linking of Ecological Data Tables. *Ecology*, 84, pp.3078–3089.

EAO, 2015. *The State of Food Insecurity in the World (SOFI)*, Rome: FAO. Available at: file:///F:/Papers/GLMM/The State of Food Insecurity in the World 2015.pdf.

Elasha, B.O. et al., 2005. *Sustainable livelihood approach for assessing community resilience to climate change: case studies from Sudan*,

Ene, M. et al., 2014. Multilevel Models for Categorical Data using SAS® PROC GLIMMIX: The Basics. *SAS Global Forum 2015*. Available at: http://analytics.ncsu.edu/sesug/2014/SD-13.pdf [Accessed June 3, 2015].

Falkingham, J. & Namzie, C., 2001. *Identifying the poor: A critical review of alternative approaches*, London: London School of Economics.

Famine Early Warning Systems Network, 2007. Southern Sudan Food Security Watch, February 9, 2007.

FAO, 2012a. *Building resilience for adaptation to climate change in the agriculture sector* A. Meybeck et al., eds., Rome: FAO.

FAO, 2008. *Food Security Update (Jan-March 2008)*, Juba.

FAO, 2012b. *Resilience Index: Measurement and Analysis model*, Rome: FAO.

FAO, 2002. *Rural development: some issues in the context of the WTO negotiations on agriculture*, Rome. Available at: ftp://ftp.fao.org/docrep/fao/004/Y3733E/Y3733E00.pdf.

FAO, 2003. Strengthening Coherence in FAO's Initiatives to Fight Hunger. *FAO Corporate Document Repository*. Available at: http://www.fao.org/docrep/MEETING/007/J0710E.HTM [Accessed May 26, 2015].

FAO, IFAD & WFP, 2013. *The State of Food Insecurity in the World 2013. The multiple dimensions of food security*, Rome: FAO.

Filmer, D. & Pritchett, L., 2001. Estimating wealth effects without expenditure data – or tears: an application of educational enrolment in states of India. *Demography*, 38(1), pp.115–132.

Filmer, D. & Pritchett, L.Z., 1998. *Estimating wealth effects without income or expenditure data - or tears: Educational enrollment in India*, Washington, D.C.

Filmer, D. & Scott, K., 2008. *Assessing asset indices*, Washington, D.C.

Flores, M., 2007. Responding to food insecurity: Could we have done it better? In A. Pain & J. Sutton, eds. *Reconstructing Agriculture in Afghanistan*. Rome: Practical Action Publishing.

Food Security Analysis Unit (FSAU), 2006. *Integrated Humanitarian and Food Security Phase Classification*, Nairobi.

Gardiner, J.C., 2013. Joint Modeling of Mixed Outcomes in Health Service Research. In *SAS Global Forum 2013*. Cary, NC: SAS Institute Inc. Available at: http://support.sas.com/resources/papers/proceedings13/435-2013.pdf.

Gitz, V. & Maybeck, A., 2012. Risks, vulnerabilities and resilience in a context of climate change. In A. Maybeck et al., eds. *Building Resilience for Adaptation to Climate Change in the Agricultural Sector*. Rome: OED Publishing.

Gordon, L., 2013. Using Classification and Regression Trees (CART) in SAS® Enterprise Miner™ for Applications in Public Health In: Data Mining and Text Analytics. In *SAS Global Forum 2013*. Lexington, KY.

von Grebmer, K. et al., 2011. *2011 Global Hunger Index: The Challenge of Hunger: Taming Price Spikes and Excessive Food Price Volatility*, Bonn, Washington, DC, and Dublin.

Greenacre, M.J., 1993. *Correspondence analysis in practice*, London: Academic Press.

Greenacre, M.J., 2000. Correspondence analysis of square asymmertric matrices. *Applied Statistics*, 49, pp.297–310.

Greenacre, M.J., 1984. *Theory and Applications of Correspondence Analysis*, London: Academic Press.

Greenacre, M.J. & Blasius, J., 2006. *Multiple corresponse analysis and related methods*, Boca Raton, FL: Chapman and Hall.

Gwatkin, D.R. et al., 2000. *Socioeconomic Differences in Health, Nutrition and Population in Turkey*, Washington, D.C.: The World Bank.

Hancioglu, A., 2002. Performance of alternative approaches for identifying the relatively poor and linkages to reproductive health. In *Reproductive Health, Unmet Needs, and Poverty:*

Issues of Access and Quality of Services. Bangkok, Thailand: CICRED.

Heeringa, S.., West, B.. & Berglund, P.A., 2010. *Applied Survey Data Analysis*, Boca Raton, FL: Chapman and Hall.

Herrero, C., Martínez, R. & Villar, A., 2010. *Improving the Measurement of Human Development*, New York.

Hosmer, D.W. & Lemeshow, S., 2000. *Applied Logistic Regression* 2nd ed., New York: John Wiley and Sons, Inc.

Hulme, D. & Shepherd, A., 2003. Conceptualizing chronic poverty. *World Development*, 31(3), pp.403–423.

IBM Corporation, 2015. IBM SPSS Amos 23.0.0 (Build 1607). Available at: http://amosdevelopment.com.

IBM SPSS, 2013. *IBM SPSS Statistics Version 22*, New York: IBM Corporation.

IFAD, WFP & FAO, 2013. *The State of Food Insecurity in the World 2013*, FAO.

IFPRI, 2014. Building Resilience for Food and Nutrition Security: Highlights from the 2020 Conference. *IFPRI 2020 Policy Consultations & Conference*. Available at: http://www.ifpri.org/sites/default/files/publications/2020resilience_synopsis.pdf.

Johnson, R.A. & Wichern, D.W., 2007. *Applied multivariate statistical analysis*, New Jersy: Pearson/Prentice Hall.

Kolev, N. & Paiva, D., 2009. Copula-based regression models: A survey. *Journal of Statistical Planning and Inference*, 139(11), pp.3847–3856. Available at: http://www.sciencedirect.com/science/article/pii/S0378375809001517.

Kumar, K., 1989. Indicators for measuring changes in income, food availability and consumption, and the natural resource base. In A.I.D. Program Design and Evaluation Methodology.

Lange, K., 1999. *Numerical Analysis for Statisticians*, New York: Springer-Verlag.

Lehtonen, R. & Pahkinen, E.J., 1995. *Practical Methods for Design and Analysis of Complex Surveys*, Chichester, UK: John Wiley and Sons Ltd.

Lemon, C. et al., 2003. Classification and Regression Tree Analysis in Public Health: Methodological Review and Comparison With Logistic Regression. *Ann Behav Med*, 26(3), pp.172–181.

Littell, C.R. et al., 2006. *SAS System for Mixed Models* Second., Cary, NC: SAS Institute Inc.

Liverpool, L.S.O. & Winter-Nelson, A., 2010. *Asset Versus Consumption Poverty and Poverty Dynamics in the Presence of Multiple Equilibria in Rural Ethiopia*, Washington, D.C.

Loh, W.-Y., 2011. Classification and regression trees. WIREs Data Mining and Knowledge Discovery.

Lokosang, L., Ramroop, S. & Hendriks, S.L., 2010. Establishing a robust technique for monitoring and early warning of food insecurity in post-conflict Southern Sudan using Ordinal Logistic Regression. *Agrekon*, 50(4), pp.101–130. Available at: http://www.tandfonline.com/doi/abs/10.1080/03031853.2011.617902.

Lokosang, L., Ramroop, S. & Zewotir, T., 2014. Indexing household resilience to food insecurity shocks: The case of South Sudan. *Agrekon:*, 53(2), pp.137–159.

Maltsoglou, I., *Household Expenditure on Food of Animal Origin: A Comparison of Uganda, Vietnam and Peru*, Available at: http://www.fao.org/ag/againfo/programmes/en/pplpi/docarc/execsumm_wp43.pdf.

Maxwell, D. & Caldwell, R., 2008. *The Coping Strategy Index: Field Methods Manual* Second., Atlanta, GA: Cooperative for Assistance and Relief Everywhere, Inc. (CARE). Available at: http://www.seachangecop.org/sites/default/files/documents/2008 01 TANGO - Coping Strategies Index.pdf.

Maxwell, D.G., 1995. *Measuring Food Insecurity: The frequency and severity of "coping strategies,"* Washington, D.C. Available at: http://ageconsearch.umn.edu/bitstream/42669/2/dp08.pdf.

Maxwell, D.G. et al., 2003. The Coping Strategy Index: a tool for rapidly measuring food security and the impact of food aid programmes in emergencies. In *International Workshop on Food Security in Complex Emergencies: building policy frameworks to address longer-term programming challenges*. Rome: FAO.

Maxwell, S., 1996. Food security: a post-modern perspective. *Food Policy*, 21, pp.155–170. Available at: http://dx.doi.org/10.1016/0306-9192(95)00074-7.

McCullagh, P., 1980. Regression Models for Ordinal Data (with discussion). *J. R. Statist. Soc. Soci.*, 42(109), p.42.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models*, London: Chapman and Hall.

McCullagh, P. & Nelder, J.A., 1989. *Generalized Linear Models, 2nd Edition*, London: Chapman and Hall.

McCulloch, C.E. & Searle, S.R., 2001. *Generalized Linear Mixed Models*, New York: Wiley.

McKenzie, D., 2004. Measuring Inequality with Asset Indicators. *Journal of Population Economics*, 18(1), pp.229–260.

Meulman, J.J., 1996. Fitting a distance model to homogeneous subsets of variables: Points of view analysis of categorical data. *Journal of Classification*, 13, pp.249–266.

Morel, G., 1989. Logistic Regression under Complex Survey Designs. *Survey Methodology*, 15, pp.203–223.

Morel, J.G., 1989. Logistic Regression under Complex Survey Designs. *Survey Methodology*, 15, pp.203–223.

Morris, S. et al., 1999. *Validity of rapid estimates of household wealth and income for health surveys in rural Africa*, Washington, D.C.

Moser, C. & Felton, A., 2007. *The construction of an asset index measuring asset accumulation in Ecuador*, Washington, D.C.: The Brookings Institute.

Mousseau, F., 2005. *Food AID or Food SOVEREIGNTY? Ending World Hunger in Our Time*, Oakland, CA: The Oakland Institute.

National Bureau of Statistics, 2010. *Poverty in Southern Sudan: Estimates from NBHS 2009*, Juba.

National Bureau of Statistics, 2008. *Sudan Household Health Survey 2006*, Juba.

Nelder, J.A. & Wedderburn, R.W.M., 1972. Generalized Linear Models. *J. R. Statist. Soc.*, 135(3), pp.370–84.

New Partnership for Africa's Development, 2009. *CAADP Pillar III Framework for Africa's Food Security (FAFS)*, Midrand, South Africa: NEPAD Secretariat. Available at: http://caadp.net/sites/default/files/documents/Resources/CAADP-guides-and-technical/CAADP Pillar III Framework for African Food Security_2009.pdf.

New Partnership for Africa's Development, 2003. *Comprehensive Africa Agriculture Development Programme*, Rome: FAO.

O'rourke, N., Hatcher, L. & Stepanks, E.J., 2005. *Step-by-Step approach to Using SAS® for Univariate & Multivariate Statistics* Second., Cary, NC: The SAS Institute®.

Olsson, U., 2002. *Generalized Linear Models. An Applied Approach*, LUND: Studentlitteratur.

Pantuliano, S., 2008. Responding to protracted crises: The principled model of NMPACT in Sudan. In L. Alinovi, G. Hemrich, & L. Russo, eds. *Beyond Relief: Food Security in Protracted Crises*. Rome and Rugby: Practical Action Publishing, pp. 25–63.

Pasteur, K., 2011. *From vulnerability to resilience: A framework for analysis and action to build community resilience*, Rugby: Practical Action Publishing.

Pradhan, M., 2000. *How many questions should be in a consumption questionnaire? Evidence from repeated experiment in Indonesia*, New York.

Prakongsai, P., 2006. An application of asset index for measuring household living standards in Thailand. In Bangkok, Thailand.

Rao, J.N.K., Wu, C.F.J. & Yue, K., 1992a. Some Recent Work on Resampling Methods for Complex Surveys. *Survey Methodology*, 18, pp.209–217.

Rao, J.N.K., Wu, C.F.J. & Yue, K., 1992b. Some Recent Work on Resampling Methods for

Complex Surveys. *Survey Methodology*, 18(209-217).

Ravallion, M., 1992. *Poverty Comparisons: A Guide to Concepts and Methods*, Washington, D.C.

Roberts, G., Rao, J.N.K. & Kumar, S., 1987a. Logistic Regression Analysis of Sample Survey Data. *Biometrika*, 74, pp.1–12.

Roberts, G., Rao, J.N.K. & Kumar, S., 1987b. Logistic Regression Analysis of Sample Survey Data. *Biometrika*, 74, pp.1–12.

Russo, L. et al., 2008. Food security in protracted crisis situations: Issues and challenges. In *Beyond Relief: Food Security in Protracted Crises*. Rugby: Practical Action Publishing, pp. 1–10. Available at: http://www.fao.org/3/a-a0778e.pdf.

Rutstein, S.O. & Kiersten, J., 2004. *The DHS Wealth Index*, Calverton, Maryland, USA.

Sahn, D.E. & Stifel, D., 2003. Exploring Alternative Measures of Welfare in the Absence of Expenditure Data. *Review of Income and Wealth*, 49, pp.463–89.

Sahn, D.E. & Stifel, D., 2000. Poverty Comparisons over Time and Across Countries in Africa. *World Development*, 28(12), pp.2123–2155.

Särndal, C.E., Swensson, B. & Wretman, J., 1992. *Model Assisted Survey Sampling*, New York: Springer-Verlag.

SAS Institute, 2011a. Knowledge Base: SAS/STAT(R) 9.22 User's Guide. *SAS/STAT(R) 9.22 User's Guide*. Available at: http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveylogistic_a0000000394.htm [Accessed June 9, 2015].

SAS Institute, 2011b. *SAS/STAT 9.3 Production GRIMMIX Procedure for Windows*, Cary, NC: SAS Institute Inc.

Schabenberger, O., 2004. *Introducing the GLIMMIX Procedure for Generalized Linear Mixed Models*, Cary, NC.

Schabenberger, O., 2005. *Introducing the GLIMMIX procedure for generalized linear mixed models*, Cary, NC: SAS Institute Inc.

Schafer, J., 2002. *Supporting livelihoods in situations of chronic conflict and political instability: Overview of conceptual issues*, Brighton, UK.

Scott, C. & Amenuvegbe, B., 1990. *Effect of Recall Duration on Reporting of Household Expenditures: An Experimental Study in Ghana*, Washington, D.C.

Sibrián, R., Ramasawmy, S. & Mernies, J., 2008. *Measuring Hunger at Subnational Levels from National Surveys Using the FAO Approach: Manual*, Rome. Available at: http://www.fao.org/fileadmin/templates/ess/documents/food_security_statistics/working_paper_series/WP005e.pdf.

Skinner, C.J., Holt, D. & Smith, T.M.F., 1989. *Analysis of Complex Surveys*, New York: John Wiley & Sons, Inc.

Slater, R. et al., 2015. *Coherence between agriculture and social protection: An analytical framework*, Rome.

Smith, L.C., Alderman, H. & Aduayom, D., 2006. *Food Insecurity in Sub-Saharan Africa: New Estimates from Household Expenditure Surveys. Research Report 146*, Washington, D.C.: International Food Policy Research Institute.

Snel, E. & Staring, R., 2001. Poverty, migration, and coping strategies: An introduction. *European Journal of Anthropology*, 38, pp.7–22. Available at: file:///C:/Users/sscu/Downloads/SOC-2001-002.pdf.

Staatz, J.M. & Dembélé, M.N., 2007. *Agriculture for Development in Sub-Saharan Africa*, Available at: https://openknowledge.worldbank.org/bitstream/handle/10986/9043/WDR2008_0037.pdf?sequence=1&isAllowed=y [Accessed June 27, 2015].

Stern, O., 2011. Women and Mariage in South Sudan. In F. Bubenzer & O. Stern, eds. *Hope, Pain & Patience: The lives of women in South Sudan*. Fanele: Jacana Media.

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

Tenenhaus, M. & Young, F.W., 1985. An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis, and other methods for quantifying categorical multivariate data. *Psychometrika*, 50, pp.91–119.

The United Nations, 2002. Millennium Development Goals. Available at: http://www.un.org/millenniumgoals/.

Tirivayi, N., Knowles, M. & Davis, B., 2013. *The interaction between social protection and agriculture: a review of evidence*, Rome: FAO. Available at: http://www.fao.org/3/a-i3563e.pdf.

United Nations, 2015. The Millennium Development Goals. Available at: http://www.un.org/millenniumgoals/ [Accessed May 20, 2015].

United Nations Development Programme, 2009. *Human Development Report 2009*, New York: Palgrave Macmillan.

United Nations Development Programme, 2010. *Human Development Report 2010: The Real Wealth of Nations - Pathways to Human Development*, New York.

Vyas, S. & Kumaranayake, L., 2006. Constructing socio-economic status. *Health Policy Plan*, 21(6), pp.459–468.

Walker, S.H. & Duncan, D.B., 1967. Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*, 54, pp.167–179.

Weiser, S.D. et al., 2009. Food Security Among Homeless and Marginally Housed Individuals Living with HIV/AIDS in San Francisco. *AIDS Behav.*, 13, pp.841–848.

Welthungerhilfe, International Food Policy Research Institute, . & Concern Worldwide, 2012. *Global Hunger Index 2012* C. von Oppeln et al., eds., Cologne, Germany: International Food Policy Research Institute.

Westreich, D., Lessler, J. & Funk, M.J., 2010. Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *J Clin Epidemiol*, 63(8), pp.826–833.

Williams, R., 2006. Generalized ordered logit/partial proportional odds models for ordinal dependent variables. *The Stata Journal*, 6(1), pp.58–82.

Wolfinger, R. & O'Connell, M., 1993. Generalized Linear Mixed Models: A Pseudo-Likelihood Approach. *Journal of Statistical Computation and Simulation*, 84(3-4), pp.233–243.

Wolter, K.M., 2007. *Introduction to Variance Estimation* Second., New York: Springer-Verlag.

World Bank, 2004. *Measuring living standards: household consumption and wealth indices*, Washington, D.C. Available at: http://siteresources.worldbank.org/INTPAH/Resources/Publications/Quantitative-Techniques/health_eq_tn04.pdf.

World Bank, 2008. *World Development Report 2008: Agriculture for development*, Washington, D.C.

World Bank & IMF, 2002. *Review of the Poverty Reduction Strategy Paper (PRSP) Approach: Early Experience with Interim PRSPs and Full PRSPs*, Washington, D.C.

World Food Programme, 2014. *South Sudan Food Security Monitoring, VAM Food Security Analysis, Round 13*, Juba.

World Food Programme, V.A. and M.B. (ODAV), 2008. *Food consumption analysis: calculation and use of the food consumption score in food security analysis*, Rome: WFP.

World Food Programme, V.A. and M.B. (ODAV), 2007. *Sudan: Southern Sudan Comprehensive Food Security and Vulnerability Analysis (CFSVA)*, Rome.

Yohannes, Y. & Hoddinott, J., 1999. *Classification and Regression Trees: An Introduction*, Washington, D.C.

Yohannes, Y. & Webb, P., 1999. *Classification and Regression Trees, CART: A User's Manual for Identifying Indicators of Vulnerability to Famine and Chronic Food Insecurity*, Washington, D.C.: International Food Policy Research Institute.

APPENDIX: SAS CODE FOR ANALYSING DATA

A. The Binary Logistic Regression Procedure

```
ods html;

ods graphics on;

DATA fsms ;

     SET mylib.fsms4 ;

RUN;


PROC FORMAT;

VALUE sexHHH 1='Male' 2 = 'Female';

VALUE ageHHH 1='(< 17' 2='18-60' 3='>60';

VALUE HHsize 1='<=3' 2='4-6' 3='7+';

VALUE CulCrops 1='Yes' 2='No';

VALUE Livestock 1='Yes' 2='No';

VALUE Fishing 1='Yes' 2='No';

VALUE IncomeSce 1='Sale of agric crops' 2='Sale of livestock products'
3='Employment/labour' 4='Petty trading' 5='Other';

RUN;

TITLE 'Stepwise Regression on Coping with Food Insecurity Data';


PROC LOGISTIC data=fsms outest=betas covout;

CLASS agehhh sexhhh HHsize CulCrops Livestock Fishing IncomeSce;

MODEL coping = agehhh sexhhh hhsize CulCrops Livestock Fishing IncomeSce

                / selection=stepwise

                  slentry=0.3

                  slstay=0.35
```

```
                details

                lackfit;

   OUTPUT OUT=pred p=phat lower=lcl upper=ucl

          predprob=(individual crossvalidate);

RUN;

PROC PRINT DATA=betas;

   TITLE2 'Parameter Estimates and Covariance Matrix';

RUN;

PROC PRINT DATA=pred;

   TITLE2 'Predicted Probabilities and 95% Confidence Limits';

RUN;

ods graphics off;

ods html close;
```

B. The Survey Logistic Procedure

```
ods html;

ods graphics on;

DATA myfsms ;

     SET mylib.fsms ;

RUN;

PROC FORMAT;

    VALUE sexHHH 1='Male' 2 = 'Female';

    VALUE ageHHH 1='(< 17' 2='18-60' 3='>60';

    VALUE HHsizecat 1='<3' 2='4-6' 3='7+';

    VALUE culcrops 1='Yes' 2='No';

    VALUE Livestock 1='Yes' 2='No';

    VALUE Fishing 1='Yes' 2='No';
```

```
    VALUE Livelihood 1='Agriculture' 2='Livestock & animal products'
3='Salaries & wages' 4='Other';

    VALUE fcgroups 1='Acceptable' 2='Borderline' 3='Poor';

RUN;

PROC SURVEYLOGISTIC DATA=myfsms TOTAL = 3692;

STRATUM clusterno /LIST ;

CLASS agehhh sexhhh hhsizecat culcrops Livestock Fishing Livelihood / PARAM
= reference ;

MODEL fcgroups = agehhh sexhhh hhsizecat culcrops Livestock Fishing
Livelihood / LINK =glogit ;

WEIGHT HHLZ_weight ;

RUN;

ods graphics off;

ods html close;
```

C. The Generalized Linear Mixed Model

```
      ods html;

      ods graphics on;

      DATA fsmsg ;

           SET mylib.fsms5 ;

      RUN;

      PROC FORMAT;

      VALUE State 1='WES' 2='EES' 3='Jonglei' 4='Lakes' 5='UNS'

                  6='WBS' 7='NBS' 8='Warrap' 9='CES' 10='Unity' ;

      VALUE sexHHH 1='Male' 2 = 'Female';

      VALUE ageHHH 1='(< 17' 2='18-60' 3='>60';

      VALUE HHsize 1='<3' 2='4-6' 3='7+';

      VALUE culcrops 1='Yes' 2='No';

      VALUE Livestock 1='Yes' 2='No';

      VALUE Fishing 1='Yes' 2='No';

      VALUE IncomeSce 1='Yes' 2='No';
```

```
RUN;


PROC GLIMMIX data=fsmsg ;
CLASS clusterno agehhh sexhhh hhsize culcrops Livestock Fishing
IncomeSce;
MODEL fcgroups (event=last) = sexhhh agehhh hhsize culcrops Livestock
Fishing IncomeSce
                / S DIST=multinomial LINK=cumlogit OR;
        OUTPUT OUT=fsmsg_pred pred=p resid=r;
RUN;


ods graphics off;
ods html close;
ods html;
ods graphics on;
PROC GLIMMIX data=fsmsg METHOD=QUAD;
CLASS clusterno sexHHH ageHHH culcrops Livestock Fishing IncomeSce;
MODEL fcgroups = SexHHH AgeHHH HHSize culcrops Livestock Fishing
IncomeSce
              / CL DIST=multinomial LINK=cumlogit SOLUTION ODDSRATIO
(DIFF=FIRST LABEL);
RANDOM intercept / SUBJECT=clusterno TYPE=VC CL;
RUN;


proc sort data=solr ; by estimate;
data solr; set solr;
    length clusterno $3;
    obs  = _n_;
    clusterno = left(substr(Subject,6,3));
run;
proc sgplot data=solr;
   scatter x=obs y=estimate /
         markerchar  = clusterno
         yerrorupper = upper
         yerrorlower = lower;
   xaxis grid label='Cluster Number' values=(1 25 50 75 100 125 150);
   yaxis grid label='Predicted Cluster Effect';
run;
```

```
ods graphics off;

ods html close;
```

D. The Joint Modelling Procedure

```
ods html;

ods graphics on;

DATA fsmsx ;

     SET mylib.fsms5 ;

RUN;

PROC FORMAT;

VALUE sexHHH 1='Male' 2 = 'Female';

VALUE ageHHH 1='(< 17' 2='18-60' 3='>60';

VALUE HHsize 1='<3' 2='4-6' 3='7+';

VALUE culcrops 1='Yes' 2='No';

VALUE Livestock 1='Yes' 2='No';

VALUE Fishing 1='Yes' 2='No';

VALUE IncomeSce 1='Yes' 2='No';

VALUE Coping 0 ='No coping' 1='Coping';

RUN;

PROC GENMOD data=fsmsx ;

CLASS agehhh sexhhh hhsize culcrops Livestock Fishing IncomeSce;

MODEL foodexp = agehhh sexhhh hhsize culcrops Livestock Fishing
IncomeSce

                / DIST=normal LINK=log ;

RUN;

PROC GLIMMIX data=fsmsx METHOD=quad ODDSRATIO ASYCOV HESSIAN;

CLASS clusterno sexhhh agehhh hhsize culcrops Livestock Fishing
incomesce;
```

```
MODEL coping (event=last) = sexhhh agehhh hhsize culcrops Livestock
Fishing IncomeSce / CL

       DIST=binary LINK=logit SOLUTION ODDSRATIO (DIFF=FIRST LABEL);

RANDOM intercept / subject=clusterno;

RUN;

DATA fsmsJ;

      LENGTH dist $6;

      SET mylib.fsms5 ;

      response = foodexp;

      dist     = "Normal";

      OUTPUT;

      response = (coping=1);

      dist     = "Binary";

      OUTPUT;

      KEEP clusterno sexhhh agehhh hhsize culcrops Livestock

            Fishing IncomeSce response dist;

   RUN;

PROC GLIMMIX DATA=fsmsJ METHOD=quad ODDSRATIO ASYCOV HESSIAN;

      CLASS clusterno agehhh hhsize culcrops Livestock Fishing
IncomeSce dist;

      MODEL response(event='1') = dist dist*agehhh dist*hhsize
dist*culcrops

               dist*Livestock dist*Fishing dist*IncomeSce /

      NOINT s dist=byobs(dist);

        RANDOM intercept / subject=clusterno ;

RUN;

PROC GLIMMIX DATA=fsmsJ METHOD=rspl ASYCOV;

      CLASS clusterno agehhh hhsize culcrops Fishing livestock
IncomeSce dist;
```

```
      MODEL response(event='1') = dist dist*agehhh dist*hhsize
dist*culcrops dist*livestock

                dist*fishing dist*IncomeSce /

      NOINT s dist=byobs(dist);

        RANDOM _residual_ / subject=clusterno type=vc;

        OUTPUT OUT=jointm_pred pred=p resid=r;

RUN;

ods graphics off;

ods html close;
```