Analysis of a Binary Response: An Application to Entrepreneurship Success in South Sudan

James Lemi John Stephen Lugga

Thesis Submitted in Fulfilment of an Academic Requirement for Degree of Master of Science in Statistics

School of Mathematics, Statistics and Computer Science

University of Kwa Zulu Natal

South Africa

2012

Abstract

Just over half (50.6%) of the population of South Sudan lives on less than one US Dollar a day. Three quarters of the population live below the poverty line (NBS, Poverty Report, 2010). Generally, effective government policy to reduce unemployment and eradicate poverty focuses on stimulating new businesses. Micro and small enterprises (MSEs) are the major source of employment and income for many in under-developed countries. The objective of this study is to identify factors that determine business success and failure in South Sudan. To achieve this objective, generalized linear models, survey logistic models, the generalized linear mixed models and multiple correspondence analysis are used. The data used in this study is generated from the business survey conducted in 2010. The response variable, which is defined as business success or failure was measured by profit and loss in businesses. Fourteen explanatory variables were identified as factors contributing to business success and failure. A main effect model consisting of the fourteen explanatory variables and three interaction effects were fitted to the data. In order to account for the complexity of the survey design, survey logistic and generalized linear mixed models are refitted to the same variables in the main effect model. To confirm the results from the model we used multiple correspondence analysis.

Declaration of Authorship

This thesis and the research work done therein was carried out in the School of Mathematics, Statistics and Computer Science, University of Kwa Zulu-Natal, Pietermaritzburg Campus. The research work contains the original work of the author and has not been submitted in any form for any degree or diploma to any other University. Where the work of others has been quoted, it is duly acknowledged and referenced in the bibliography.

Mr. James Lemi John Stephen Lugga	Prof. Temesgen Zewotir
Signed:	Signed: .....
Date	Date .....

Dedication

This work is dedicated to Keji James Lemi and the Family

Acknowledgements

My sincere gratitude to my supervisor, Professor Temesgen Zewotir for his supervision, guidance, advice and assistance in all the stages of this study. Special thanks to the Staff and Colleagues of the School of Mathematics, Statistics and Computer Science for their moral and academic support through-out the period of my study. To mention only a few; Laila Barnaba; Art Luanda; Robert Mutwiri; Oscar Ngesa; Tsepang; Sifiso; Thulani; Aminata; Lea; Asnake; Dawid; Anisa; and Olina among others.

Special gratitude also goes to the management of the South Sudan National Bureau of Statistic for providing me the opportunity to acquire a postgraduate qualification especially when the country is in need of qualified cadres. Special thanks to my colleagues at the office Gennaro Joseph, Yacob Walla, Alex Tiangwa, and Vinayak for helping me with the dataset.

Lastly, I would like in a very special way to acknowledge the moral support of my family and everyone who has contributed in one way or an other to my success.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The South Sudan economy is expected to grow substantially in the next few years following independence. Most of this growth is likely to take place through the private business sector. Given a secure and peaceful business environment, there will be plenty of investment opportunities. Private sector development is an important engine for sustainability of economic growth (Bray, 2007). Private business enterprises enable the creation of new jobs, as well as generating income that assists in improving standards of living. Thus, policies that encourage and support the development of private business enterprise are very crucial. There is no standard definition of micro, small, medium, or large enterprise developed in South Sudan. Whereas there are several international definitions of small and medium enterprises, they vary from country to country. However, according to the Statistics Division of the United Nations Department of Economic and Social Affairs, private business enterprise can be classified as micro, small, medium, and large enterprises (Abor and Quartey, 2010). If an enterprise is composed of $1-9$ employees and has an annual turnover of 2 million Euros, it is regarded as a micro enterprise. An enterprise with 10 to 49 employees and an annual turnover of $2-10$ million Euros is regard as a small enterprise, whereas an enterprise with 50 to 249 employees and an annual turnover of $10-50$ million Euros is classified as a medium enterprise. A large enterprise is an enterprise with more than 250 employees and an annual turnover in excess of 50 million Euros. South Sudan's enterprise classification is solely with respect to the number of employees and not to the annual turnover. This is because

businesses in South Sudan have low turnovers compared to the United Nations' standards.

Understanding which factors lead to business success and which lead to failure is a primary, and as yet unfulfilled, purpose of business research (Rogoff *et al*, 2004). Grunert (1994) remarked that almost all scientific research in business administration is concerned with understanding what makes some businesses more successful than others. Finding out what influences an enterprise to grow is obviously very important for policy makers, investors, and advisers, as well as for business owners. Although there has been much research and commentary in this field, no single theory that adequately explains the interplay of all the factors influencing business success and growth has been developed (Longenecter *et al*, 2006 and Stokes and Wilson, 2010).

In order for businesses to survive and successfully make a profit, a conducive business environment is required, in addition to effective government policy, provision of security, and property rights (economic freedom). Where democracy and good governance are not in place, businesses are always hampered from growth, and the attraction of investors is discouraged, as such businesses in such an environment fail to prosper (Bray, 2007). Lack of empirical information on business performance is also a problem, in that investors may not be able to make decisions as to where to invest and how to invest. Despite poor infrastructure, inadequate services, and some level of insecurity in some parts of South Sudan, micro and small enterprise performance must be considered important as it forms a crucial part of the economy regarding creating new jobs and increasing trade performance. Growth is important for poverty reduction. For instance, the poverty line in South Sudan was estimated at 72.9 South Sudanese pounds (31.5 USD) per person per month. Just over half of the population (50.6%) was found to fall below the poverty line or/to live on less than one dollar per day (National Bureau of Statistics, 2010). The need for private business development is therefore crucial to assist in the process of poverty reduction in South Sudan.

Positive economic growth can be attained when business in the private sector grows and prospers. The Government of South Sudan and development partners such as the Multi Donor Trust Fund, the United Nations Development Programme (UNDP), and other partners, are striving to support private sector development. For the Government, business

community, and other stakeholders to facilitate and oversee business growth, they may need empirical studies (statistical findings) on factors related to business performance in order to provide vital information that can be used to form sound policies. However, such studies are limited. This study is therefore an attempt to investigate factors that determine business success and failure. It is aimed at highlighting the aspects needed to support successful business investment in the country. Hence, this study's results will assist policy makers on the profile of successful and unsuccessful micro and small business practice in South Sudan. It will also act as a baseline for future studies on business survey in the South Sudan. Besides identifying the factors that contribute to business success or failure, we develop a predictive model for factors that determine business success or failure. In other words the study aims to provide valid baseline information about the determinants of business success or failure, and to recommend appropriate policies regarding how to mitigate the factors of failure of businesses.

Micro and small business success or failure was measured using profit and loss over the period of one year. The profit and loss concept was driven from the theory of profit maximization from the theory of the firm. The total profit and total cost approach was considered. Thus, profit or loss was calculated from the difference between business total income and total cost over the period of one year. Therefore businesses that gained a profit are considered successful, whereas those that sustained losses are considered to have failed. Some factors were investigated so as to examine whether they have an influence on the success or failure of micro and small businesses in South Sudan.

The study is organised as follows Chapter 1 gives a general introduction about the study problem. In Chapter 2 we review related literature on the determinants of success and failure of micro and small enterprises. Chapter 3 focuses on the description of the data and survey methodologies. Chapter 4 deals with the basic statistical approaches of binary response analysis, called logistic regression. The results of the logistic regression are also discussed in Chapter 4. The extension of logistic regression to survey data, and random effect modelling, are discussed in Chapters 5 and 6. The application of correspondence analysis is given in Chapter 7. The study conclusion and recommendations are discussed in Chapter 8.

Chapter 2

Literature Review on the Theory of the Determinants of Business Success

A firm or business is referred to as an institution or organization that is engaged in the trade of goods and services to consumers. Businesses are mostly privately owned and operate to earn a profit (Spulber, 2009).

The neoclassical theory of the firm assumes that all the decisions of the firm are aimed at profit maximization and that firms maximize profits by producing outputs or goods and services at the point at which marginal revenue equals marginal cost. In the short run, however, firms are subject to diminishing returns (Varian, 2010). During this period the capital is fixed, therefore marginal cost is upwardly sloping after diminishing returns set in (Varian, 2010). The argument of this theory is that, the main objective of the business is profit maximization and that the ultimate goal is to make a decision on price in order to maximize market share. This is because in the long run, businesses hope that an increase in market share will enable higher monopoly power and therefore higher profits in the future (Varian, 2010). Another important factor in the neoclassical theory is the "separation theorem". According to this theorem, the objective of the firm is separate from the objectives of its consumers (Spulber, 2009). That is to say the idea that businesses maximize profit is based upon the objectives of its consumers and not up on those of the firm. In this context, consumers choose to purchase their most preferred consumption bundles within the limits

of their budget, taking market prices as fixed and not depending on firms' input and output choices. Generally, the neoclassical separation theorem makes three assertions: first, firms maximize profits; second, firms generate gains from trade; and third, firms' decisions are separate from consumer decision (Spulber, 2009). The neoclassical theory's main weakness is that it ignores firms' principle agent problem (Kantarelis, 2007). That is that, the owners may wish to maximize profits, but the employees may not, and as a result, ignore the agency problem for transaction costs due to conflict between owners and suppliers of inputs in the market system. Another weakness of the neoclassical theory is that it does not allow firm evolution (Kantarelis, 2007). Entrepreneurship is then seen as a fundamental and important part of modern economics and social life, which overcome the neoclassical theory's weaknesses (Stokes and Wilson, 2010).

Entrepreneurship is referred to as the manifest ability and willingness of individuals, on their own and as a team within and outside an existing organization to perceive and to create new economic opportunities (new products, new production methods, and new organizational schemes) as well as to introduce their ideas to the market, in the face of uncertainty and other obstacles, by making decisions on location and the use of resources (Stokes and Wilson, 2010). Entrepreneurship theory explains: about the environment within which entrepreneurship occurs; the people engaged in the entrepreneurship; entrepreneurship behaviours displayed by the entrepreneurs; the creation of organizations by the entrepreneurs; opportunities identified and explored; innovation; assuming risk at personal, organizational and even societal levels; and adding value for the entrepreneur and society (refer to Stokes and Wilson, 2010 for more details).

Generally, there are internal and external factors that influence business performance. The internal factors are mainly business and entrepreneur characteristics. The external factors are the socio-economic setup of the Country. We discuss these factors briefly in the sequel. Figure 2.1 gives the summary of the scheme of the determinants of business success/failure.

Size: One of the characteristics of businesses is size. Rogers (2004), findings on the relation between firm size and innovation, suggests that large firms have stronger cash flows

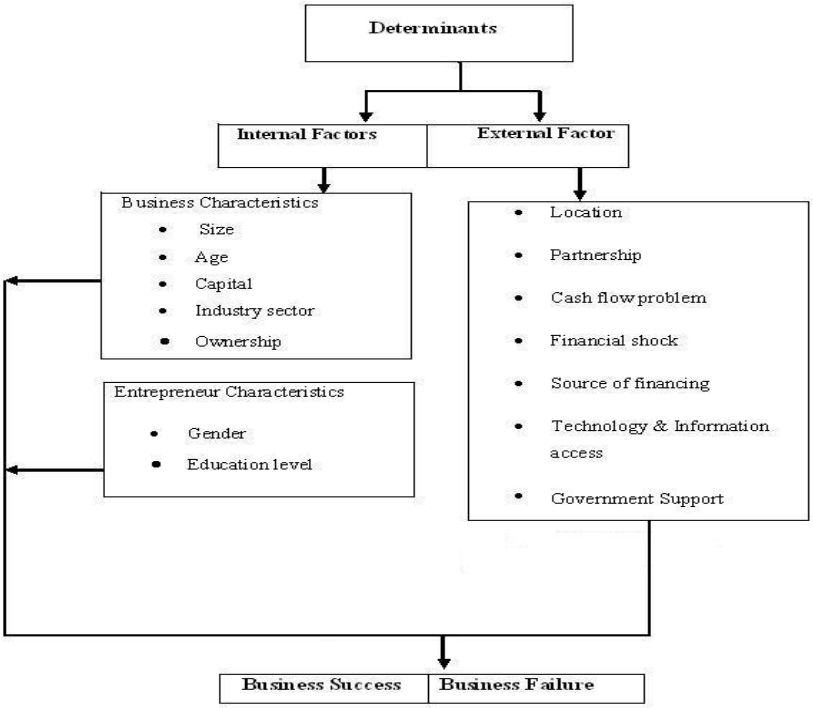

Figure 2.1: Analytical Framework of the Determinants of Business Success or Failure

to fund innovation compared to small firms. The argument is that large businesses may be better than small businesses in terms of assets that can be used as collateral for loans. Bartelsman *et al* (2005) on the other hand found that there was a significant heterogeneity of business' dynamics. This heterogeneity was manifested in large disparities in business size, business growth, and business productivity performance (Bartelsman *et al*, 2005).

Age: Business age plays a significant role in understanding business survival and growth. Studies have found that duration of firm operation was significantly related to business success (Kristianse *et al*, 2003). The finding asserted that older firms tend to gain more experience in business than new firms and thus foster their success and growth, that is to say that new (young) firms are more vulnerable to exit than old firms. The argument is that among small firms, older firms grow faster than younger firms, whereas among the larger firms there is a tendency for growth to decline with age (Papadaki and Chami, 2007). Coad *et al* (2010) also found that businesses improve with age. However, Loderer and Waelchli (2010) found that young firms perform better in profitability than older firms. Their finding suggested that, when profit starts to fall over time, firms eventually fail to sustain better profit (Loderer *et al*, 2010).

Start up Capital: Start up capital for entry may limit business operation. Robb and Fairlie (2008) found that the strength of the relationship between start-up capital and business success is strong for each type of business outcome. Robb and Fairlie (2008) argue that business start-up problems are usually due to limited equity to finance a new business. Banks are often reluctant to lend money to small businesses because of low expected profit margins and expected high risk of business failure (Robb and Fairlie 2008).

Industry Sector: Business industry sector is crucial in determining the success or failure of busineses. For instance, investing in a specific industry may be influenced by the consumer's objective to purchase a specific product or service produce by that specific industry (Spulber, 2009). As stated by Papadaki and Chami (2007), start up barriers may be lower in industries that are engaged in retail and personal services. These industries are usually characterized by more intense competitive pressure. Products or services in these sectors may be easily imitated. In contrast, participation in industry sector offering

more professional services may be highly dependent on a very specific sets of capabilities or requirements, developed through prior experience or education which renders imitation difficult. Some studies have found that there are two apparent problems when examining business industrial sector; first, blurring of business activities, and second, variation in industry definition. The argument is that business activities are blurred because larger firms are inadequately represented by a single industry classification. That is to say that one company may be classified as a manufacturing firm but that it has many divisions (subsidiaries) that deliver services to customers that are in different industrial classifications. Therefore, grouping different activities within a single business sector identity code may blur sector differences among firms (Davidsson *et al*, 2002).

Ownership: Business ownership also has an impact on business performance. There is a contrast in the nature of business ownership when it comes to the question of national owned and foreign owned companies. For instance, foreign owned companies may tend to prosper better than national owned companies due to experience and capital variations or limitations (Bewley *et al*, 2010). The type of ownership structures used in this study, as defined by the dataset, are: general partnership; limited partnership; public company; cooperative sole proprietorship; and private company.

Gender: In their study of the growth determinants of micro businesses in Canada, Papadaki and Chami (2007) found that women are just as successful as men in business ownership. On the other hand, significant differences were found between men versus women in avoidance of uncertainty. Gupta *et al* (2009) found the existence of many differences between male and female entrepreneurs due to gender stereotypical characteristics attributed to both genders in society, that influence the classification of various occupations as either masculine or feminine. These differences, according to Gupta *et al* (2009) indicate that men are at an advantage when compared to women because of gender stereotypical perception's of entrepreneurship. Men seem to benefit from many factors, chief among them being a better structured business plan, easier access to financial resources, and better networking abilities. Kepler and Shane (2007) stated the existence of differences in motivation between male and female entrepreneurs. Male entrepreneurs were more likely than female entrepreneurs to

start a business and make money, and believed that starting a business is more important than spending time with one's family.

Education: Another most crucial aspect of entrepreneur's characteristics is the entrepreneur level of education. Papadaki and Chami (2007) reported that education is presumably related to knowledge, skills, motivation, self confidence, problem solving ability, commitment, and discipline. The argument is that higher education is likely to increase an entrepreneur's ability to cope with problems and to seize opportunities that are important to the firm's success and growth. Papadaki and Chami's (2007) report indicates that businesses of owners with higher education exhibit higher growth than those of owners with lower educational qualifications.

Location: A business geographical location may affect its performance; this usually depends on it's market sector and on demand in a particular location. For instance, Stoke and Wilson (2010) reported that a firm's growth depends on the economic health of the local population. Indarti (2004) reported differences of specific location factors from one business sector to other. Bewley *et al* (2010) on the other hand, reported that the probability of success is high for businesses that are located in areas with a diverse industrial structure and economic activity. However, differences in infrastructure are also relevant; the argument is that a good transport network can enable effective competition, even when economic activity is geographically dispersed (Bewley *et al*, 2010). Both Storey (1994) and Davidsson *et al* (2002) have observed that some locations are more likely to foster business growth than others. In both studies (UK firms and Swedish firms), the empirical results suggest that businesses in smaller (rural) areas are slow to grow compared to businesses in big (urban) centres (Storey, 1994). Mead and Liedholm (1998) also found that location played a significant role in determining micro-and small enterprises' chances of success and survival. Generally it can be agreed that decisions on business location are crucial to investors.

Partnership: Empirical studies have found that partnership is an important factor of business success. For example, Bewley *et al* (2010) reported that partnership plays the role of out-sourcing that provides intermediate goods and services to an enterprise. The argument is that partnership forms a platform for collaboration between businesses. For instance, a

partnership can be formed with international partners in order to facilitate entry into a new market. This may foster trade in importing inputs with low cost and in exporting products that are produced internally to the international market.

Cash flow problem: Financial access is very important for business operation. All businesses need to spend some money before they can earn money. A supermarket, for instance, needs to purchase goods before it can sell goods. In this process, the business needs liquid capital for operation. If a business runs out of available liquid capital and is not able to purchase needed inputs, it has cash flow problems (liquidity problems). Saridakis *el al* (2007) highlight that self-reported liquidity constraints by new businesses in the first year of trading, have enduring impacts upon business success and survival. These liquidity constraints are typically those of younger businesses that were discouraged from accessing finances from the bank.

Financial shock: Businesses may suffer from financial loss due to shock. Shock can be from: fire; theft; flooding; car accidents; personnel injuries during working hours; and eviction. These shocks may cause financial losses to businesses. Financing these shocks may force businesses to borrow, or sell business assets to recover. Generally, shocks affect a business's operation.

Source of Financing: Access to capital is very crucial for business entry and sustainability. Generally, micro and small enterprises rely heavily on banks for the provision of finance. That is to say most micro and small firms use external sources to finance their operations, usually through financial institutions (banks), venture capitalists, and individual investors (Papadaki and Chami, 2007). Ricketts (2002) stipulated that due to the problems of information asymmetry (lack of dual information exchange) which confront entrepreneurs and financiers depends to some extent on the nature of the banking system. Ricketts (2002) argued that banks may avoid the risk of businesses failure to refund credit. Saridakis, *et al* (2007) also found that financial constraints at business start-up not only significantly impacts business growth, but also negatively affects survival. Stokes and Wilson (2010) stipulated that some enterprises may have limited ability to raise external equity finance from shareholders. Some sole traders may not have shareholders and thus, the equity route is

closed . Generally, it can be argued that capital accessibility fosters business success.

Technology and Information Access: Technology can either be purely hi-tech (processes carried out by machines / software) or low-tech (involving changes in patterns of human behaviour) (Stokes and Wilson, 2010). Technologies can be shared among industries or businesses; transfer of technology, for instance, can help to facilitate the production of new products, materials, and services. Dibrell *et al* (2008) found that the impact of innovation on business performance was primarily indirect and was fuelled by information technology. Dibrell *et al* (2008) argued that in order to maximize investment in innovation activities, information technology initiatives should be aligned with innovation. It was further highlighted that micro-and small enterprises should consider how to apply information technology to other strategic initiatives, such as consumer responsiveness, in order to compete with larger businesses. As stipulated by Chrysostome and Rosson (2004), technology and information access has changed many business practices and has provided companies, particularly small and medium enterprises with the benefit of low communication costs, exposure to foreign markets, and the use of e-commerce. However, these advantages generally hide many problems that are related to engaging in doing international business, for instance, through the internet. That is to say, except for standardized products, the internet has not been able to replace the person-to-person contact that is generally required to build trust for successful international transactions. Generally, technology improves business productivity, and access to information helps businesses to innovate and to compete efficiently in local and global markets.

Government Support: To ensure best practice, governments have to establish proper legal systems and create a better environment for business. Promotion of the private business sector is crucial for economic growth. Governments in most cases give contracts to businesses and this strengthens busineses' capital for entry and sustainability. Among other aspects government intervention in situations of market failure is crucial (Ricketts, 2002). On the other hand, lack of openness (corruption) within government and complicated bureaucracy can hinder business success.

Chapter 3

The Data and Descriptives

3.1 The data

The data used in this study was obtained from the South Sudan business survey conducted in early October 2010. The Business Survey 2010 covered formal businesses in the ten state capitals of South Sudan. The objective of the survey was to obtain information on how the business sector is operating in South Sudan. The data from the survey provided information on the economy that is crucial in understanding the functioning of the free market; it also provided vital information for foreign investors who are looking for investment opportunities in South Sudan. It is likely that this information will enhance economic growth and lead to an increase in foreign investment. This information is also vital in order to advise the government, the private business community, and development partners on the formulation of private sector policy. The data was also intended to be used as an input in estimating private investment for South Sudan's GDP.

The methodology used in the business survey follows the International Standards for Industrial Classifications (ISIC) of economic activities. Due to the fact that South Sudan currently has few businesses operating, a category of one digit ISIC (Class A-S) was used to classify the business industry. From the pilot survey it was found that the majority of businesses in South Sudan are concentrated in just two classifications (G and I), thus a combination of international and location-specific classifications was also used. Formal busi-

nesses are defined as all businesses operating from a fixed structure and that have a business name. This definition includes small shops and stalls in markets that are operating under fixed structures and that have a business name. Traders selling from moveable structures (market tables) are not included as formal businesses (NBS, 2010).

The process of designing the business survey questionnaire took several months. The discussion started in mid-January 2010, where a Business Survey Technical Working Group (BSTWG), which comprised of the National Bureau of Statistics (NBS) staff and stakeholders, formed the main forum for discussion of the questionnaire. The questionnaire development took into account the main questions required by the members of the BSTWG, and also reflected the priorities and concerns of primary users. The collaborative process in the design of the questionnaire was identified as one of the main successes of the business survey (NBS, 2010). A first-draft of the questionnaire was tested in a small study early in the process of the design, where a few business enterprises in Juba were interviewed. The results from the pre-test were used as guidelines in the fieldwork. Later in the process, a pilot study was carried out in August 2010, during which a nearly-final-draft of the questionnaire was tested. The feedback from the pilot was incorporated in the final design of the questionnaire. Generally, the questionnaire consists of 11 modules (leveled A to K), each covering a specific topic: Module A is an identification module which covers business names, activities and geographical codes; Module B concerns ownership structures; Module C covers labour statistics, including employment and remuneration of employees; Module D covers the income and sales, which also includes production, trade margins, and export; Module E concerns intermediate consumption and business imports; Module F covers investments and sales of fixed assets which is also intended for calculation of gross fixed capital formation; Module G concerns the value of businesses' stocks. Module H covers business environment, business development in the three years prior to the survey, the expectations for the next three years and what were considered to be the major constraints for the business; Module I concerns shocks, and financial losses due to shocks; Module J concerns finance availability. The intention of this module was to obtain information on the lending of money; Module K is the last module and it focuses on taxes and registration of the business. Module H is to a

large extent shaped by the stakeholders, where some of the ministries gave many inputs on what kind of information they would find useful.

The sample used in the business survey was drawn from the business listing done in May 2010. From the result of the business listing, 7 333 businesses were listed in the ten state capitals of South Sudan. 2 000 businesses were then selected as a sample from the sampling frame comprising of 7 333 businesses. The stratification sampling technique was used to design the sample. Stratification is a process of grouping a population (businesses) from population (busineses) listing, into a relatively homogenous subgroup before sampling. Thus when subpopulations (stratum) vary considerably, it is advantageous to sample each subpopulation (stratum). These subpopulations (strata) are non-overlapping (Lehtonen *el al*, 1995). For this survey, within each state the frame was stratified by economic activity (ISIC Rev. 4) and number of employees. The strata were formed as follows: Stratum 1 consists of 6 or more employees; Stratum 2, $3-5$ employees, ISIC group G (Wholesale and retail trade); Stratum 3, $3-5$ employees, ISIC group I (Accommodation and food service activities); Stratum 4, $3-5$ employees, all other ISIC groups. Stratum 5, $1-2$ employees, ISIC group G (Wholesale and retail trade); Stratum 6, $1-2$ employees, ISIC groups I (Accommodation and food service activities); Stratum 7, $1-2$ employees, all other ISIC groups (NBS, 2010). The final sample is presented in Table 3.1

Table 3.1: Final sampling design

	Stratum							
State	***1***	***2***	***3***	***4***	***5***	***6***	***7***	**Total**
Upper Nile	**32**	39	41	28	20	20	20	200
Jonglei	**22**	**39**	**23**	**16**	30	**9**	**11**	150
Unity	**13**	**30**	**7**	**11**	45	13	31	150
Warrap	**21**	**19**	**40**	**14**	22	17	17	150
Northern Bahr el Ghazal	**17**	**23**	**33**	**15**	32	15	15	150
Western Bahr el Ghazal	**29**	**54**	**23**	**13**	41	20	20	200
Lakes	**25**	42	24	14	15	15	15	150
Western Equatoria	**36**	35	19	**15**	15	15	15	150
Central Equatoria	**314**	72	41	63	20	20	20	550
Eastern Equatoria	**25**	**31**	**14**	**21**	20	19	20	150
Total Sample								**2000**

Before the data collection (fieldwork) started, a training of field officers (enumerators) took place in September 2010, which lasted for nine days. The trainees were divided into two

training locations; one group in Rumbek, and the other in Juba. The fieldwork was launched in early October 2010. A total of 41 enumerators and 10 Field Operation Managers were involved in the process of data collection. The fieldwork generally ran smoothly; a total of 2 000 businesses were interviewed (NBS, 2010).

After the fieldwork, the questionnaires were scanned for a period of three weeks at the NBS head office in Juba. The scanning syntax was written remotely in Oslo by a consultant from Statistics Norway (NBS, 2010). Following the scanning of the questionnaires from all the ten states, the data was compiled in SPSS format for data cleaning (editing). The draft SPSS syntax for the data cleaning was prepared in advance and was ready by the time the data scanning was completed. This was tested and adjusted where necessary with the actual data, and included a number of manual visual checks on possible outliers. This process was done module by module, and at each stage a check was done to ensure that the information was in line with the other modules. The whole process was completed in Juba by NBS staff by February 2011.

3.2 The Variables of Interest

The response variable used in this study was business profitability. This was based on the profits gained by businesses over the period of one year. The profit gained was calculated as the difference between the business revenue (income) and the business cost over the period of one year. Businesses that gained a profit are considered successful, whereas those that sustained losses are considered to have failed.

The business survey dataset has captured a number of characteristics that are assumed to have influence on business success or failure. However, due to some missing values, a number of these explanatory variables were excluded from our study. Generally, after exploratory analysis, we identified fourteen independent variables that are suspectable to business success, as discussed in Chapter 2. The fourteen variables are briefly discussed as follows: 1 - State (location); explains more about the states/regions in which an investor may wish to invest. 2 - Ownership; describes the type of ownership structure that the business is engaged in. 3 - Technology; assesses the use of internet by business. 4 - Government

Support; highlights the availability and use of government support to sustain business, and also the creation of a conducive environment for business. 5 - Cash flow (liquidity) problems; explains financial limitations businesses experience. 6 - Financial loss due to shock; describes the unexpected and accidental financial loss experience by business this was included as a measure of financial shock . 7 - Business size; refers to the number of employees businesses have, it is used to measure micro-and small enterprises. 8 - Business age; refers to the number of years of business operation. 9 - Outstanding loan; refer to the loans that are due to lenders which are not yet settled by the business. 10 - Start up capital; represents the amount of starting capital of the businesses. 11 - Gender; refers to gender of business owner. 12 - Education level; investigates the highest level of education of the employees/owners. 13 - Stakeholders; describes the type of partnership the business is involved in. 14 - Industry; refers to the type of business sector. The coding and the levels of the data variables summary is given in Table 3.2.

3.3 Preliminary Data Analysis

The exploratory analysis presented in this section constitutes a descriptive statistics analysis. The aim is to present a preliminary analysis of some variables in the dataset. SPSS version 19 was used to construct frequency tables, cross tabulation, and graphs.

From Table 3.2, the businesses that are successful constitute about 71.9% whereas 28.1% are unsuccessful. Juba, Malakal, and Wau have the highest proportion of respondends which constitutes about 27.1%, 10.1%, and 10% respectively. 66.8% of the business ownerships are sole proprietorships, 28.1% are partnerships, and 14% are companies. 85.4% of the businesses have not been using internet, while 14.6% have been using internet. 95.3% of the businesses reported that government was supporting them, whereas, 4.7% reported no support from the government. 61.6% of the businesses reported cash flow problems, as compared to 38.4% of the businesses that reported no cash flow problems. Businesses that experienced financial loss due to shock are 34.9%, compared to 65.1% of businesses that have not experienced financial loss due to shock. Micro-enterprises constitute about 88.4% of the businesses, whereas small-enterprises constitute about 11.4%. Most of the businesses have a median age of 6 years.

Table 3.2: Variable Levels and Frequency Distribution

Characteristic	Levels	N	%	Q1	Median	Q3
Business Success	Successful	1421	71.9			
	Not successful	555	28.1			
Business Location	Malakal	200	10.1			
	Bor	148	7.5			
	Bentiu	148	7.5			
	Kuajok	150	7.6			
	Aweil	147	7.4			
	Wau	197	10.0			
	Rumbek	150	7.6			
	Yambio	149	7.5			
	Torit	150	7.6			
	Juba	535	27.1			
Ownership	Partnership	297	28.1			
	Company	276	14.0			
	Others	83	4.2			
	Sole Proprietorship	1320	66.8			
Use of Internet	Yes	288	14.6			
	No	1688	85.4			
Government Support	Yes	1883	95.3			
	No	93	4.7			
Cash Flow Problem	Yes	759	38.4			
	No	1217	61.6			
Financial Shock	Yes	689	34.9			
	No	1287	65.1			
Business Size	Microenterprise	1746	88.4			
	Smallenterprise	230	11.6			
Business Age	Scale			3.9	6.3	8.7
Outstanding Loan	Yes	468	23.7			
	No	1508	76.3			
Startup Capital	Scale			3,890.9	7,780.2	11,669.6
Gender of Owner	Male	1215	61.5			
	Female	761	38.5			
Education Level	No Schooling	524	26.5			
	Primary school	338	17.1			
	Secondary school	722	36.5			
	University degree	281	14.2			
	Vocational training	111	5.6			
Stakeholders	Local	1897	96.0			
	Foreigner	79	4.0			
Industry Type	Mining, Energy, Manufacturing and Construction	122	6.2			
	Trade and Transport service	1478	74.8			
	Administration, Professional and Scientic service	98	5.0			
	Education, Social, and Health Service	192	9.7			
	Other Service	86	4.4			

N=Frequency, %=Percentage Q1=First Quartile, Q3=Third Quartile. The Median, Q1 and Q3 are used for the scale variables.

Businesses with outstanding loans are 23.7%, compared to 76.3% of businesses with no outstanding loans. Most of the businesses have a maximum startup capital of 7 780 South Sudanese pounds. 61.5% of the entrepreneurs are males, whereas 38.5% are females. 26.5% of business entrepreneurs have no qualification, 17.1% have completed primary school, 36.5% have completed secondary school, 14.2% have a university degree, and 5.6% have undergone vocational training. 96% of the businesses stakeholders are local South Sudanese, whereas 4% are foreigners. Most of the businesses are engaged in trade and transport services which consist of wholesale and retail trade, transportation and storage, accommodation and food service, and real estate activities; this constitutes about 74.8%% of the total industry sector.

Table 3.3: Distribution of Business Success and Failure by Location

Business Location	Succeed (%)	Fail (%)
Malakal	69	31
Bor	87.8	12.2
Bentiu	51.4	48.6
Kuajok	49	51
Aweil	40.1	59.9
Wau	46.2	53.8
Rumbek	61.3	38.7
Yambio	89.9	10.1
Torit	88	12
Juba	92.4	7.6

Table 3.3 presents a distribution of business success and failure by location. It can be observed that Juba, Yambio, Torit, and Bor have the highest percentages of successful businesses, which constitute about 92.4%, 89.9%, 88%, and 87.8% respectively. Aweil, Wau, and Kuajok on the other hand, have the highest percentages of unsuccessful businesses which constitute about, 59.9, 53.8%, and 51% respectively.

From Figure 3.1, male entrepreneurs who are engaged in the sole proprietorship type of business ownership are 67.2%, compared to 32.8% of female entrepreneurs. Female entrepreneurs who have companies are 64.9%, compared to 35.1% of male entrepreneurs. 62.6% of male entrepreneurs are engaged in the partnership type of business ownership, compared to 37.4% of female entrepreneurs.

From Table 3.4, most of the business ownerships are dominated by sole proprietorships across the locations. Bentiu has about 83.1% of sole proprietorship business ownerships

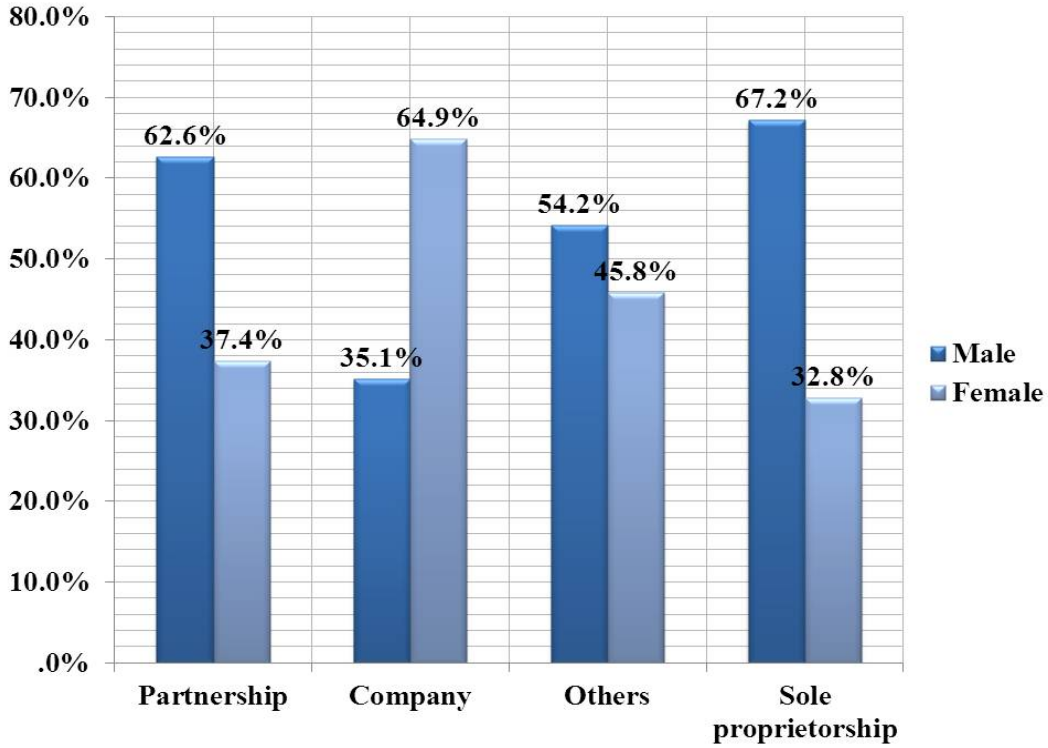

Figure 3.1: Distribution of Business Ownership by Gender

Table 3.4: Distribution of Business Ownership by Location

Business Location	Partnership (%)	Company (%)	Sole Proprietorship(%)	Others(%)
Malakal	10	3.5	76.5	10
Bor	25	2	70	2
Bentiu	10.8	3.4	83.1	2.7
Kuajok	19.2	2	74.2	4.6
Aweil	22.4	2	75.5	0.0
Wau	22.3	2.5	6.1	69
Rumbek	17.3	6	75.3	1.3
Yambio	5.4	22.1	62.4	10.1
Torit	2.7	27.3	65.3	4.7
Juba	14.9	31.2	51.5	2.4

which is the highest compared to the other locations, followed by Malakal (76.5%) and Rumbek (75.3%). Bor has the highest percentage of partnerships (25%), followed by Aweil (22.4%) and Wau (22.3%). Juba has the highest percentage of companies (31.2%), followed by Torit (27.3%) and Yambio (22.1%).

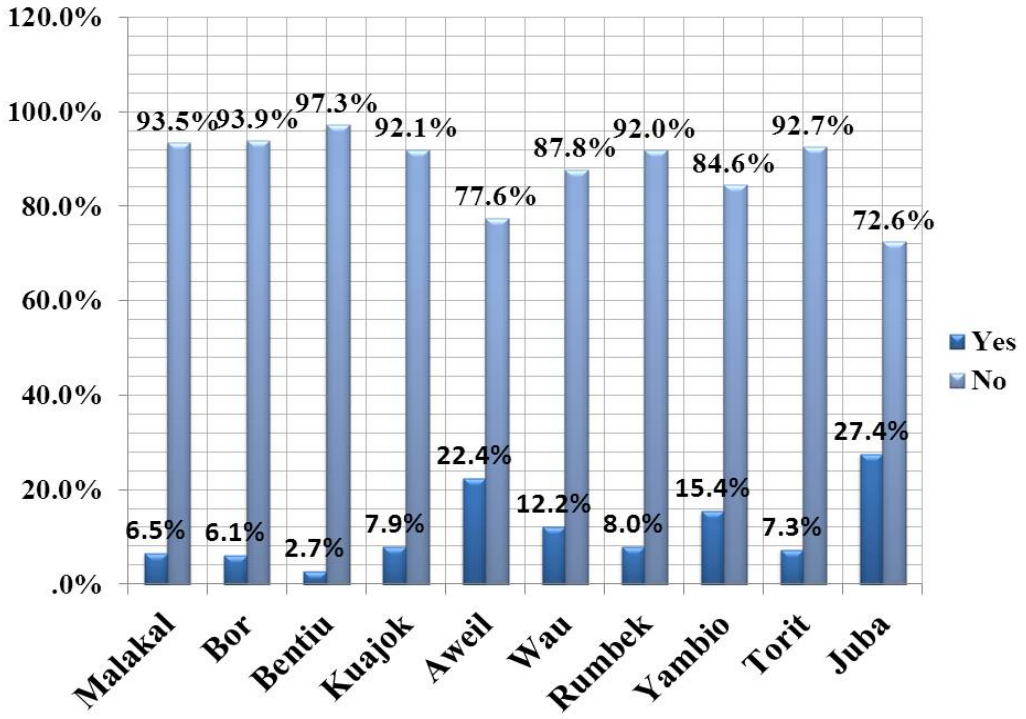

Figure 3.2: Distribution of Internet usage by Location

Figure 3.2 displays the distribution of internet usage by location. It can be seen that Juba has the highest percentage of businesses that use the internet, followed by Aweil, Yambio, and Wau, which constitute about 27.4%, 22.4%, 15.4%, and 12.2% respectively. The use of the internet can be considered as a technological input for business success as it is needed for business innovation.

Figure 3.3 displays the distribution of cash flow problems by location. Yambio has the highest percentage of businesses with cash flow problems (57%), followed by Malakal (46.5%) and Kuajok (42.4%). Cash flow problems can cause business to fail. For instance, if a business runs out of available liquid capital, it may not be able to purchase all the needed intermediate goods, and hence experiences a cash flow problem. As discussed in the lit-

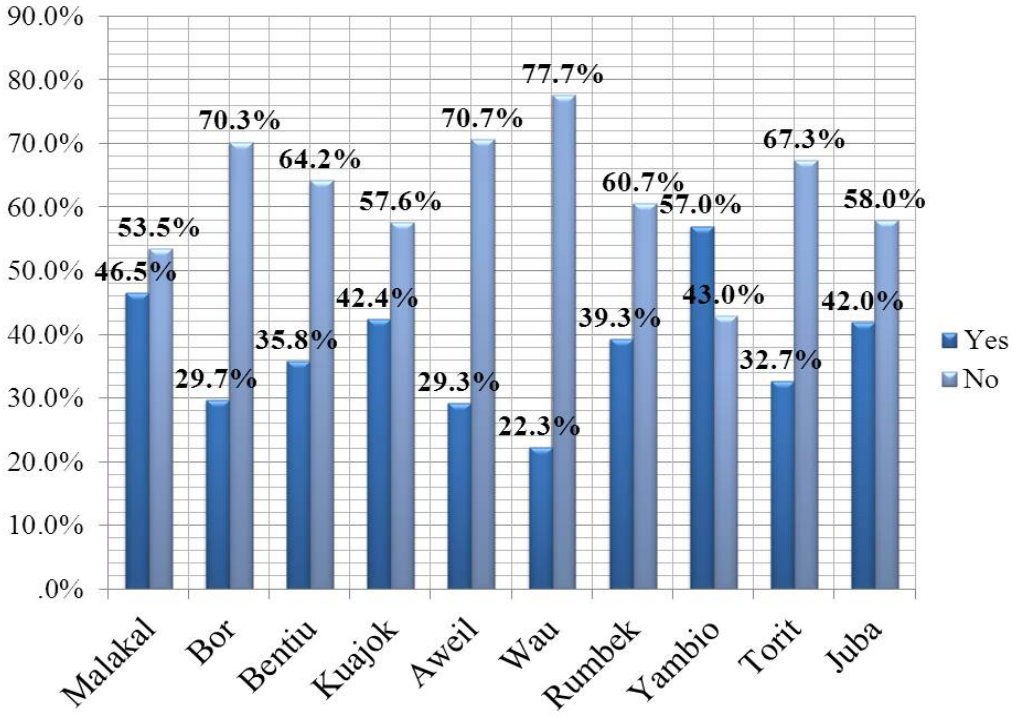

Figure 3.3: Distribution of Cash Flow Problems by Location

erature, most micro and small businesses experience cash flow problems due to lack of/or limited access to external finances, for instance from banks. This is because of a lack of proper collateral experience by micro and small businesses.

It can be observed from Figure 3.4, that female entrepreneurs have more cash flow problem than their male counterparts. That is to say that about 42.7% of female entrepreneurs reported cash flow problems, compared to 35.7% of male entrepreneurs.

Figure 3.5 presents the distribution of business financial loss due to shock, by location. Bor reported the highest percentage of businesses that have experienced financial loss due to shock (59.5%), followed by Rumbek (51.3%) and Kuajok (43.7%). The financial shock mentioned here may be fire, theft, flooding, car accidents, personnel injuries during working hours, and eviction.

From Table 3.5, Bentiu reported the highest percentage of businesses with outstanding loans (50%), followed by Malakal (40.5%) and Wau (29.4%). Like cash flow problems, outstanding loans can cause businesses to fail. For instance if a business is in debt and is

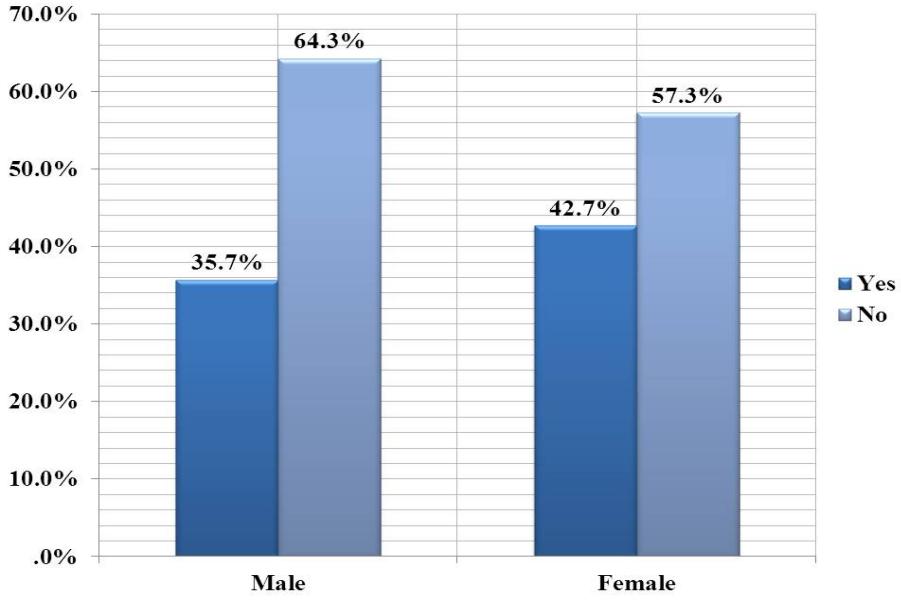

Figure 3.4: Distribution of Cash Flow Problem by Gender

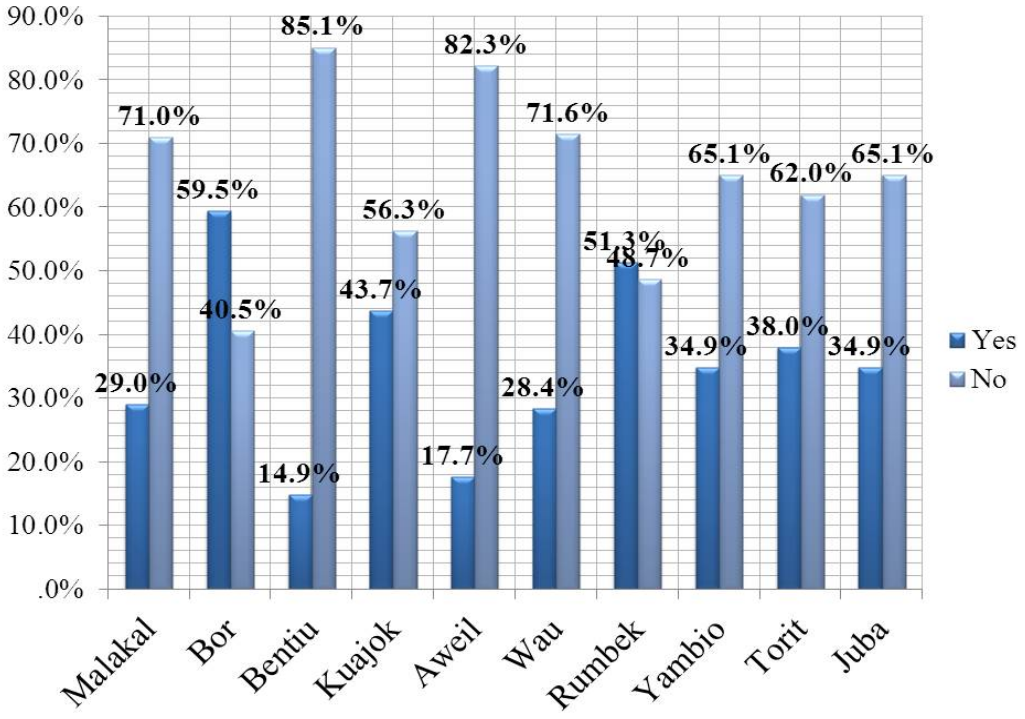

Figure 3.5: Distribution of Financial Loss due to Shock by Location

Table 3.5: Distribution of Business Outstanding loans by Location

Business Location	Yes (%)	No (%)
Malakal	40.5	59.5
Bor	26.4	73.6
Bentiu	50	50
Kuajok	17.9	82.1
Aweil	4.8	95.2
Wau	29.4	70.6
Rumbek	22	78
Yambio	12.1	87.9
Torit	13.3	86.7
Juba	20.7	79.3

not able to settle its outstanding loan, this may block the business's path of borrowing.

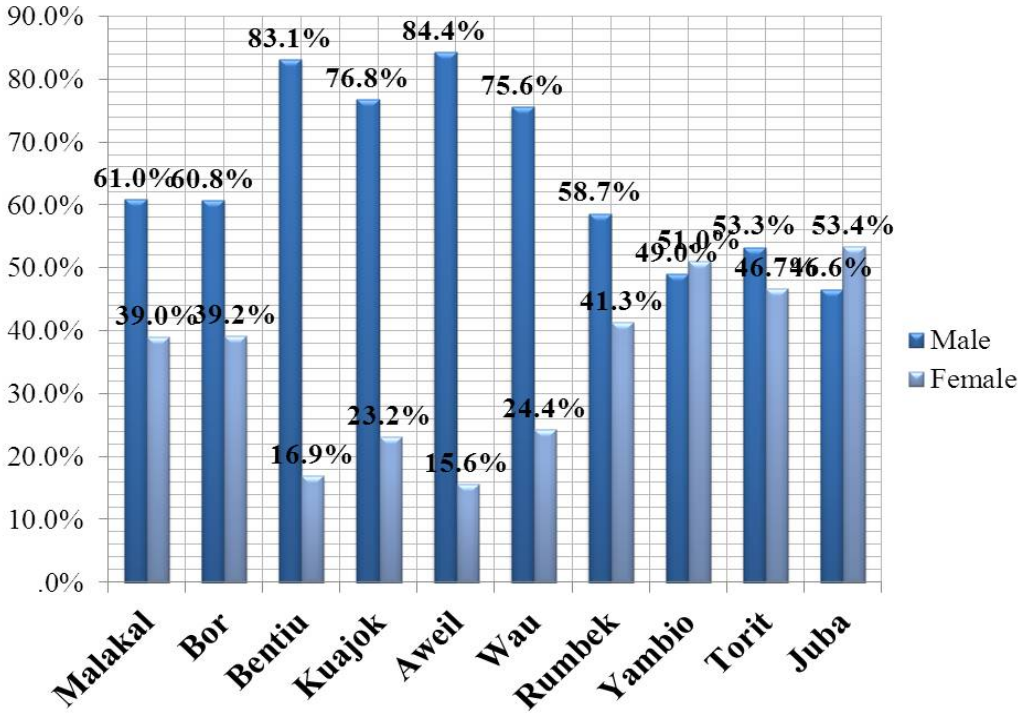

Figure 3.6: Distribution of Entrepreneur Gender by Location

On the discussion of entrepreneur gender, Figure 3.6 indicates that Aweil has the highest percentage of male entrepreneurs (84.2%), compared to the other locations, followed by Bentiu (83.1%), Kuajok (76.8%), and Wau (75.6%) respectively. Juba on the other hand has the highest percentage of female entrepreneurs compared to the other locations, which constitute about 53.4%, followed by Yambio (51%) and Torit (46.7%).

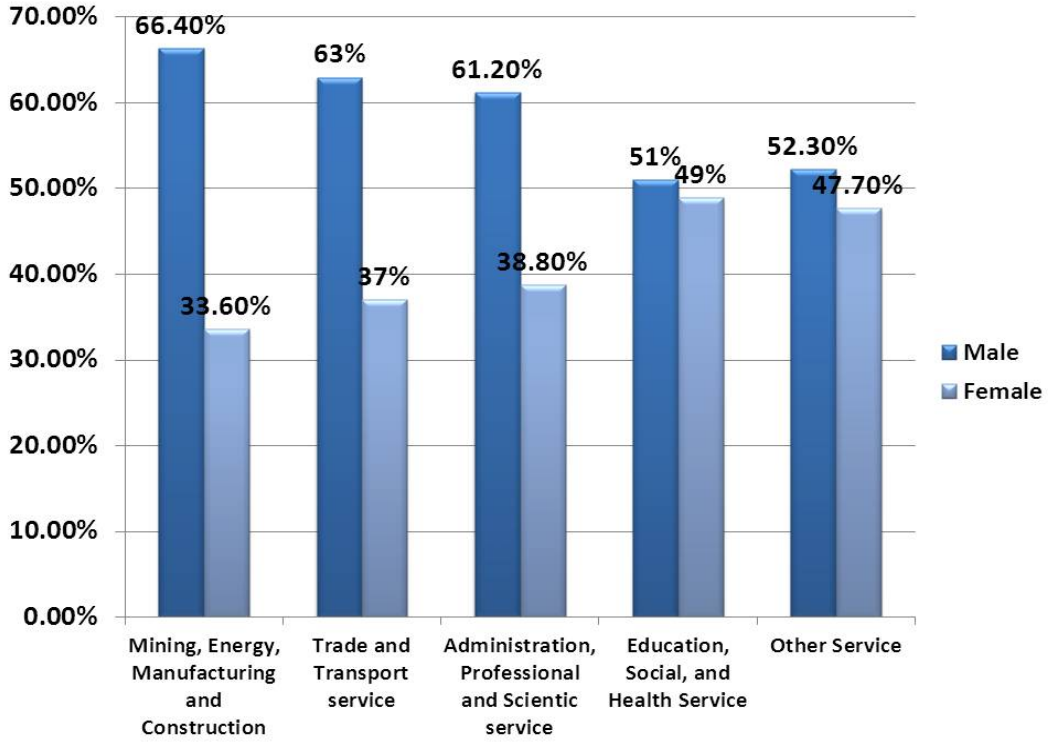

Figure 3.7: Distribution of Entrepreneur Gender by Industry Type

From Figure 3.7, male entrepreneurs who are engaged in mining, energy, manufacturing, and construction activities consitute 66.4%, compared to 33.6% for female entrepreneurs. 63% of male entrepreneurs are engaged in trade and transportation activities, compared to 37% of female entrepreneurs. 61.2% of male entrepreneurs are engaged in administrative, professional and scientific services compared to 38.8% of the female entrepreneurs. Whereas, male entreprenuers who are investing in education, social, and health services are 51% compared to 49% of the female entreprenuers. Likewise Male entreprenuers who are investing in other services are 52.3% compared to 47.7% of the female entreprenuers.

From Table 3.6, business entrepreneurs with no qualification constitute the highest percentages of 52.3% for Kuajok and 40.7% for Rumbek, as compared to the other locations. The highest percentages of business entrepreneurs who have completed primary school are reported in Malakal (26%), Wau (24.2%), and Bentiu (23%). Percentages of business entrepreneurs who have completed secondary school are the highest for Yambio (55%), for Torit (44.7%), and for Juba (41%). The highest percentages for business entrepreneurs who have

Table 3.6: Distribution of entrepreneurs Education level by Location

Business Location	No Schooling (%)	Primary School (%)	Secondary School (%)	University Degree (%)	Vocational training (%)
Malakal	19.5	26	32	20	2.5
Bor	33.1	13.5	39.5	5.4	8.8
Bentiu	34.5	23	28.4	8.1	6.1
Kuajok	52.3	15.2	27.2	5.3	0
Aweil	26.5	22.4	40.1	8.8	2
Wau	36.5	24.4	25.9	7.3	12
Rumbek	40.7	14.7	25.3	7.3	12
Yambio	12.8	16.1	55	14.8	1.3
Torit	18	16	44.7	6	15.3
Juba	16.4	10.8	41	24.6	7.1

university degrees are reported for Juba (24.6%), followed by Malakal (20%) and Yambio (14.8%) . The highest percentages are reported for Torit (15.3%), Bor (8.8%), and Juba (7.1%) for business entrepreneurs who have completed vocational training.

Chapter 4

Modelling Binary Response Variable

From previous Sections it was emphasized that the response variable in our data set is a binary variable. A special case of the generalized linear model, called the logistic regression model, is then the best option to deal with such a data set. Thus, we first highlight theories on the generalized linear models, and then we will illustrate the logistic regression model as a special case of the generalized linear model.

4.1 Generalized Linear Models

The generalized linear models (GLMs) have features that are generally applicable to many regression problems. GLMs extended the ordinary regression model to encompass non-normal response distribution and modelling functions of the mean (Agresti, 2002). The linear models are intended to model a continuous response variable y, as a function of one or more factor variables (Dunteman, 2006). This types of regression follows the normal distribution assumptions. However, there are situations where the response variable is not a continuous variable, but rather is categorical (dichotomous or binary) in nature hence, does not satisfy the normality assumption. For instance, in a situation where business success or failure is measured as a response variable, which can be coded as 0 or 1 (0 = failure, and 1= success). This type of response variable is not continuous but binary. There are also situations where the response variable is a count, such as a count of defects of company products. These types of regression problems fall under the natural exponential family of distributions (Agresti, 2002). The generalized linear model was developed by Nelder and

Wedderburn in 1972. It can be used to fit binary response data that follows a very general distribution of the exponential family. The exponential family includes the normal, binomial, Poison, geometric, negative binomial, exponential, gamma, and inverse normal distribution (Mayers *et al*, 2002). Generally, the most unifying concept underlying the GLM is the exponential distribution. The reason for restricting the GLM to the exponential family of distribution for the response variable y, is the algorithm applied to the entire family for any choice of the link function. The members of the exponential family of distributions all have the probability density function for an observed response y, of the form

$$f\left(y_i;\theta_i,\phi\right)=\exp\left[\frac{y_i\theta_i-b(\theta_i)}{a_i(\phi)}+\ c(y_i,\phi)\right],i=1,2,\ldots,n \tag{4.1}$$

where ϕ is called the scale (dispersion) parameter, and θ_i is a natural parameter, $b(\theta_i)$ is a normalizing function. The function $a_i(\phi)$ has the form $a_i\left(\phi\right)=\frac{\phi}{w_i}$ and is called the dispersion parameter for a known weight w_i. Generally, $a_i(\phi)$, $b(\theta_i)$, and $c(\phi)$ are referred to as specific functions (Agresti, 2002). Therefore the generalized linear model is given by

$$g(\mu_i)=\mathbf{x'}_i\beta=\beta_0+\beta_1x_{i1}+\beta_2x_{i2}+\ldots+\beta_px_{ip} \tag{4.2}$$

Generalized linear models have three components:

1. A random component that identifies the response variable y, and its probability distribution,

2. A systematic component that specifies the explanatory variables used as a linear predictor function and;

3. The link function g, that links the linear predictor to the natural mean of the response variable y.

The appropriate link function and the analysis techniques depends upon the level of the measurement of the response variable, for instance, if the response variable is normal (continuous), then the link between the response and the linear predictor of the model is direct

(identity). Thus, the ordinary regression model can be used. However, if the response variable is categorical, then other links can be used. Generally, the random components of the GLM consist of response variable Y, with independent observation $(y_1, y_2, \ldots, y_N)$ from a distribution in the natural exponential family.

The expected value $[E(Y_i)]$ of the mean response variable y_i, and the variance $[var(Y_i)]$, can be derived from the log likelihood function Ł $= \sum_i L_i$, where

$$L_i = \left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \ c(y_i, \phi)\right]$$

which follows that

$$E\left[\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right] = 0 \qquad 0r \qquad \mu_i = E(Y_i) = b'\ \theta_i$$

and

$$\frac{b''(\theta_i)}{a_i(\phi)} = E\left[\frac{Y_i - b'(\theta_i)}{a_i(\phi)}\right]^2 = \frac{Var(Y_i)^2}{[a(\phi)]}$$

So that

$$Var(Y_i) = b''(\theta_i)a(\phi)$$

The GLM links η_i *to* $\mu_i = E(Y_i)$ by a link function g. Thus, μ_i is related to the explanatory variables by a linear predictor given by

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}, \quad i = 1, 2, \ldots, N$$

This can be written in matrix notation as

$$\eta_i = g(\mu_i) = \mathbf{x}'_i\beta, \quad i = 1, 2, \ldots, N.$$

where $\mathbf{x}'_i = [1, x_{i1}, x_{i2}, \ldots, x_{ip}]$, and $\beta' = [\beta_0, \beta_1, \beta_2 \ldots, \beta_p]$. The link function g, for which $g(\mu_i) = \theta_i$ is the canonical link. The direct relationship

$$\theta_i = \sum_j \beta_j x_{ij}$$

occurs between the natural parameter and the linear predictor. Since $\mu_i = b'(\theta_i)$, then the natural parameter is a function of the mean, $\theta_i = \left(b^{'}\right)^{-1}(\mu_i)$, where $\left(b^{'}\right)^{-1}$ denotes the inverse function of $b^{'}$. Thus, the canonical link is the inverse of $b^{'}$.

Each member of the exponential family of distributions has a unique canonical link function. The canonical link is given by

$$g(\mu_i) = \theta_i = \eta_i$$

This transforms the mean to the natural parameter. The other links include, the probit link, which is given by

$$\eta_i = \Phi^{-1}[E(y_i)]$$

where Φ represents the cumulative standard distribution function and, the complementary log-log link, which is given by

$$\eta_i = \ln[\ln(1 - E(y_i)]$$

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

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

The 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

$$\frac{\partial L(\beta)}{\partial \beta_j} = \sum_i \frac{\partial L_i}{\partial \beta_j} = 0, \quad \text{for all } j = 0, 1, 2, \ldots p. \tag{4.3}$$

Using the chain rule we obtain

$$\frac{\partial L_i}{\partial \beta_j} = \frac{\partial L_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}.$$

Since

$$\frac{\partial L_i}{\partial \theta_i} = \frac{\left(y_i\theta_i - b\left(\theta_i\right)\right.}{a_i\left(\phi\right)}$$

$$\mu_i = b'(\theta_i)$$

and

$$Var\left(y_i\right) = b''(\theta_i)a(\phi)$$

it then becomes

$$\frac{\partial L_i}{\partial \theta_i} = \frac{(y_i - \mu_i)}{a(\phi)}$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''\left(\theta_i\right) = \frac{Var\left(y_i\right)}{a(\phi)}$$

also since

$$\eta_i = \sum_j \beta_j x_{ij}$$

then

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}$$

by substituting all the above, the results can be rewritten as

$$\frac{\partial L_i}{\partial \beta_j} = \frac{(y_i - \mu_i)}{a(\phi)} \frac{a(\phi)}{Var\left(Y_i\right)} \frac{\partial \mu_i}{\partial \eta_i} x_{ij} = \frac{(y_i - \mu_i)\, x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}$$

Therefore the likelihood equation is given by

$$\frac{\partial L_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \ldots p. \tag{4.4}$$

These equations are nonlinear in $\hat{\beta}$; usually these equations are solved iteratively. First an initial solution of the equations denoted by $\hat{\beta}^{(.)}$ is guessed and then updated until the iterative algorithm converges to a solution $\hat{\beta}$, which is the maximum likelihood estimate of β (Agresti, 2002).

The likelihood function for the GLM also determines the asymptotic covariance matrix of the maximum likelihood estimator $\hat{\beta}$. This matrix is the inverse of the information matrix $\mathcal{J}$, which has elements

$$-E\left[\frac{\partial^2 l(\beta)}{\partial \beta_h \partial_{\beta_j}}\right].$$

Recall that the likelihood equation is given by

$$\frac{\partial L_i}{\partial \beta_j} = \sum_{i=1}^{N} \frac{(y_i - \mu_i)\, x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 0, 1, 2, \ldots p$$

It then follows that

$$E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) = -E\left[\frac{(y_i - \mu_i)\, x_{ih}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i} \frac{(y_i - \mu_i)\, x_{ij}}{Var(Y_i)} \frac{\partial \mu_i}{\partial \eta_i}\right]$$

$$E\left(\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) = \frac{-x_{ih} x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Since $L(\beta) = \sum_i L_i$ then

$$E\left(-\frac{\partial^2 L_i}{\partial \beta_h \partial \beta_j}\right) = \sum_{i=1}^{N} \frac{x_{ih} x_{ij}}{Var(Y_i)} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2$$

Generally, the information matrix has the form

$$\mathcal{J} = \mathbf{X}'\mathbf{W}X$$

where $\mathbf{X}$ is a design matrix and $\mathbf{W}$ is the diagonal matrix with main diagonal elements

$$w_i = \frac{(\partial \mu_i / \partial \eta_i)^2}{Var(Y_i)} \tag{4.5}$$

And the asymptotic covariance matrix of $\hat{\beta}$ is then estimated by

$$\hat{cov}\left(\hat{\beta}\right) = \hat{\mathcal{J}}^{-1} = \left(\mathbf{X}'\hat{\mathbf{W}}\mathbf{X}\right)^{-\mathbf{1}} \tag{4.6}$$

where $\hat{\mathbf{W}}$ is estimated $\mathbf{W}$ evaluated at $\hat{\beta}$**.** Clearly $\mathbf{W}$ also depends on the link function. Thus, the asymptotic sampling distribution of $\hat{\beta}$ is then given by

$$\hat{\beta} \sim \mathbf{N}(\beta, \hat{\mathcal{J}}^{-1})$$

Generally, the weighted least square method, the Newton-Raphson method, and the Fisher Scoring method are the iteration methods used for determining the maximum likelihood function for the parameter β (Agresti, 2002).

The Newton Raphson method is an iterative method for solving nonlinear equations whose solution determines the point at which a function takes its maximum. It begins with

an initial guess for the solution and the process continues until the method generates a sequence of guesses which are then converged to the location of the maximum, when the function is suitable and the initial guess is good. In general, Newton Raphson determines the value of $\hat{\beta}$ at which a function $\mathbf{L}(\beta)$ is maximized. Let

$$\mathbf{u}^{'} = \left(\frac{\partial L(\beta)}{\partial \beta_0}, \frac{\partial L(\beta)}{\partial \beta_1} \, , \ldots , \frac{\partial L(\beta)}{\partial \beta_p} \right)$$

and Let $\mathbf{H}$ denote the matrix entries

$$h_{ab} \; = \frac{\partial^2 L(\beta)}{\partial \beta_a \partial \beta_b}$$

Which is referred to as Hessian matrix. Let $\mathbf{u}^{(t)}$ and $\mathbf{H^{(t)}}$ be $\mathbf{u}$ and $\mathbf{H}$ evaluated at $\beta^{(t)}$, the guess t for $\hat{\beta}$. Step t in the iteration process $(t = 1, 2, \ldots,)$ approximates $\mathbf{L}(\beta)$ near $\beta^{(t)}$ by the terms up to the second order in Taylor series expansion. Thus

$$L\left(\beta\right) \approx L\left(\beta^{(t)}\right) + \mathbf{u^{(t)}}^{'}(\beta - \beta^{(t)}) + \left(\frac{1}{2}\right)\left(\beta - \beta^{(t)}\right)^{'} \mathbf{H}^{(t)} \beta - \beta^{(t)}$$

Solving

$$\frac{\partial L(\beta)}{\partial \beta} \approx \mathbf{u}^{(t)} + \mathbf{H^{(t)}}\left(\beta - \beta^{(\mathbf{t})}\right) = \mathbf{0}$$

For β yields the next guess which is expressed as

$$\beta^{(t+1)} = \beta^{(t)} - \left(\mathbf{H}^{(t)}\right)^{-\mathbf{1}} \mathbf{u}^{(t)} \tag{4.7}$$

Iterations proceed until changes in $L(\beta^{(t)})$ between successive cycles are sufficiently small. Therefore the maximum likelihood estimator is the limit of β^t, as $t \to \infty$.

Fisher scoring is an alternative iterative method for solving likelihood equations. According to this argument, Fisher scoring uses the expected value of the matrix, called the expected information, whereas the Newton Raphson method uses the matrix itself, which is referred to as the observed information (Agresti, 2002). Let $\mathcal{J}^{(t)}$ denote the approximation at iteration t for the maximum likelihood estimate of the expected information matrix; that is to say, $\mathcal{J}^{(t)}$ has elements

$$-E(H)$$

Evaluated at $\beta^{(t)}$. The Fisher scoring is then given by

$$\beta^{(t+1)} = \beta^{(t)} + (J^{(t)})^{-1}\mathbf{u}^{(t)} \tag{4.8}$$

Model selection in the GLM faces almost the same issues as in the general linear regression model. Agresti (2002) reported that the selection process becomes harder as the number of explanatory variables increases, this is due to a rapid increase in possible effects and interactions. Generally, there are two main goals for model selection: first, the model should be complex enough to fit the data well; and second, it should be simple to interpret (Agresti, 2002). The selection of variables that enter the model is done through three procedures: forward selection; backward elimination; and stepwise selection. Forward selection adds variables sequentially into the model until addition of new variable does not improve the fit. Backward elimination on the other hand begins with a complex model and sequentially removes variables. At each stage, backward elimination selects the variable that has the least damaging effect on the model (i.e. has the largest p-value) when it is removed from the model. The process stops when any further deletion of variables leads to a significantly poorer fit. Stepwise selection is similar to forward selection, however when there are many variables under consideration in a model, the stepwise procedure is preferable to forward selection because it has an advantage of minimizing the chance of keeping redundant variables out of the model. The Akaike Information Criterion (AIC), is another criterion besides significant tests, used to select a good model in terms of eliminating quantities of interest. The AIC judges a model by how close its fitted values tend to be true values in terms of a certain expected value. In a given sample, Akaike showed that this criterion selects the model that minimizes

$$AIC = -2LogL + 2p$$

Where L is the maximum likelihood, and P is the number of parameters in the model. This penalized a model for having many variables. Usually, the model with the small AIC value is the best (Agresti, 2002).

Generally after fitting a model, a test of goodness of fit is needed to assess whether the statistical model used is the best for fitting the data. This test measures the discrepancy between observed values and expected values for the model used in the analysis. Dobson

(2002) stated that adequacy of a model fit can be assessed by comparing the model of interest with a more general model that has the maximum number of parameters that can be estimated. This model is referred to as a "saturated model". It uses log-likelihood ratio statistics, which is a generalized linear model with the same distribution and link function as the model of interest. The most common log-likelihood ratios are the deviance and Pearson's chi-square. These log-likelihood ratios measure the discrepancy of fit between the maximum log-likelihood of the model of interest, and the log-likelihood of the fitted saturated model. A saturated model explains all variation by systematic components of the model. Let $\hat{\theta}$ denote the estimate of θ for the saturated model corresponding to the estimated means $\hat{\mu}_i = y_i$, for all i. Then for a particular unsaturated model, the corresponding maximum likelihood estimates are denoted by $\hat{\theta}$ and $\hat{\mu}_i$. Thus, the deviance is given by

$$D = -2[l\,(\hat{\mu};y) - l(y;y)]$$

where $l\,(\hat{\mu};y)$ maximized the log-likelihood under the current model or model of interest, and $l(y;y)$ maximized the log-likelihood in the saturated model. It is the log-likelihood ratio statistics for testing the null hypothesis that the model holds against the alternative hypothesis, that a more general model holds.

Over-dispersion is a situation that sometimes occurs in data that are modelled with binomial or Poison distribution. The general idea is that in binomial and Poison distributions, the scale parameter ϕ is expected to be 1. If the value of the scale parameter ϕ is less than 1, there is under-dispersion, whereas if the value of ϕ is greater than 1, there is over-dispersion. Olsson (2002) argued that a common effect of over-dispersion is when the estimates of standard errors are under-estimated. A simple way to model over-dispersion is to introduce the scale parameter ϕ into the variance function. Generally, there are two solutions for over-dispersion:

- The data is remodelled by imposing $var(\mu) = \phi\mu(1-\mu)$ for binomial, or $var(\mu) = \phi\mu$ for Poison; this forces the model to 1, or;
- If $\hat{\phi}$ is different from 1, then the distribution of the data is neither binomial nor Poison thus another distribution can be used.

Generally, over-dispersion may occur due to a lack of homogeneity in the data. This lack of homogeneity may occur between groups of individual or within individuals observations (Olsson, 2002).

The residual examination shows us where our model (GLM) is poorly fitted to an overall goodness of fit (Agresti, 2002). Therefore the deviance and the Pearson residual examination are essential to asses the goodness of fit. Generally for the residual test of model fit the standardized residual is plotted against the fitted values, or the deviance/Pearson residual is plotted against the linear predictor. The decision from the plot is that if there is no systematic pattern in the residual plot against the fitted (predicted) values, then the fit is good.

When assessing model goodness of fit or model adequacy, it is also important to check the accuracy of the choice of the link function. When the link function is not appropriately chosen, the resultant estimates will be wrong and conclusions drawn from such estimates will be misleading. Let $g(\mu)$ be the link function, then

$$Z = g(\mu_i) + g'(\mu_i)\epsilon_i$$

is the working variate. If the link is correct, then the fitted Z against the $g(\mu_i)$ must be a straight line. Alternatively, as discussed by Vittignhoff *et al.* (2005), the link function can also be tested by refitting the model with the linear predictor, and the square of the linear predictor obtained from the original model as factor variables. The decision on the appropriateness of the link function is as follows: if the link function is appropriate, then the linear predictor will be statistically significant, and the square of the linear predictor will be statistically insignificant. This implies that the prediction given by the linear predictor is not improved by adding the square linear predictor term, which is basically used to evaluate the null hypothesis that the model is adequate (Vittignhoff *et al*, 2005).

After analysing the data under study, some observations in the data may have an influence on the parameter estimates $\hat{\beta}$ (Olsson, 2002). Thus, inclusion or exclusion of such observations into or out of the model may change the model parameter estimates substantially. Outliers are observations that don't follow the pattern of other observations. Outliers can usually be detected by different types of plots; among others are residual plots, whereas

influential observations are detected using Cook's distance. The approximation of Cook's distance C_i, is obtained by

$$C_i \approx \frac{r_p^2 h_{ii}}{p(1-h_{ii})} \tag{4.9}$$

where p is the number of model parameters, r_p is the Pearson residual, and h_{ii} is the leverage which are the elements of the hat matrix. A high C_i implies that there are influential observations in the data set. Usually this is caused by a large deviance residual and leverages. Generally, the assessment of influential observations is done by an index plot of C_i against the i^{th} observation. For the i^{th} observation to have influence on the parameter estimate, its Cook's distance should be greater than 1.

4.2 Logistic Regression Model

The logistic regression model is a special case of the generalized linear model used to model categorical response variables. Logistic regression is a very powerful statistical technique for modelling binary response such as; "alive or dead", "success or failure". Such binary responses are used as generic terms of the two categories usually coded as 0 or 1. Let the binary response variable be defined as

$$y = \begin{cases} 1 & \text{if the outcome is a success} \\ 0 & \text{if the outcome is a failure} \end{cases}$$

and let the explanatory variable be denoted by $\mathbf{x}$ $(x_1, x_2 \ldots x_n)$. We then let $\pi_i = p(y = 1)$ be the probability of success, and $\pi_i = 1 - p\,(y = 0)$ be the probability of failure. Thus, the logistic regression model is given by

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} \ = \mathbf{x'}_i\beta \ , \quad i = 1, 2, \ldots, n \tag{4.10}$$

Where $\mathbf{x}_i^{'}$ is a vector of the covariates and the dummy variable corresponding to the i^{th} observation, and β is the vector of the unknown parameters.

Generally, there are three links used to model a binary response. These are the logit link, the probit link, and the complementary log-log link. The logistic regression model, when the

canonical link function is used, is given by

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1-\pi_i} \ = \mathbf{x'}_i\beta \ , \quad i = 1, 2, \ldots, n$$

The term

$$\log \frac{\pi_i}{1-\pi_i}$$

is called the logit function. The ratio

$$\frac{\pi_i}{1-\pi_i}$$

is the odds of success. The logit is generally referred to as the natural parameter of the binomial distribution and the logit link is its canonical link.

The probit model is an alternative method for modelling a binary response. It follows the assumption that the tolerance term is normally distributed with mean μ , and variance σ^2, i.e. $N(\mu, \sigma^2)$ for unknown μ and σ (Olson, 2002). The probability π_i, is a cumulative distribution function (cdf) for a standard normal distribution. Generally, the probit model is given by

$$\Phi^{-1}(\pi_i) = \mathbf{x'}_i\beta \quad i = 1, 2, \ldots, n$$

The probit link function $\Phi^{-1}(.)$ maps the (0, 1) range probabilities onto $(-\infty, \infty)$ a range of linear predicators. The curve of the model has the shape of the normal cdf when Φ is the standard normal cdf.

The complementary log-log function is similar to the logistic and probit models used for modelling binary response. Dobson (2002) argued that if the values of π are near 0.5, then the complementary log-log function is the same as the logit and probit models. Generally, complementary log-log is applicable if the value of π is near 0 or 1 (when there are extreme values). The arugment is that for complementary log-log, the cdf needs not be symmetric about the mid-point $\pi = 0.5$, as is the case for the logit and the probit models. The complementary log-log model is given by

$$\log(\pi_i) = \log[-\log(1 - \pi)]$$

Since logistic regression is a special case of the generalized linear models, the model selection and goodness of fit test is the same as discussed in Section 4.1. Generally, the Pearson chi-square or the likelihood ratio do not have limiting chi-squared distributions, but they are still useful for comparing models and can be applied in an approximate manner to grouped observed and fitted values for a partition of the space of x values. The argument here is that the number of explanatory variables increases, when simultaneous grouping of values for each variable can produce a contingency table with a large number of cells. Thus, it is then useful to use the Hosmer-Lemeshow goodness of fit statistic, which suggests the partition of observed and fitted values according to the estimated (predicted) probabilities of success using the original ungrouped data. This approach forms g groups in the partition of observed frequencies of the response variable $y = 1$, which have approximately equal size preferably 10 groups (Hosmer and Lemeshow, 2000). Generally, the estimated (predicted) probabilities are ordered in an ascending order before groupings can be performed. According to Hosmer and Lemeshow (2000) there are two grouping strategies that are proposed: first, the table is collapsed based on a percentile of the estimated probabilities; and second the table is collapsed on fixed values of the estimated probability. Thus, the first method used of g group, i.e $g = 10$ forms the first group containing $n'_1 = n/10$ subjects having the smallest estimated probabilities, and the last group containing $n'_{10} = n/10$ subjects having the largest estimated probability. In the second method, the use of $g = 10$ groups results in cutpoints defined at the value $k/10$, where $k = 1, 2, \ldots, 9$; this group contains all subjects with estimated probabilities between the adjacent cutpoints. Therefore for the row $y = 1$, the estimates of the expected values are obtained by summing the estimated probabilities over all subjects in the group. Whereas for the row $y = 0$, the estimated expected values are obtained by summing over all subjects in a group, one minus the estimated probabilities (Hosmer and Lemeshow, 2000). Let X^2_{HL} denote the Hosmer and Lemeshow goodness of fit statistic test. Then the Hosmer-Lemeshow test statistic X^2_{HL}, has a chi-square distribution with $g - 2$ degress of freedom. This statistic X^2_{HL}, is compared with the critical value of the chi-square distribution, with $g - 2$ *df* ($\chi_{g-2,\alpha}$), for checking model goodness of fit. The interpretation is as follows, if the X^2_{HL} statistic is significant, it implies that there is model

lack of fit. Whereas if the X^2_{HL} statistics is insignificant, it implies that the model fitted the data well, hence there is goodness of fit (Hosmer and Lemeshow, 2000).

Usually after a general diagnostic test for model goodness of fit, it is also important that validation of model predicted probabilities (measurement of model predictive power), is done so as to assess whether the model has really fitted the data well. In this study we used the classification table and the *ROC* curve to do this assessment.

Generally, a classification table cross classifies a binary response (y), with a prediction of whether it equals 0 or 1 ($y = 0$ or 1). The prediction is $\hat{y} = 1$, when $\hat{\pi}_i > \pi_o$, and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_o$, for some cut-off π_o. Most classification tables use $\pi_0 = 0.5$ (as a threshold) and summarise the predictive power by Sensitivity $= P(\hat{y} = 1/y = 1)$, and specificity $= P(\hat{y} = 0/y = 0)$. Table 4.1 is a plot of sensitivity against 1−specificity. It is the plot of true positive rate against false positive rate.

Table 4.1: Specificity and Sensitivity Classification

Outcome of diagnosis Test	**Correct Classification**		
	Y=1 (+ve)	Y=0 (-ve)	Total
$\hat{Y}$=1 (+ve)	S	F	S+F
$\hat{Y}$=0 (-ve)	F_n	S_n	$F_n + S_n$
Total	S+ F_n	F+ S_n	N= S+F+ F_n + S_n

From Table 4.1, the following probabilities can be estimated:

$$Sensitivity = \frac{S}{S + F_n}$$

The estimated probability of correctly classifying an observation with an outcome of success, or is the number of true positive assessments over the number of all positive assessments.

$$Sensitivity = \frac{S_n}{F + S_n}$$

The estimated probability of correctly classifying an observation with an outcome of failure, or is the number of true negative assessments over the number of all negative assessments.

$$Sensitivity = \frac{F}{F + S_n}$$

The estimated probability of incorrectly classifying an observation with an outcome of failure, or is the number of false positive assessments over the number of all negative assessments.

$$Sensitivity = \frac{F_n}{S + F_n}$$

The estimated probability of incorrectly classifying an observation with an outcome of success, or is the number of false negative assessments over the number of all positive assessments.

A receiver operating characteristic (ROC) curve is a plot of sensitivity as a function of $1 -$ specificity for the possible cut-offs π_o. This curve usually has a concave shape with a 45^0 line connecting the points $(0, 0)$ and $(1, 1)$; at this point the area under the curve (AUC) is 0.5. The higher the area under the curve, the better the predictions. PROC LOGISTIC in SAS can be used to plot the ROC curve for the model. The area under a ROC curve is identical to the value of another measure of predictive power, which is referred to as concordance index (Agresti, 2002). For instance, let's consider all pairs of observations $(i,\ j)$, such that $y_i = 1,$ and $y_j = 0$. Then the concordance index c estimates the probability that the predictions and the outcomes are concordant. A value $c = 0.5$ means that the predictions were no better than random guessing. Generally, the ROC curves are a popular way of evaluating diagnostic tests. Sometimes such tests have $J > 2$ ordered response categories, rather than (positive, negative). The ROC curve then refers to the various possible cut-offs for defining a result to be positive. It plots sensitivity against $1 -$ specificity for the possible collapsing of the J categories to a (positive, negative) scale (Agresti, 2002). Decisions on the area under the curve, which is also referred to as the probability that the predicted probability assigned to the event $y = 1$ is higher than the non-event $y = 0$, is as follows: if the area under the curve (AUC) is less than 0.6, this suggests that the prediction accuracy of the model is poor; and if the area under the curve is between $0.6 - 0.7$, this suggests that the prediction of the model is good; where as if the area under the curve is between $0.7 - 0.8$ this suggests that the predication accuracy of the model is very good, and when the area under the curve is between $0.9 - 0.1$ this suggests that the prediction accuracy of the model is excellent. Figure 4.1 displays a ROC curve for two models. The AUC for model 1 is less than the AUC for model 2. This implies that

Model 2 has a better predictive accuracy power than model 1. Thus, model 2 is preferable to model 1. Also refer to Satchell and Xia (2006) for more discussion.

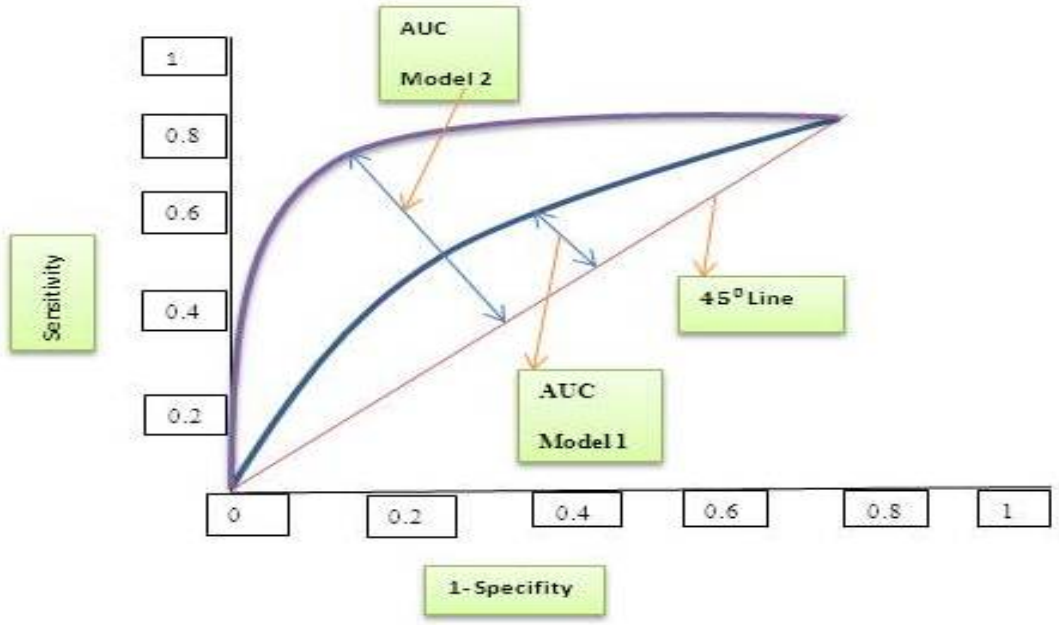

Figure 4.1: Roc Curves Plot (Sensitivity against 1-specificity)

4.3 Results

In this Section we present the results of the analysis after fitting the logistic regression model to the data. First, we run a full main-effect model consisting of all the fourteen explanatory variables discussed in Section 3.2. In an attempt to capture interaction effect that affects business success as well as improve the deviance in the analysis, two-way interaction tests were done and all insignificant interactions dropped. The inclusion process of the interaction effects are done step by step until the last significant interaction effects are retained in the model. Therefore the final model includes all the main-effects, and three two-way interaction effects presented in Tables 4.6 and 4.7. Cash flow problems and type of industry were found to be significant main effects. In addition to the main effect, three two-way interaction

effects were also found to be significant. The significant two-way interaction terms are: the use of internet and stakeholders; outstanding loans and education; and state and gender. We then use PROC GENMOD SAS procedure for further model checking; that is to say, model checking of goodness of fit, over-dispersion, influential, and model predictive accuracy power are discussed.

4.4 Model Diagnostics

The logit link, the probit link, and the complementary log-log link were used to select the model of interest. It is clear from Figures 4.2 and 4.3 that there are no extreme values, that is to say the that residual plot indicates no significant model inadequacy, and the Cook's distance plot also indicates no influential observations and outliers. In practice, as discussed by Dobson (2002) and Agresti (2002), if there are no extreme values, then the logistic and the probit regression models provide similar fits. That is to say that the logit and probit link functions are appropriate when the values of π_i are around 0.5. However with the logit model, one can estimate the odds ratio for the data, unlike with the probit model. Recall that for the binomial distribution, the natural parameter ($g[\mu]$) is the logit of the probability of success, and the logit link is its canonical link. Table 4.4 shows that the logit link is an appropriate link function. We therefore prefer the logit model to the probit model, and since there are no extreme values, the complementary log-log link, which is appropriate for modelling data with extreme values, is dropped. The overall fitness of the model is significant, as presented in Table 4.2. We checked the model goodness of fit using the Hosmer and Lemeshow goodness of fit test, the Pearson chi-square, the Deviance scale, and the AIC, the residual and influence tests, as discussed in Section 4.2.

Table 4.2: Overall Model Significance Test

Test	**Chi-Square**	**DF**	**p-value**
Likelihood Ratio	469.2326	44	< 0.0001
Score	438.0516	44	< 0.0001
Wald	342.5799	44	< 0.0001

Recall that the response variable is binary which follows a binomial distribution. It is

then important to check for over-dispersion, as discussed in Section 4.2. From Table 4.3, the scale parameter ϕ was estimated by the square root of the deviance over the degree of freedom, and it is equal to 0.9771, which implies that there is no over-dispersion. To check for consistency, we compared the scale parameter with Pearson chi-square in Table 4.3, which equals 1.075 this value is approximately equal to 1. This also implies that there is no over-dispersion.

Table 4.3: Criteria for Assessing Model Goodness of Fit

Criterion	**DF**	**Value**	**Value/DF**
Deviance	1538	1502.8462	0.9771
Scale Deviance	1538	1502.8462	0.9771
Pearson Chi-square	1538	1653.3647	1.075
Scale Pearson X2	1538	1653.3647	1.075
Log Likelihood		-947.8383	
Full Log Likelihood		-826.6851	
AIC (small is better)		1725.3701	
AICC (small is better)		1726.7441	
BIC (small is better)		1926.568	

The choice of the link function is very crucial for the consistency of the model. As discussed in Section 4.2, we refitted the model using the linear predictor and the squared linear predictor as explanatory variables using SAS PROC GENMOD. The results shown in Table 4.4 suggest that the link function is appropriate since the linear predictor is highly significant and the squared linear predictor is insignificant. This implies that the prediction given by the linear predictor is not improved by adding the square linear predictor term, and thus suggests consistency of the choice of the link function.

Table 4.4: Criteria for Assessing the Link Function

Effect	**DF**	**Estimate**	**Standard Error**	**Wald χ^2**	***p*-value**
Intercept	1	-0.0028	0.068	0.0001	0.9666
Linear predictor	1	1.0717	0.1381	60.25	< 0.0001
Squared linear predictor	1	-0.0111	0.0647	0.03	0.8637

Assessment of the model goodness of fit was done using the Hosmer-Lemeshow goodness of fit statistic test. As discussed in Section 4.2, Hosmer and Lemeshow recommended that observations be partitioned into 10 groups, according to their probabilities. The Hosmer

and Lemeshow partition, together with their observed and expected probabilities, are shown in Table 4.5. The Hosmer and Lemeshow chi-square for the logit model is 4.9323, with $d.f$ 8 and a p-value of 0.7648. The insignificant p-value and the high value for the chi-square, implies that there was no lack of fit when the model was fitted to the data. This result is consistent with results from the over-dispersion check and the choice of the link function, which suggests consistency of the model fit.

Table 4.5: Partition for the Hosmer and Lemeshow Test

Group	**Total**	**Y = 1**		**Y = 0**	
		Observed	**Expected**	**Observed**	**Expected**
1	198	73	72.91	125	125.09
2	199	89	89.52	110	109.48
3	198	95	102.55	103	95.45
4	198	129	119.4	69	78.6
5	199	143	143.47	56	55.53
6	198	171	168.35	27	29.65
7	198	176	176.37	22	21.63
8	198	176	180.47	22	17.53
9	198	184	184.46	14	13.54
10	192	185	183.5	7	8.5

As discussed in Section 4.2, Figure 4.2 shows the deviance residual plot against the linear predictors. Since the residual plot shows no systematic pattern, we therefore conclude that there is no significant model inadequacy, and that the presence of the influential observations and outliers detected by the residual plot are not influential, and thus are observed in the covariates space. This check also suggests consistency of the model fit.

Model fit can also be assessed using leverage and influence diagnosis. For instance, Agresti (2002) reported that the fit could be different if an observation that appears to be an outlier on y, which has large leverage, is deleted from the model. Recall from Section 4.1, that the influence of a single observation on the parameter is measured by Cook's distance. Figure 4.3 indicates that there are no influences on the parameter, since the outliers of the Cook's distance are all less than 1. However from our plot three outliers are detected from the observations. These observations are: 1 679 with predicted probability of 0.25; observation 1 711 with predicted probability of 0.25; and observation 1 378 with predicted probability of 0.3. To confirm whether these observations have any influence on the estimated coefficients,

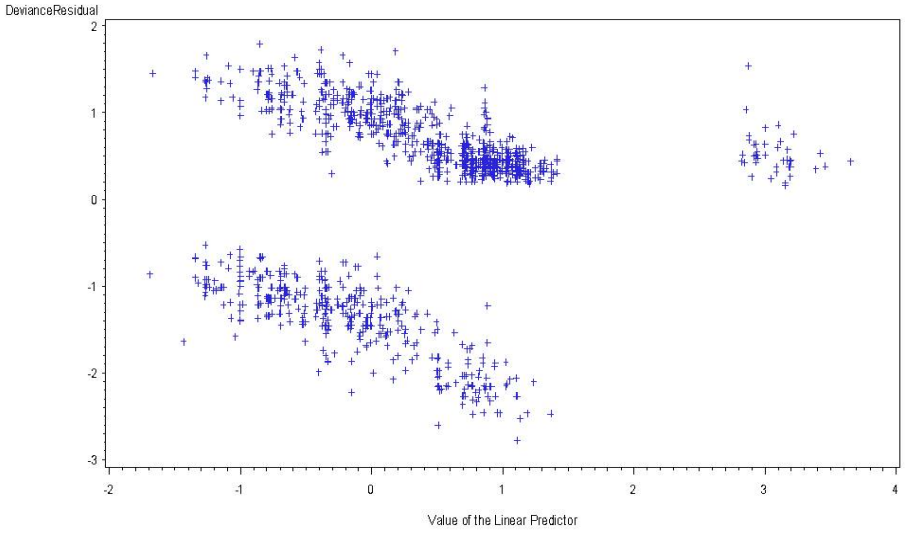

Figure 4.2: Residual Plot for Logit Model

we refit the model by deleting these observations one step at a time and still find that their inclusion or exclusion from the model has no significant influence on the model estimated coefficients. This model check confirms the consistency of the model fit and suggests that the logit model fitted the data well.

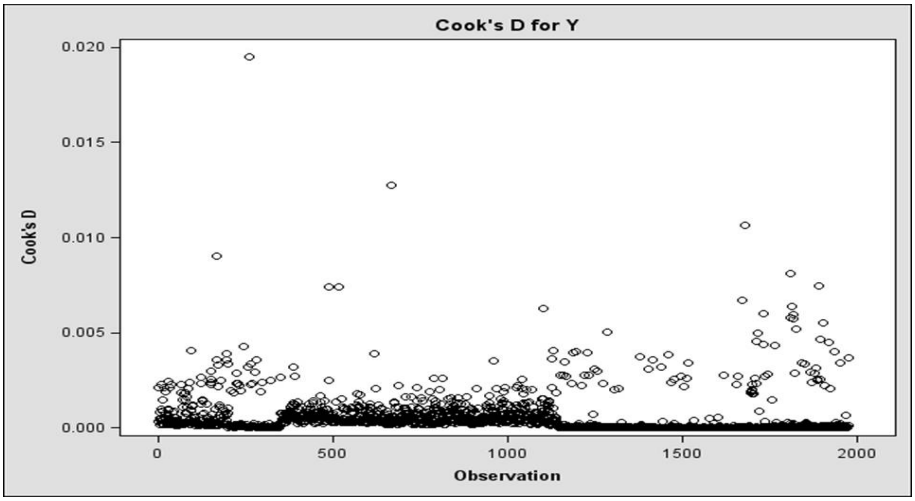

Figure 4.3: Cook's Distance Plot for Logit Model

The receiver operating characteristic (ROC) curve was used to plot sensitivity as a function of 1- specificity for the possible cut-offs π_o, as discussed in Section 4.2. The predictive

accuracy tells us how best the model can predict business success. Recall that a value of 0.5 of the area under the curve means that the predictions are poor, whereas if it approaches 1, the better the predictive power. The area under the curve for the logit model is 0.7911; this value assesses the predictive accuracy of the model in predicting business success. The area under the curve is the proportion of the correctly predicated probabilities, thus it implies that about 79.11% of the probabilities of business success are predicted correctly, which is a very good predictive accuracy of the model. This model check also confirms the consistency of the model fit. The *ROC* curve is given in Figure 4.4.

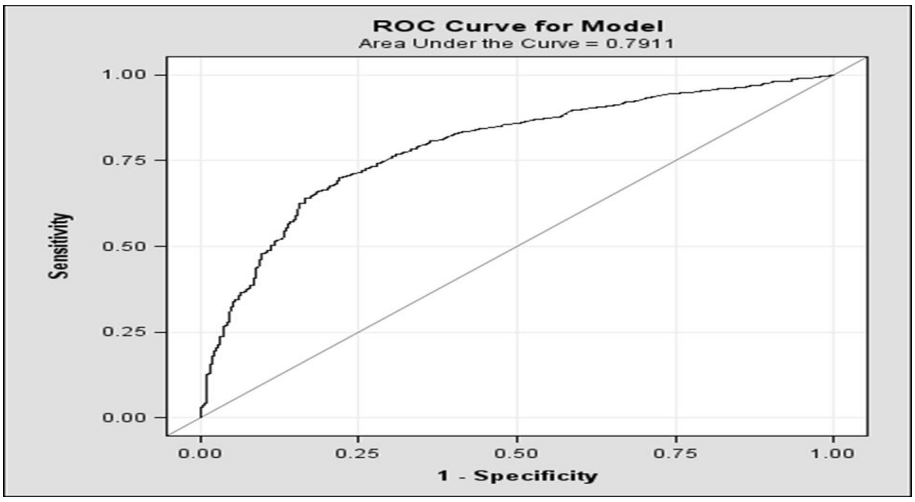

Figure 4.4: Roc Curve for Logit Model

Interpretation of the (Odds Ratio) Model Coefficients

The odds ratios were calculated from the estimated model coefficients in Table 4.6. The odds ratio for business location and the other covariates, together with their standard errors and p-values, are also presented in Table 4.6. The odds ratio for business location is compared with Juba as a reference. This is because Juba is the capital city and has the largest sample size compared to the other locations (Table 3.1).

From Table 4.6, it can be observed that the effect of business ownership structure is insignificant. This implies that controlling for the other covariates, business success does not differ between businesses that are classified as partnerships, companies, sole proprietorships, and others. The effect of government support to businesses is also insignificant. This implies

Table 4.6: Parameter Estimates and Odds Ratio for the Main Model

Effect	**Estimate**	**Odds Ratio**	**Standard Error**	P**-value**
Intercept	2.4503		0.8058	0.0024
State(Ref=Juba)				
Malakal	-1.6793	0.18	0.2462	< 0.0001
Bor	0.4104	0.51	0.8230	0.6180
Bentiu	-2.6003	0.07	0.5112	< 0.0001
Kuajok	-2.690	0.07	0.5012	< 0.0001
Aweil	-3.7470	0.02	0.5779	< 0.0001
Wau	-2.4525	0.09	0.4306	< 0.0001
Rumbek	-1.9334	0.14	0.4268	< 0.0001
Yambio	-0.7587	0.47	0.5002	0.1293
Torit	-0.0126	0.99	0.569	0.9824
Ownership(Ref=Sole Proprietorship)				
Partnership	0.0799	1.08	0.1668	0.6320
Company	0.0874	1.09	0.2546	0.7314
Others	0.3561	1.42	0.304	0.2414
Internet(Ref=No)				
Yes	-0.833	0.43	0.2783	0.0028
Government Support(Ref=No)				
Yes	-0.016	0.98	0.2964	0.9569
Cash flow Problem(Ref=No)				
Yes	0.7156	2.05	0.2223	0.0013
Financial Loss due to shock(Ref=No)				
Yes	-0.2391	0.79	0.3387	0.4802
Business Size(Ref=Small Enterprise)				
Micro enterprise	0.1789	1.95	0.2433	0.4622
Business Age	0.0382	1.03	0.1804	0.8321
Outstanding Loan(Ref=No)				
Yes	-0.2655	0.76	0.1385	0.0553
Startup Capital	0.0221	1.02	0.0269	0.4130
Gender(Ref=Female)				
Male	-0.3146	0.73	0.1323	0.0174
Education Level(Ref=V. training)				
No schooling	0.3271	1.38	0.3486	0.3481
Primary school	0.0431	1.04	0.8000	0.9051
Secondary school	0.0049	1.00	0.3368	0.9882
University Degree	0.3627	1.44	0.3794	0.3391
Stakeholders(Ref=Foreign)				
Local	-0.6873	0.50	0.498	0.1676
Type of Industry(Ref=Other Services)				
MEMC	1.1429	3.13	0.4112	0.0054
TT	0.5565	1.74	0.278	0.0453
APSA	0.717	2.05	0.3921	0.0675
ESHS	0.674	1.96	0.3368	0.0454

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

Table 4.7: Continuation of Parameter Estimates and Odds Ratio for the Main Model

Effect	**Estimate**	**Odds Ratio**	**Standard Error**	*P*-**value**
Internet*Stakeholders(Ref=No and Foreign)				
Yes and Local	1.717	5.568	0.7136	0.0161
Loan*Education(Ref=No and V.training)				
Yes and No schooling	-1.9562	0.141	0.8827	0.0267
Yes and Primary school	-1.4217	0.241	0.8973	0.1131
Yes and Secondary school	-1.1603	0.313	0.8792	0.1869
Yes and university degree	-1.8401	0.159	0.9303	0.0479
State*Gender(Ref=Juba and Female)				
Malakal and Male	-1.0253	0.359	0.4920	0.0372
Bor and Male	-1.8047	0.165	0.8480	0.0333
Bentiu and Male	0.0606	1.062	0.5693	0.9152
Kuajok and Male	-0.8884	0.411	0.5360	0.0974
Aweil and Male	0.6194	1.858	0.6242	0.3210
Wau and Male	-0.4364	0.646	0.4836	0.3669
Rumbek and Male	-0.1142	0.892	0.4974	0.8184
Yambio and Male	0.3281	1.388	0.6554	0.6167
Torit and Male	-0.4995	0.607	0.6222	0.4221

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

that the odds of success among businesses that receive support from the government for business investment is not different from businesses that did not received support from the government. On the other hand, the effect of cash flow problems is found to be a significant factor for business success. That is to say that the odds of sucess for businesses with no cash flow problems is twice the odds (2.05, p-value 0.0013) of success for businesses with cash flow problems. Since the effect of financial loss due to shock is insignificant, this implies that, controlling for the other covariates, the odds of success among businesses with financial loss due to shock is not different from that of businesses with no financial loss due to shock. Similarly, the effect of business size on business success is insignificant. This implies that controlling for the other covariates, the odds of success among micro-enterprises is not different from small enterprises. Likewise, the effect of business age on business success is found to be insignificant, implying that controlling for the other covariates, the odds of success for businesses that have long been long in operation is not different from those businesses that shorly been in operation. The effect of startup capital on business success is insignificant, implying that controlling for the other covariates, the odds of success among

businesses with low startup capital is not different from businesses with higher startup capital for micro and small enterprises. The odds ratio for type of industry indicates that the odds of success for businesses investing in mining, energy, manufacturing, and construction is 3.13 (p-value 0.0054) times greater than the odds of success for businesses investing in the other industry sectors. Likewise, the odds of success for businesses investing in trade and transportation activities is 1.74 (p-value 0.0453) times greater than the odds of success for businesses investing in the other industry sectors. On the other hand the odds of success for businesses investing in administrative, professional, and scientific services is insignificant, implying that controlling for the other covariates, the odds of success is not different from businesses investing in the other industry sectors. The odds of success for businesses investing in education, social, and health services is 1.96 (p-value 0.0454) times greater than the rate of success for businesses engaged in the other industry sectors. Note that other industry sectors comprises of those businesses that are investing in agriculture, forestry, fisheries, water supply, sewage and waste management, and other services.

The results shown in Table 4.6 suggest that liquidity problems which are also referred to as cash flow problems are the cause of business failure. As discussed in Chapter 2, most micro and small businesses used external sources to finance their operations usually through financial institutions, venture capitalists, and individual investors. Generally, cash flow problems may be caused by poor credit arrangements, or a lack of proper sources of financing which is crucial for business success. Our results support the argument by Ricketts (2002) and the findings by Mole (2007), that due to lack of genuine collateral, banks may avoid the risk of businesses failing to refund credit, and hence hinder small enterprises from accessing finance. Industry sector was also found to be a significant factor in business success. Our findings suggest that investors' decisions on which industry sector to invest in are very crucial. As reported by Spulber (2009), investing in a specific industry sector may be influenced by consumers choosing to purchase a specific product or service of that specific industry. The study results support the findings of Papadaki and Chami (2007) which stated that products and services from businesses that are engaged in the retail and personal services sectors may be easily imitated, thus fostering success. In contrast, participation in industry

sector of more professional services that require specific sets of capabilities or requirements, developed through prior experience or education, render imitation difficult.

Interaction Effects

The odds ratio for the two-way interaction effects, that is to say the use of internet by stakeholders, outstanding loans by education, and state by gender, are presented in Table 4.7.

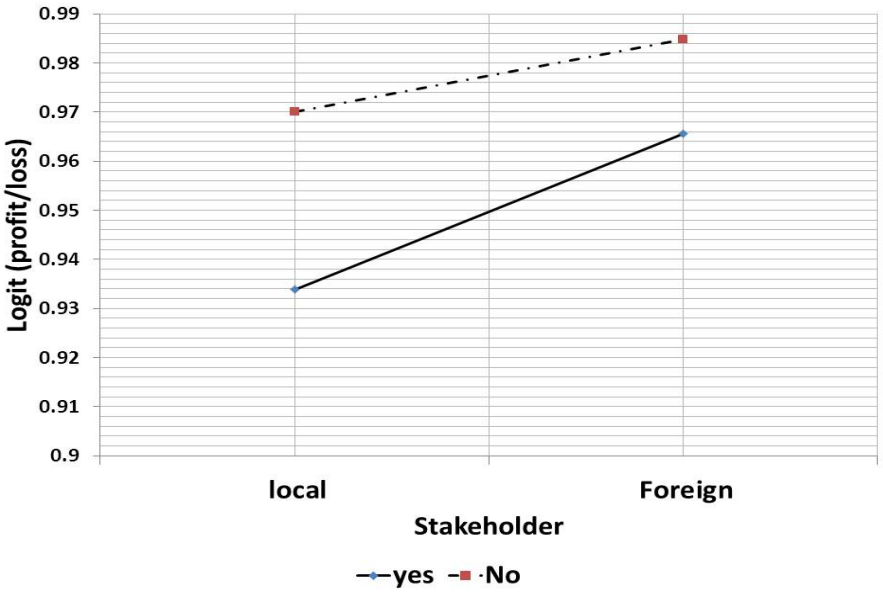

Figure 4.5: Log-odds Associated with Use of Internet by Stakeholders

The graphical display of the effect of the log-odds of internet use and stakeholders is given in Figure 4.5. The odds of success of local business stakeholders who are internet users is 5.568 (p-value 0.016) times the odds of success of foreign business stakeholders who are not using the internet. The discussion of technology as a key factor for business innovation has been investigated by some researchers. Dibrell *el al* (2008) findings for instance, suggested that the impact of innovation on business performance was indirectly influenced by information technology. Our findings suggest that use of the internet by local and foreign stakeholders is significantly related to business success.

Figure 4.6 displays the log-odds effect of business outstanding loans and entrepreneur level of education. The odds of success for entrepreneurs with no qualifications and who have

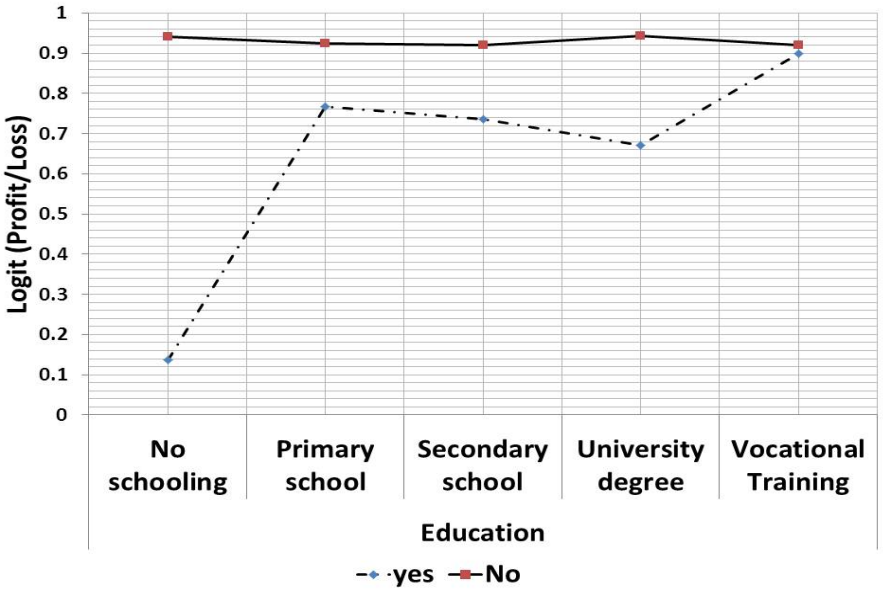

Figure 4.6: Log-odds Associated with Business Outstanding Loans by Education

outstanding loans is 0.141 (p-value 0.0267) times the rate of success for entrepreneurs who have undergone vocational training and who don't have outstanding loans. The success rate for entrepreneurs who have a university degree and who have outstanding loans is 0.159 (p-value 0.0479) times the success rate for entrepreneurs who have undergone vocational training and who don't have outstanding loans. On the other hand, the effect of entrepreneurs who have completed primary school and those who have completed secondary school and have outstanding loans is insignificant, implying that, controlling for the other covariates, their success rate is not different from entrepreneurs who have undergone vocational training and who don't have outstanding loans. Papadaki's and Chami's (2007) findings suggest that higher education is likely to increase an entrepreneur's ability to cope with problems and to seize opportunities that are important to business success and growth. The study found a significant effect of outstanding loans by different levels of education, which suggests that the rate of success is higher for entrepreneurs with higher education, who have the ability to cope with problems related to loans.

Figure 4.7 displays the log-odds of the effect of state (business location) and gender. The rate of success for male entrepreneurs in Malakal is 0.359 (p-value 0.0372) times the rate of success for female entrepreneurs in Juba. Similarly, the success rate for male entrepreneurs

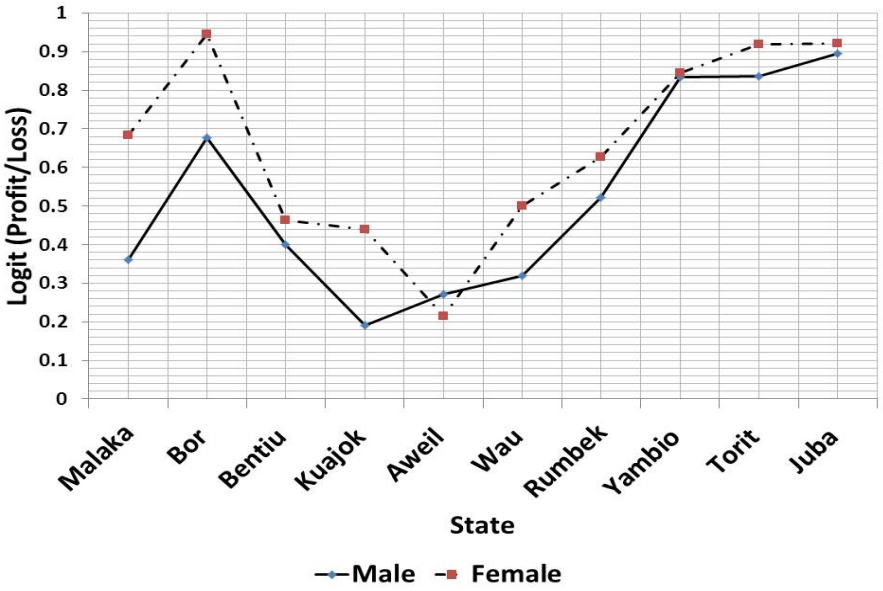

Figure 4.7: Log-odds Associated with Business Location by Gender

in Bor is 0.165 (p-value 0.0333)times the success rate for female entrepreneurs in Juba. The effect of male entrepreneurs in Bentiu, Kuajok, Aweil, Wau, Rumbek, Yambio and Torit is insignificant. This Implies that controlling for the other covariates, the success rate for male entrepreneurs in these locations is not different from female entrepreneurs in Juba. The study findings suggest that the rate of business success among male entrepreneurs and female entrepreneurs differs from location to location across the country. Moreover, Figure 4.7 reveals that for Yambio and Juba, both male and female entrepreneurs have the same rate of business success.

It must be noted that the results discussed in this Chapter do not take into account the effect of data structure, that is to say the effects of the sampling design, the random selection of the sampling units (PSU), and the weights from which the data is obtained. In order to take into account the effects of the sampling design, the sampling weights, and the random and fixed effects for business survey data, we introduce the survey logistic model in Chapter five and the generalized linear mixed model in Chapter six. The survey logistic model is suitable for modelling data from a complex survey design. On the other hand, the GLMM is useful for solving the problem of over-dispersion, which is commonly observed among non-normal distributed responses, as well as for modelling dependence among responses in

repeated measures (longitudinal data), by incorporating both the random and the fixed effect estimates.

Chapter 5

Survey Logistic Regression Model

5.1 Introduction

The survey logistic regression model is a logistic regression model used to model data from a complex survey design. It usually accounts for the complexity of survey design, that is to say 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 the variance estimates. If the data is from a simple random sampling then the survey logistic and the ordinary logistic give identical estimates. But if the data is from a complex survey design, then the estimates of the coefficients and the standard errors will be different because of the effects of stratification and clustering. In this chapter we discuss the effects of both sampling-survey design and weights on the data structure. The sampling technique used in this study, as discussed in chapter 3, is based on stratified random sampling which is done in two stages. In the first stage, enumeration areas (EAs) (which are also referred to as PSU) are sampled in each stratum. The stratums are made up of business locations (state capitals), industry, and number of employees. In the second stage, businesses are sampled. Let the response variable be denoted by y_{ijh} ($i = 1, 2, \ldots, m_{hj}$, $j = 1, 2, \ldots, n_h$, and $h = 1, 2, \ldots, H$), which equals 1 if profit is present in the i^{th} business within j^{th} PSU, nested within h^{th} stratum, and 0 otherwise. Noted that, h is the stratum, j is the cluster and i is the individual business. Let $\pi_{ijh} = p(y_{ijh} = 1)$ be the probability that profit is present in the i^{th} business within j^{th} PSU

nested within h^{th} stratum. Then the survey logistic model is given by

$$\log(\pi_{ijh}) = x'_{ijh}\beta, \quad i = 1, 2, \ldots, m_{hj}, \quad j = 1, 2, \ldots, n_h, \quad h = 1, 2, \ldots, H \tag{5.1}$$

Where x_{ijh} is the row of the design matrix corresponding to the characteristics of the i^{th} business in j^{th} PSU, nested within h^{th} stratum; and β is the vector of unknown parameters of the model. Hence, the log likelihood function is given by

$$l(\beta; y) = \sum_{h=1}^{H} \sum_{j=1}^{n_h} \sum_{i=1}^{m_{hj}} \left[y_{ijh} log \left(\frac{\pi_{ijh}}{1 - \pi_{ijh}} \right) - log \left(\frac{1}{1 - \pi_{ijh}} \right) \right] \tag{5.2}$$

Parameter estimation for the survey logistic regression model can be calculated through a number of methods. According to Heeringa *et al.* (2010), the likelihood function for a simple random sampling of n observation on a binary response variable y with possible values 0 and 1, is based on the binomial distribution

$$l(\beta|x) = \prod_{i=1}^{n} \pi(x)^{yi} [1 - \pi(x)]^{1-yi} \tag{5.3}$$

where $\pi(x)$ is linked to the regression model coefficients and evaluated through the logistic CDF

$$\pi(x_i) = \frac{exp(x_i\beta)}{1 + exp(x_i\beta)} \tag{5.4}$$

For this case, the logistic regression model parameters and standard errors can be estimated using the method of maximum likelihood discussed in Section 4.1. 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,...,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 PROC SURVEYLOGISTIC procedure in SAS (9.2). For our case, the default Taylor series (Linearization) method is used because it estimates variance from among the PSU.

When survey data is collected using a complex sample design with unequal cluster size, most of the statistics of interest will not be simple linear functions of the observed data (Heeringa *et al.* 2010). Thus, Taylor's (Linearization) method is applied, in order to derive an approximate form of estimator that is linear in statistics and for which variances and covariances can be directly and easily estimated. Taylor's method is the most commonly used method to estimate the covariance matrix of the regression coefficients for complex survey data. It is the default variance estimation method used by PROC SURVEYLOGISTIC in SAS (9.2). Generally, variance estimation can be estimated using the Taylor series (Linearization) method as follows; the estimated covariance matrix of the model parameter $\hat{\beta}$ by the Taylor series method is given by

$$\hat{V}(\hat{\beta}) = \hat{Q}^{-1}\hat{G}\hat{Q}^{-1} \tag{5.5}$$

where

$$\hat{Q} = \sum_{h=1}^{H}\sum_{j=1}^{n_h}\sum_{i=1}^{m_{hj}} w_{hji}\hat{D}_{hji}(diag(\hat{\pi}_{hji} - \hat{\pi}_{hji}\hat{\pi'}_{hji})^{-1}\hat{D'}_{hji}$$

$$\hat{G} = \frac{n-1}{n-p}\sum_{h=1}^{H}\frac{n_h(1-f_h)}{n_h-1}\sum_{j=1}^{n_h}(e_{hj} - \overline{e}_h)(e_{hj} - \overline{e}_h)'$$

$$e_{hj} = \sum_{i=1}^{m_{hj}} w_{hji}\hat{D}_{hji}(diag(\hat{\pi}_{hji} - \hat{\pi}_{hji}\hat{\pi'}_{hji})^{-1}(y_{hji} - \hat{\pi}_{hji})$$

$$\overline{e}_h = \frac{1}{n_h}\sum_{j=1}^{n_h} e_{hj}$$

- D_{hji} is the matrix of the partial derivatives of the link function g, with respect to β and $\hat{D}_{hji}$, and the response probabilities $\hat{\pi}_{hji}$ evaluated at $\hat{\beta}$.
- $h = 1, 2, \ldots, H$ is the stratum index, $j = 1, 2, \ldots, n_h$ is the cluster index, $i = 1, 2, \ldots, m_{hj}$ is the observation index within cluster j of stratum h.
- n is the total sample size.
- y_{hji} is a D-dimensional column vector whose elements are indicator variables for the first D categories for variable Y. If the response of the j^{th} unit of the i^{th} cluster in stratum h 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_{hji} is the sampling weight.
- π_{hji} is the expected vector of the response variable.
- f_h is the sampling rate for stratim h
- Finally, p is the number of covariates in the model.

Also see the work of 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 $z_{1-\frac{\alpha}{2}}$ is the $100(1-\frac{\alpha}{2})^{th}$ percentile of the standard normal distribution, and σ_j^2 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 (PROC LOGISTIC) and the survey logistic model (PROC SURVEYLOGISTIC) 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.

Generally, model selection and evaluation is based on a subjective combination of theoretical issues, inspection of parameter estimates, goodness of fit test, interpretability, and a comparison of the performances of competing models (Marsh and Balla, 1994). The process of model selection for the survey logistic regression model follows the same procedures as for the ordinary logistic regression model. The only difference is that the selection procedures (forward, backward, and stepwise) are not yet incorporated in SAS PROC SURVEYLOGISTIC. Model selection can however, be done using the type 3 analysis of effects, where insignificant variables are excluded from the model one at a time, and the contribution of the remaining effects to the deviance reduction, are then observed. Generally, as in all regression modelling, identification of a best logistic regression model for survey data should follow a systematic and scientific governed process (Heeringa *et al.* 2010). Hosmer and Lemeshow (2000) reported that the process for model selection should start by specifying the initial model (saturated model), refining the set of predictors, and then determining the final form of the logistic regression model that explains the data well. Among many model selection procedures that exist, no single one is always the best. Thus one should accept that any model is a simplification of reality (Agresti, 2002). Other criteria for model selection include the Akaike's Information Criteria (AIC) and the Schwarz Criteria (SC), which are used to impose penalties to the likelihood ratio statistic $-2logL$ (Agresti, 2002). Generally, the decision on either whether 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. However, if the interest is in the order of the model, then a model based on SC is preferred (refer to Buckland *et al.* (1997), and Burnham and Anderson, (2002) for further discussion on model selection).

The goodness of fit test for logistic regression for data from complex sampling designs is not yet developed or implemented in many available softwares; usually simulation studies and results were compared to the ordinary goodness of fit statistics discussed in Section 4.1 (Archera *et al.* 2006). For instance, the Hosmer and Lemeshow goodness of fit test statistic, the Pearson residual, and the deviance residual test discussed in Section 4.1, are not yet incorporated in SAS PROC SURVEYLOGISTIC. Thus for the assessment of our model goodness of fit, we used the Akaike Information Criterion (AIC), the Schwarz Criterion (SC), and the -2log likelihood statistic as approximations for the goodness of fit test. These three criteria are displayed using the SURVEYLOGISTIC procedure (SAS 9.2). Let s denote the explanatory effect in the model, and let $\hat{\pi}_j$ be the probability of the observed response for the j^{th} observation. Then using the SAS notation, the model evaluation of goodness of fit can be done through the three criteria as follows:

-2Log-likelihood

The -2Log-likelihood statistic test is given by

$$-2LogL = -2\sum_j w_j f_j log(\hat{\pi}_j)$$

where w_j and f_j are the weights and the frequency values of the j^{th} observation. For the binary response model, the -2Log-likelihood is an equivent to

$$-2LogL = -2\sum_j w_j f_j [r_j log(\hat{\pi}_j) + (n_j - r_j) log(1 - \hat{\pi}_j)]$$

where r_j is the number of events, n_j is the number of trials, and π_j is the estimated event probability. This can be displayed using the event/trial syntax in SAS (9.2).

Akaike Information Criterion

The AIC measures the discrepancy or symmeteric distance between a full (saturated) model and a reduced (selected) model of interest. The AIC is given by

$$AIC = -2LogL + 2p$$

where p is the number of parameters in the model. The AIC strategy is used to select the best model among two or more competing nested models. Thus the model with the smallest AIC is the best.

Schwarz Criterion

The Schwarz Criterion is another criterion used to asses/evaluate model goodness of fit. Like the AIC, this method adjusts the -2LogL statistic for the number of parameters. The Schwarz Criterion is given by

$$SC = -2LogL + plog(\sum_j f_j)$$

where p is the number of parameters in the model. It can be noted here that for cumulative response models for both the SC and the AIC assessments of fit, $p = k + s$, where k is the total number of response levels minus one, and s is the number of explanatory effects. For the generalized logit model, $p = k(s + 1)$.

The assessment of model predictive accuracy power in the survey logistic model can be achieved through statistics such as the concordance index (c), Sommer's D (SD), Goodman-Kruskal Gamma (GKG), and Kendall's Tau-a (KT). Unlike in the logistic model, where we used the ROC curve to assess the predictive power of the model, in the survey logistic model we used the concordance index c, which is equal to the area under the receiver operating characteristic (ROC) curve, to assess the model predictive power. The concordance index, as discussed in Section 4.2, estimates the probability that the predictions and the outcomes are concordant (Agresti, 2002). The concordance index as displayed by SAS PROC SURVEYLOGISTIC is given by

$$c = [(n_c + 0.5)(t - n_c - n_d)t^{-1}]$$

where n is the total number of observations in the data set, t is the total number of pairs given by $n(n-1)/2$, n_c is the number of concordant pairs, n_d is number of discordant and $t - n_c - n_d$ is the number of tied pairs. According to Agresti (2002), a value $c = 0.5$ means that the predictions were no better than random guessing. However as c approaches 1, the better the model predictive power. Refer to Section 4.2 for decisions on the AUC for the ROC, which is equivalent to the concordance index c.

5.2 Results

Generally, stratification and clustering of the complex sample design involved in most research 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 model in Table 4.3, and Table 4.4 using PROC SURVEYLOGISTIC. First we present a type 3 analysis of effect in Table 5.1, and then present the parameter estimates for the main effects model in Table 5.3 and table 5.4. The model fit statistics and prediction of model accuracy power is presented in Table 5.2. The odds ratios together with their standard errors and p-values are presented in both Table 5.3 and Table 5.4. Lastly we present comparisons of parameter estimates, together with their standard errors produced by both PROC LOGISTIC and PROC SURVEYLOGISTIC, in Table 5.5 and Table 5.6 respectively. The results from Table 5.1 indicate that business location, internet-use, government support, cash flow problems, outstanding loans, and startup capital are significant factors for business success.

After fitting the survey logistic model, we presented some diagnostic statistics tests using the AIC and the SC selection criteria, which are presented in Table 5.2. The AIC and the SC are basically statistics used for model selection. However, these statistics can also be used as approximations to compare two or more competing models. Generally, the model with the lowest AIC value is selected, especially when the objective of the study is to check or measure the consistency of the model (Burnham and Anderson, 2002). Whereas a model based on the SC is selected if the interest is on the order of the model.

As discussed in Section 5.1, we used the concordance index (c) which is equivalent to

Table 5.1: Type 3 Analysis of Effects for the Survey Logistic Model

Effect	**DF**	**Wald χ^2**	***P*-value**
State (business Location)	9	39.3583	$< .0001$
Ownership	3	10.6831	0.0987
Internet	1	82.4689	$< .0001$
Government Support	1	9.9475	0.0069
Cash flow Problem	1	8.2024	0.0166
Financial Loss due to shock	1	3.3357	0.0678
Business Size	1	0.5045	0.4775
Business Age	1	0.1848	0.6673
Outstanding Loan	1	4.2526	0.0392
Startup Capital	1	24.0054	0.0076
Gender of Owner	1	0.0408	0.8398
Education Level	4	1.7200	0.7871
Stakeholders	1	2.4238	0.1195
Type of Industry	4	6.4981	0.1649
Internet*Stakeholders	1	82.2042	$< .0001$
Loan*Education	4	7.4505	0.1139
State*Gender	9	6.8227	0.6556

the area under the curve for the the *ROC*, to check the model predictive accuracy power. The interpretation 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. The concordance index presented in Table 5.2 suggests that 78.0% of the probability of business success is predicted correctly this is a very good prediction of accuracy for the survey logistic model.

Table 5.2: Model Fit Statistics

Criterion	**Intercept only**	**Survey Logistic Model**
AIC	8554.535	6595.337
SC	8560.123	6846.811
-2 LogL	6526.111	6505.337
c		0.780

Interpretation of the odds ratio for the survey logistic model is the same as for that of the logistic regression model. The only difference is in the confidence interval for the coefficients. That is to say that in the logistic regression model, the confidence interval is narrow due to underestimation of the standard errors. The odds ratios are obtained from the estimates of the model coefficients in Table 5.3, and Table 5.4 respectively.

From Table 5.3, it can be observed that the effects of both partnership and company type

Table 5.3: Parameter Estimates for the Main Effects for the Survey Logistic Model

Effect	**Estimate**	**Odds Ratio**	**Standard Error**	P**-value**
Intercept	3.7028		1.1445	0.0012
State(Ref=Juba)				
Malakal	-1.3191	0.27	0.4349	0.0024
Bor	-0.1775	0.84	0.4246	0.6759
Bentiu	-1.9339	0.14	0.3477	< 0.0001
Kuajok	-2.0873	0.12	0.3857	< 0.0001
Aweil	-2.485	0.08	0.3954	< 0.0001
Wau	-2.3891	0.09	0.3976	< 0.0001
Rumbek	-2.2775	0.10	0.4072	< 0.0001
Yambio	-0.1709	0.84	0.5398	0.7516
Torit	-0.3768	0.67	0.4216	0.3715
Ownership(Ref=Sole Proprietorship)				
Partnership	0.0425	1.04	0.2669	0.8735
Company	0.2999	1.35	0.3717	0.4198
Others	1.0647	2.89	0.5011	0.0336
Internet(Ref=No)				
Yes	-2.1457	0.12	0.9841	0.0292
Government Support(Ref=No)				
Yes	-0.8362	0.43	0.4695	0.0479
Cash flow Problem(Ref=No)				
Yes	0.1619	1.17	0.2075	0.4354
Financial Loss due to shock(Ref=No)				
Yes	0.3669	1.44	0.2141	0.0866
Business Size(Ref=Small Enterprise)				
Micro enterprise	-0.1468	0.86	0.4454	0.7417
Business Age	0.1429	1.15	0.2627	0.5867
Outstanding Loan(Ref=No)				
Yes	2.2952	9.93	1.0091	0.0229
Startup Capital	-1.3928	0.25	0.7082	0.0492
Gender(Ref=Female)				
Male	-0.3796	0.68	0.2427	0.1178
Education Level(Ref=V. training)				
No schooling	0.3418	1.41	0.5574	0.5397
Primary school	0.0286	1.02	0.5470	0.9582
Secondary school	0.2177	1.24	0.4917	0.6579
University Degree	0.3409	1.41	0.6460	0.5977
Stakeholders(Ref=Foreign)				
Local	-1.2244	0.29	0.657	0.0624
Type of Industry(Ref=Other Services)				
MEMC	1.6316	5.11	0.5552	0.0033
TT	0.7055	2.02	0.3622	0.0515
APSA	0.5943	1.81	0.6414	0.3542
ESHS	0.8176	2.26	0.4634	0.0777

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

Table 5.4: Continuation of Parameter Estimates for the Main Effects for the Survey Logistic Model

Effect	**Estimate**	**Odds Ratio**	**Standard Error**	P**-value**
Internet*Stakeholders(Ref=No and Foreign)				
Yes and Local	1.985	7.279	1.0001	0.0472
Loan*Education(Ref=No and V.training)				
Yes and No schooling	-3.0383	0.048	1.0859	0.0051
Yes and Primary school	-2.2612	0.104	1.1084	0.0413
Yes and Secondary school	-2.6349	0.072	1.1096	0.0176
Yes and university degree	-2.8844	0.056	1.257	0.0218
State*Gender(Ref=Juba and Female)				
Malakal and Male	-1.2282	0.293	0.7109	0.0841
Bor and Male	-1.8171	0.162	1.0786	0.0921
Bentiu and Male	-0.3345	0.716	0.8040	0.6774
Kuajok and Male	-0.1026	0.902	0.6755	0.8793
Aweil and Male	0.4082	1.504	0.8249	0.6207
Wau and Male	-1.0681	0.344	0.6094	0.0796
Rumbek and Male	-0.4420	0.643	0.6958	0.5253
Yambio and Male	-0.2950	0.745	1.0914	0.7869
Torit and Male	-0.6891	0.502	0.8074	0.3934

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

of business ownership are insignificant, implying that controlling for the other covariates, the odds of business success is not different from sole proprietorship. However, the odds of success for other types of business ownership is 2.89 (p-value 0.0336) times greater than the odds of success for sole proprietorship. Similarly, the odds of success for businesses that received support from the government is 0.43 (p-value 0.0479) times less than the odds of success of businesses that did not received support from the government. On the other, the effect of cash flow problems on business success is insignificant, which implies that controlling for the other covariates, the odds of success among businesses with cash flow problems is not different from businesses with no cash flow problems. Likewise, the effect of financial loss due to shock is insignificant. This implies that controlling for the other covariates, the odds of business success among businesses with financial loss due to shock is not different from businesses with no financial loss due to shock. The effect of business size on business success is insignificant, implying that controlling for the other covariates, the odds of success among micro-enterprises is not different from that of small enterprises.

Similarly, the effect of business age on business success is insignificant. This implies that controlling for the other covariates, the odds of business success among businesses with few years in operation is not different from businesses with many years in operation. The odds ratio for startup capital indicates that the odds of success for businesses with less startup capital is 0.25 (p-value 0.0492) times less than the odds of success for businesses with more startup capital. The odds ratio for type of industry indicates that the odds of success for businesses that are operating in mining, energy, manufacturing and construction is 5.11 (p-value 0.0033) times greater than the odds of success for businesses operating in the other industry sectors. The odds of success for businesses operating in trade and transport is 2.02 (p-value 0.0515) times greater than the odds of success for businesses investing in the other business industry sectors. The effect of businesses operating in administrative, professional, scientific, educational, social, and health services is insignificant. This implies that controlling for the other covariates, the odds of success for businesses operating in these industry sectors is not different from the other industry sectors. The odds ratio for the

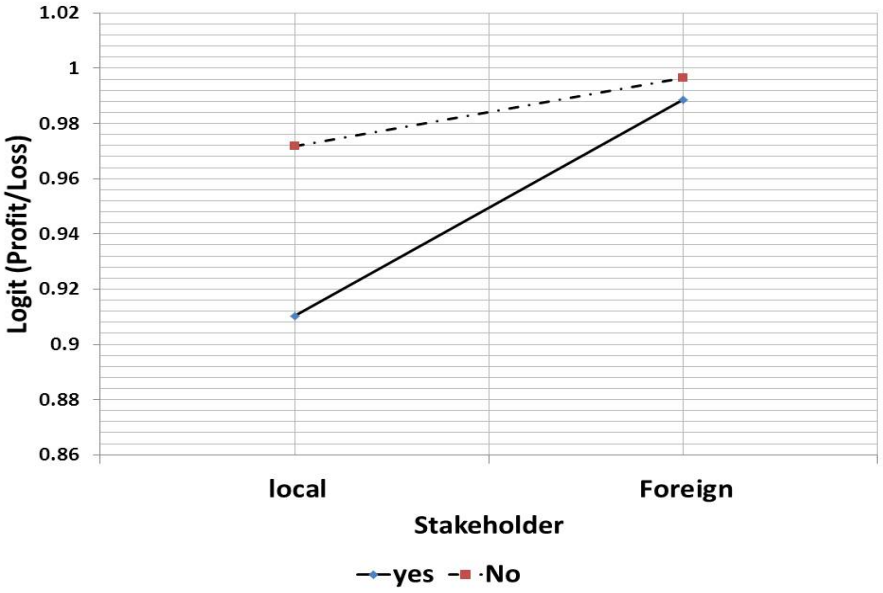

Figure 5.1: Log-odds Associated with Use of Internet by Stakeholders

interaction effects is presented in Table 5.4. It can be observed from Table 5.4 that the odds ratio for the use of the internet and stakeholders' interaction effect indicates that the odds of success for local stakeholders who use the internet is 7.279 (p-value 0.0472) times greater

than the success odds of foreign stakeholders who are not internet users. A graphical display of the log-odds effect of internet-use and stakeholders is presented in Figure 5.1.

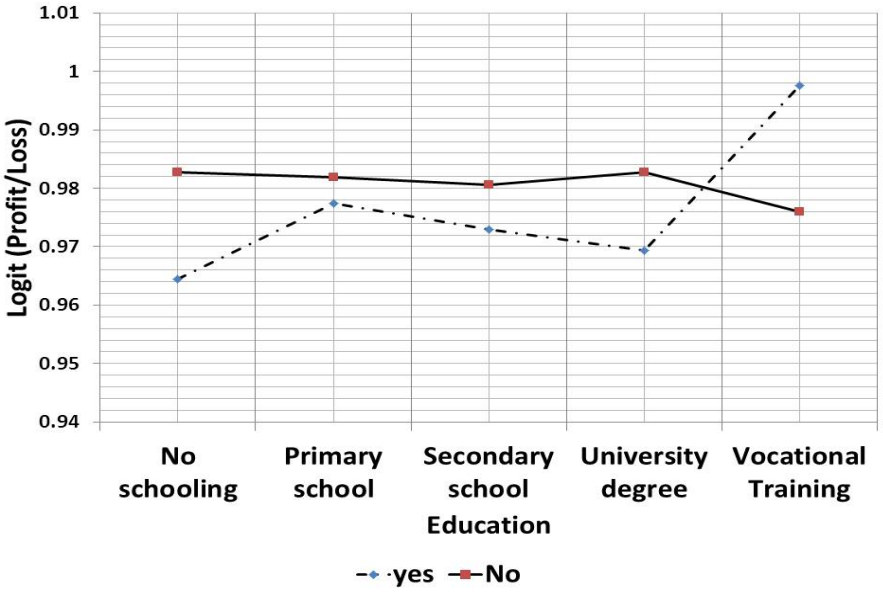

Figure 5.2: Log-odds Associated with Outstanding Loans by Level of Education

Figure 5.2 displays the log-odds effect of outstanding loans by entrepreneur's level of education. The odds of success of entrepreneurs with no qualifications who have outstanding loans is 0.048 (p-value 0.0051) times less than the odds of success of entrepreneurs who have undergone vocational training and who don't have outstanding loans. Similarly, the odds of success for entrepreneurs who have completed primary school and who have outstanding loans is 0.104 (p-value 0.0413) times less than the odds of success of entrepreneurs who have undergone vocational training and who don't have outstanding loans. Likewise, the odds of success for entrepreneurs who have completed secondary school and who have outstanding loans is 0.072 (p-value 0.0176) times less than the odds of success of entrepreneurs who have undergone vocational training and who don't have outstanding loans. The odds of success for entrepreneurs who have university degrees and who have outstanding loans is 0.056 (p-value 0.0218) times less than the odds of success of entrepreneurs who have undergone vocational training and who don't have outstanding loans.

Figure 5.3 displays the log-odds effect of state (business location) by gender. It can be observed that the effect of male entrepreneurs in Malakal, Bor, Bentiu, Kuajok, Aweil,

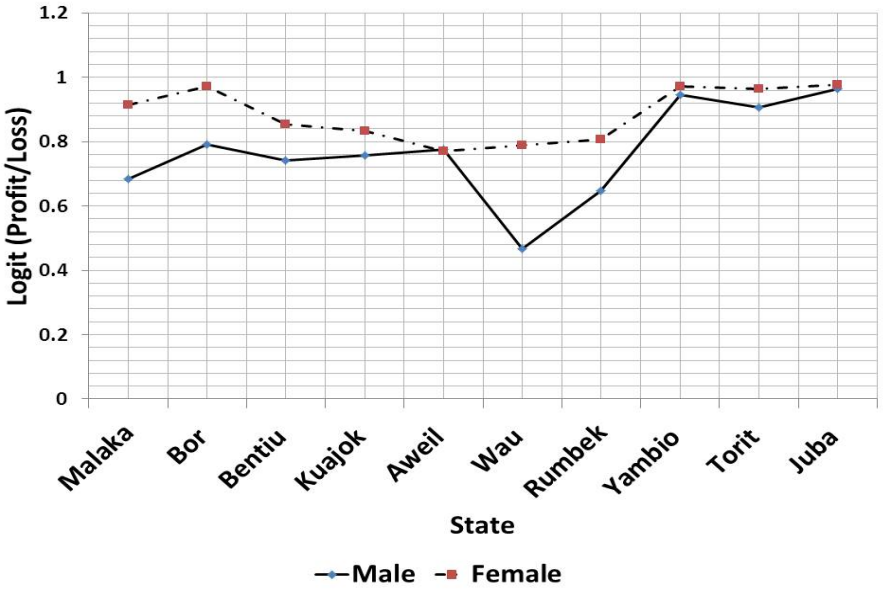

Figure 5.3: Log-odds Associated with Business Location by Gender

Wau, Rumbek, Yambio, and Torit is insignificant. This implies that controlling for the other covariates, the odds of success for male entrepreneurs in these locations is not different from that of female enterpreneurs in Juba. However, from Figure 5.3 it can be observed that the success rate for both Wau male entrepreneurs and Rumbek male entrepreneurs is significantly different from that of Juba female entrepreneurs.

5.3 Comparison of the Logistic Model and the Survey Logistic Model

In this section we present a comparison of the estimates and the standard errors under the logistic regression model, with those under the survey logistic regression model, produced by both PROC LOGISTIC and PROC SURVEYLOGISTIC respectively. In our analysis we found that sampling stratification and the sampling weight have an impact on the estimates as well as on their standard errors. As presented in Table 5.5 and Table 5.6, we found that the standard errors in the survey logistic model are higher than those in the ordinary logistic model. The unequal distribution of the sample size across states for the business survey (Table 3.1) has an impact on the parameter estimates and on the relative standard

errors. The problem of unequal sample size and its impact on parameter estimates has been investigated by many researchers. For instance, Jackson (2001) found that sample size has a practical, significant effect on the variance of the parameter estimates. Jackson (2001) argued that small sample estimates are practically unbiased. Cramer (1999) on the other hand, reported that, a binary logit analysis with unequal sample frequencies of two outcomes, suggested that the less frequent outcome always has lower estimated predicted probabilities than the other outcome especially for survey analyses in social sciences and epidemiology. Korn and Graubard (1995) also reported differences in the estimates of unweighted and weighted estimators. According to Korn and Graubard (1995), there is a trade-off cut between the potential large bias of the unweighted estimators, and the potential larger variability of the weighted estimators, due to stratification of the sampling design. Generally, if the effect of the survey design is not considered in the analysis, the resultant estimates will not be reliable.

Although the survey logistic regression model can be used to account for complexity in the survey design for the data under study, it still faces a number of limitations. Among these are:

1. Model fit diagnostics such as the deviance residual, the Pearson residual, and the Hosmer-Lemeshow goodness of fit statistics, are not yet developed in PROC SURVEYLOGISTIC. As such, further model diagnostics can not be done to check the consistency of the model fit.

2. As discussed in the literature, the AIC and the SC are used as approximations for model goodness-of-fit tests. However, these statistics are best used for variable selections for models with too many parameters. Both the AIC and the SC involve -2logL to penalize models with too many parameters (Agresti, 2002). The use of the AIC and the SC as approximations for model fit is not enough for model-check diagnostics.

3. Unlike with PROC LOGISTIC, where further diagnostics such as the outlier, leverage, and influencial tests can be done, PROC SURVEYLOGISTIC has not yet implemented commands for these diagnostics. As such, test for influential observations (Cook's distance plot) can not be done using PROC SURVEYLOGISTIC.

Table 5.5: Model Comparison for the Logistic and the Survey Logistic Main Effect

	Logistic Model		Survey Logistic Model	
Effects	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Intercept	2.4503	0.8058	3.7028	1.1445
State(Ref=Juba)				
Malakal	-1.6793	0.2462	-1.3191	0.4349
Bor	0.4104	0.823	-0.1775	0.4246
Bentiu	-2.6003	0.5112	-1.9339	0.3477
Kuajok	-2.69	0.5012	-2.0873	0.3857
Aweil	-3.747	0.5779	-2.485	0.3954
Wau	-2.4525	0.4306	-2.3891	0.3976
Rumbek	-1.9334	0.4268	-2.2775	0.4072
Yambio	-0.7587	0.5002	-0.1709	0.5398
Torit	-0.0126	0.569	-0.3768	0.4216
Ownership(Ref=Sole Proprietorship)				
Partnership	0.0799	0.1668	0.0425	0.2669
Company	0.087	0.2546	0.2999	0.3717
Others	0.3561	0.304	1.0647	0.5011
Internet(Ref=No)				
Yes	-0.833	0.2783	-2.1457	0.9841
Government Support(Ref=No)				
Yes	-0.016	0.2964	-0.8362	0.4695
Cash flow problem (Ref=No)				
Yes	0.7156	0.2223	0.1619	0.2075
Financial Loss due to shock(Ref=No)				
Yes	-0.2391	0.3387	0.3669	0.2141
Business Size(Ref=Small Enterprise)				
Micro enterprise	0.1789	0.2433	-0.1468	0.4454
Business Age	0.0382	0.1804	0.1429	0.2627
Outstanding Loan(Ref=No)				
Yes	-0.2655	0.1385	2.2952	1.0091
Startup Capital	0.0221	0.0269	-1.3928	0.7082
Gender(Ref=Female)				
Male	-0.3146	0.1323	-0.3796	0.2427
Education Level(Ref=V. training)				
No schooling	0.3271	0.3486	0.3418	0.5574
Primary school	0.0431	0.8000	0.0286	0.547
Secondary school	0.0049	0.3368	0.2177	0.4917
University Degree	0.3627	0.3794	0.3409	0.646
Stakeholders(Ref=Foreign)				
Local	-0.6873	0.498	-1.2244	0.657
Type of Industry(Ref=Other Services)				
MEMC	1.1429	0.4112	1.6316	0.5552
TT	0.5565	0.278	0.7055	0.3622
APSA	0.717	0.3921	0.5943	0.6414
ESHS	0.674	0.3368	0.8176	0.4634

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

Table 5.6: Continuation of the Comparison of Logistic Model and Survey Logistic Model

	Logistic Model		Survey Logistic Model	
Effects	**Estimate**	**Standard Error**	**Estimate**	**Standard Error**
Internet*Stakeholders(Ref=No and Foreign)				
Yes and Local	1.717	0.7136	1.985	1.0001
Loan*Education(Ref=No and V.training)				
Yes and No schooling	-1.9562	0.8827	-3.0383	1.0859
Yes and Primary school	-1.4217	0.8973	-2.2612	1.1084
Yes and Secondary school	-1.1603	0.8792	-2.6349	1.1096
Yes and university degree	-1.8401	0.9303	-2.8844	1.257
State*Gender(Ref=Juba and Female)				
Malakal and Male	-1.0253	0.4920	-1.2282	0.7109
Bor and Male	-1.8047	0.8480	-1.8171	1.0786
Bentiu and Male	0.0606	0.5693	-0.3345	0.8040
Kuajok and Male	-0.8884	0.5360	-0.1026	0.6755
Aweil and Male	0.6194	0.6242	0.4082	0.8249
Wau and Male	-0.4364	0.4836	-1.0681	0.6094
Rumbek and Male	-0.1142	0.4974	-0.4420	0.6958
Yambio and Male	0.3281	0.6554	-0.2950	1.0914
Torit and Male	-0.4995	0.6222	-0.6891	0.8074

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

Chapter 6

Generalized Linear Mixed Models

In this chapter we introduce the generalized linear mixed models (GLMM). Generally, as discussed by Leeuw *et al* (2008), many surveys' data are not from simple random samples or a relatively homogeneous population, but are obtained from nested sampling in heterogeneous subgroups, as is the case with the business survey data. As discussed in Section 4.1, the generalized linear models are appropriated for data from a simple random sampling. When the data is from a complex survey, application of the generalized linear models becomes difficult. The data used in this study, as discussed in Section 3.1, is obtained from a business survey which was designed using stratified random sampling. Therefore, in order to capture the complexity effects for the business survey data, we introduce the generalized linear mixed models which account for both the fixed and random effects of the data.

The linear mixed models have been used in situations where the observations (response variables) are continuous (Jiang, 2007). However, in practice there are cases where the response variable is discrete or categorical, as discussed in Section 4.1, and which follows a general distribution of the exponential family which can be solved using generalized linear models. The generalized linear models as discussed in Section 4.1 include a variety of models that include normal, binomial, Poisson, and multinomial distributions, to mention but a few (Agresti, 2002). When data is obtained from a complex survey design, generalized linear mixed models are preferred to generalized linear models. This is because the GLMM account for the complexity of the survey design. Generally, generalized linear mixed models have the same features as the generalized linear models. The only difference is the inclusion of the

random effect to the model. Suppose that a vector of random effect u, and the response $y_1, y_2, \ldots, y_n$, are (conditionally) independent, such that the conditional distribution of y_i, given u, is a member of the exponential family with probability density function (pdf)

$$f\left(y_i|u\right) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + \ c_i(y_i, \phi)\right], i = 1, 2, \ldots, n \tag{6.1}$$

Where $b(.)$, $a_i(.)$, $c_i(.,.)$ are known functions, and ϕ is a dispersion parameter which may or may not be known. Then the quantity θ_i is associated with the conditional mean μ_i=E$(y_i|u)$, which is in turn associated with a linear predictor

$$\eta_i = \mathbf{x}'_i\beta + \mathbf{z}'_i u \tag{6.2}$$

Where $\mathbf{x}_i$ and $\mathbf{z}_i$ are known vectors for the design matrix for the fixed and random effects respectively, and β is a vector of unknown parameters for the fixed effects, through a link function $g(.)$, such that

$$g(\mu_i) = \eta_i$$

It is assumed that $y_i \sim N(x'_i\beta + z'_i u, \delta^2)$, and $u \sim N(0, G)$, where δ^2 is unknown variance and the covariance matrix G may depend on a vector of unknown variance components. The general argument is that the linear predictor for the generalized linear model is given by

$$\eta_i = \mathbf{x}'_i\beta$$

Hence, when the random effects are included in the model, then the linear predictor will have a form of a generalized linear mixed model given by

$$\eta_i = \mathbf{x}'_i\beta + \mathbf{z}'_i\mathbf{u} \tag{6.3}$$

Where η_i=$g(\mu_i)$, g is a link function, and μ=E $(y|u)$. This gives the general form of the generalized linear mixed model

$$g(\mu_i) = \mathbf{x}'_i\beta + \mathbf{z}'_i\mathbf{u} \tag{6.4}$$

Where:

- $g(.)$ is the link function which links together the conditional mean of y_i and the linear form of predictors;
- $\mathbf{x'}_i$ is the i^{th} row of the model design matrix for the fixed effects;
- β is the vector of the fixed effect parameter;
- $\mathbf{z'}_i$ is the i^{th} row of model design matrix for the random effect; and
- $\mathbf{u}$ is the vector of random effects.

Refer to McCulloch *et al.* (2001) for further discussion.

The mixed logistic model is given by

$$\text{logit}(p_i) = \mathbf{x'}_i\beta + \mathbf{z'}_i\mathbf{u} \quad i = 1, 2, \ldots, n \tag{6.5}$$

According to Jiang (2007), this is a special case of the GLMM in which the (conditional) exponential family is bernoulli and the link function is $g(\mu) = \text{logit}(\mu)$. This is the case where the dispersion parameter $\phi = 1$.

Recall that the objective of the study is to model determinants of business success and failure, where the response variable y is binary (0=loss, 1=profit), also refer to Section 3.2. Thus the distribution of the response variable y, belongs to the exponential family of distribution. Note that if the model has only fixed effects, generalized linear models are applicable, but if the model has both fixed and random effects, then generalized linear mixed models are applicable. Therefore since the effects of PSUs were randomly selected, as discussed in Section 3.1, they enter the model as random effects, and the same variables selected in Section 3.2 as explanatory variables, are now the fixed effects. Since we have only one random effect (i.e PSU), a simplest form of the GLMM, which is referred to as the random intercept model, is used to fit the business survey data. This model is given by

$$\eta = \mathbf{X}\beta + \mathbf{u}, \quad \mathbf{u} \sim N(0, \delta^2\mathbf{I}) \tag{6.6}$$

Where X and β are as defined in 6.4, and $\mathbf{u}$ is a random vector of PSU effects whose i^{th} element represents the influence of the i^{th} PSU on business observations not captured by the

observed covariates. Refer to Pendergast *et al* (1996), Agresti *et al* (2000), and Littell *et al* (2006) for more details on the random intercept model.

When data is from a complex survey design, parameter estimation for the GLMM becomes difficult. Generally, unlike the linear mixed models, the likelihood function under the GLMM does not have a closed-form expression (Jiang, 2007). The argument is that the likelihood may involve high-dimensional integrals that can not be evaluated analytically. Littell *et al* (2006) on the other hand, report that obtaining the marginal distribution is not easy if the conditional distribution of response y, given the random effect u, is not normal. Generally, when the exact likelihood function is difficult to compute when data is non-normal, approximation becomes one of the natural alternatives (Jiang, 2007). Different statisticians have proposed different methods for approximation. For instance, Wolfinger and O'Connell (1993) proposed the pseudo-likelihood, also sometimes referred to as the restricted pseudo-likelihood (RPL), and Breslow and Clayton (1993) proposed the penalized quasi-likelihood (PQL). These approximation methods can be used to determine the maximum likelihood. As discussed by Jiang (2007), for simple models the likelihood function may be evaluated by numerical integration techniques. Such techniques are trackable if the integrals involved are of low-dimension. McClloch *et al* (2001) reported that the quasi-likelihood does not specify a distribution, only the mean-to-variance relationship; thus it is not a sufficient base on which one can estimate the variance - covariance structure. Hence a penalty function is added to the quasi-likelihood, of the form

$$\frac{1}{2}u'D^{-1}u$$

Such that the penalty quasi-likelihood is given as

$$PQL = \sum Q_i - \frac{1}{2}u'D^{-1}u \tag{6.7}$$

The maximum quasi-likelihood equations would then be obtained by differentiating equation (6.7) with respect to β and u. This gives

$$\frac{1}{\delta^2}X'WD(y-\mu) = 0$$

and

$$\frac{1}{\delta^2} Z'WD(y-\mu) - D^{-1}u = 0$$

Justification of this approach is via laplace approximation (integral approximation method). As reported by McCulloch *et al* (2001), despite the number of ways in which the same approaches have been justified, the PQL approach has not been found to work well in practice. Also refer to Jiang (2007), Lee *et al* (2007), and Breslow and Clayton (1993) for further discussion. Thus, given the inconsistence of the penalized quasi-likelihood, as discussed earlier, we use the pseudo-likelihood which is also a default approximation method in SAS (9.2), to estimate the best linear unbiased predictors (BLUP) of the random effects. The pseudo-likelihood is based on the linearization approach, which is best for a model with only one random effect (Wolfinger *et al*. 1993). Generally, the estimating equations for the GLMM are solved iteratively to obtain the parameter estimates. For the binary response, the estimating equations are given as

$$\begin{bmatrix} \mathbf{X'WX} & \mathbf{X'WZ} \\ \mathbf{Z'WX} & \mathbf{Z'WZ} + \mathbf{G^{-1}} \end{bmatrix} \begin{bmatrix} \beta \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X'Wy^*} \\ \mathbf{Z'Wy^*} \end{bmatrix} \tag{6.8}$$

Where:

- $\mathbf{y^*} = \hat{\eta} + (\mathbf{y} - \hat{\pi})\mathbf{D^{-1}}$ is the working (pseudo) dependent variate;
- $\mathbf{X}$ and $\mathbf{Z}$ are as defined in (6.4);
- $\hat{\eta} = \mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\mathbf{u}}$;
- $\mathbf{W} = \mathbf{D}'\mathbf{R}^{-1}\mathbf{D}$;
- $\mathbf{D} = \frac{\partial \mu}{\partial \eta} = \text{diag}[\pi_i(1-\pi_i)]$;
- $\mathbf{R} = \text{var}(\mathbf{y}|\mathbf{u}) = \mathbf{R}_\mu^{\frac{1}{2}}\mathbf{A}\mathbf{R}_\mu^{\frac{1}{2}} = \mathbf{R}_\mu^{\frac{1}{2}}\mathbf{I}\mathbf{R}_\mu^{\frac{1}{2}}$;
- $\mathbf{R}_\mu = \text{diag}[\pi_i(1-\pi_i)$;
- $\mathbf{A} = \mathbf{I} =$ identity matrix; and

- $\mathbf{G}$=var($\mathbf{u}$)=$\mathbf{I}\delta_u^2$.

Here the features of the ungrouped binary conditional model are observed:

1. The conditional mean: $\mu_i = \pi_i = \frac{1}{1+\exp(-\eta)}$;
2. The natural parameter: $\theta(\mu_i) = -\log(\pi_i^{-1} - 1)$;
3. The variance function: $\mathrm{V}(\mu_i) = \pi_i(1 - \pi_i)$; and
4. The dispersion parameter: a(ϕ)=1.

Therefore, from (6.8) we can obtain the following:

1. The profiled parameter (Fixed effects) estimates given by

$$\hat{\beta} = (\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\theta)^{-1}y^*$$

2. The best linear unbiased predictor (BLUP) of the random vector effect u given by

$$\hat{\mathbf{u}} = \hat{\mathbf{G}}\mathbf{Z}'\mathbf{V}(\theta)^{-1}\hat{\mathbf{r}}$$

Where:

- $\hat{\mathbf{r}} = y^* - (\mathbf{X}'\mathbf{V}(\theta)^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}(\theta)^{-1}y^*$ for $\mathbf{y}^* = \hat{\eta} + (\mathbf{y} - \hat{\pi})\mathbf{D}^{-1}$

- θ is a qx1 vector of parameters containing all unknowns in **G and R**

- $\mathbf{V}(\theta) = \mathbf{ZGZ}' + \mathbf{D}^{-1}\mathbf{R}_\mu^{\frac{1}{2}}\mathbf{A}\mathbf{R}_\mu^{\frac{1}{2}}\mathbf{D}^{-1}$ and

- $\mathbf{D} = \left(\frac{\partial\mu}{\partial\eta}\right)_{\hat{\beta},\hat{u}}$.

The pseudo-response and weights are updated using the parameter and the random effects estimates, and vice versa. This process continues until it converge into a solution.

Recall from Section 4.1 that for binomial and poisson distribution, the scale parameter ϕ is expected to be 1. For the case of the GLMM, the variance is a known function of the

mean, and so $\phi = 1$. However, when the scale parameter ϕ is different from 1 for the GLMM, the parameter estimates can be profiled from the logPL. Generally, the parameter ϕ for the GLMM is estimated by

$$\hat{\phi} = \hat{\mathbf{r}}'\mathbf{V}^{-1}\hat{\mathbf{r}}/m \tag{6.9}$$

Where $m = n$, and n is the number of individuals used in the analysis for the Maximum Pseudo Likelihood (MPL), and $m = n - p$, where p is the rank of X for the RPL.

As was the case with the generalized linear models discussed in Section 4.1, statistical inference (test of hypothesis) about the model parameters is done using the likelihood ratio test, the Wald test, and the score test (McCulloch *et al*. 2001). Littell *et al* (2006) on the other hand, reported that inferences in GLMM can also be done using the F- test statistic, if the conditional variance $\mathbf{R}$ depends on a known or unknown scale parameter matrix $\mathbf{A}$. The likelihood ratio test, the Wald test, and the score test for the GLMM are discussed in detail in McCulloch *et al* (2001).

Unlike the GLM, model selection in the GLMM is difficult. Some researchers have used Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) for generalized linear mixed model selection, using longitudinal data (Pan and Lin, 2005). However, it is often difficult to compare the pseudo values with the AIC statistics, because the AIC is based on maximum likelihood computation for simple random sampling, whereas the pseudo-likelihood is based on complex survey design (Duijn *et al*, 2008). In this study, the same variables used to fit the main effect model using PROC LOGISTIC discussed in Section 4.3, are also used to fit the generalized linear mixed model.

In general terms, as discussed by Cook and Weisberg (1994), residual plots are the best to assess adequacy of model goodness of fit. However, it is often difficult to determine whether the observed pattern reflects model misspecification or random fluctuation. For binary data for instance, the residual plots tend to be uninformative because all the points lie on one of the two curves according to the two possible values of the response (Pan and Lin, 2005). Thus, it becomes difficult to interpret the individual residual plots for the GLMM. Pan and Lin (2005) proposed cumulative sums of residuals with respect to covariate values or predicted values, for longitudinal data for the GLMM. In this study, the log pseudo-likelihood

and the generalized chi-square statistics are used to assess model fit. We also present least - squares means differences to make inferences about the parameters of the current fitted model.

The least-squares means differences (LSMD) is another form of inference in GLMM. The least-squares means are used to estimate the marginal means over a balanced population. Inference using the LSMD in the context of GLMM, with examples for binomial and poisson data, is discussed in detail by Littel *et al.* (2006). Let the least-squares means be denoted by μ, then:

1. $\mu_i - \mu_j$, for all $i \neq j$ represent the pairwise comparison of the factor level least-squares means; and

2. $\mu_i - \hat{\mu}$, where $\hat{\mu}$ is the overall least-squares means, represent the comparison of each factor level least-squares means against the overall average of all factor levels least-squares means.

Generally, factor least-squares means can be displayed using graphs or tables. In this study, the diffogram (mean-mean scatter plot of least-squares means) concept discussed by Littell *et al.* (2006) and Hsu *et al.* (1994) is implemented (Figure 6.1). For the pairwise comparisons, we used the Tukey-Kramer method for adjustment for multiplicity, whereas for the comparison of each factor level-squares means against the average factor least-squares mean, we used the Nelson method of adjustment for multiplicity (Hsu *et al.* 1994) and (Littell *et al.* 2006).

As discussed earlier, inference for the GLMM parameter can be done using the least-squares means differences of the response measured at different factor levels. As displayed in Figure (6.1), the axes of the diffogram plot are the least-squares means, where the $y-$axis $= \hat{\mu}$, and the x-axis $= \hat{\mu}$. The 45^0 line is the reference line from the origin corresponding to the set of points satisfying $\hat{\mu_i} = \hat{\mu_j}$, for all i and j. Thus, as discussed by Hsu *et al.* (1994), the directional distance of any point from the 45^0 line is given by the difference of the two corresponding least-squares means divided by the square root of 2. For instance, the directional distance of the point $(\hat{\mu_i}, \hat{\mu_j})$ from the 45^0 line is given by $\hat{\mu_i}, \hat{\mu_j}/\sqrt{2}$. For the Tukey-Kramer's confidence interval, the interpretation is the same as discussed by Hsu *et al.* (1994) and Littel *et al.* (2006). That is to say, the Tukey-Kramer's confidence interval

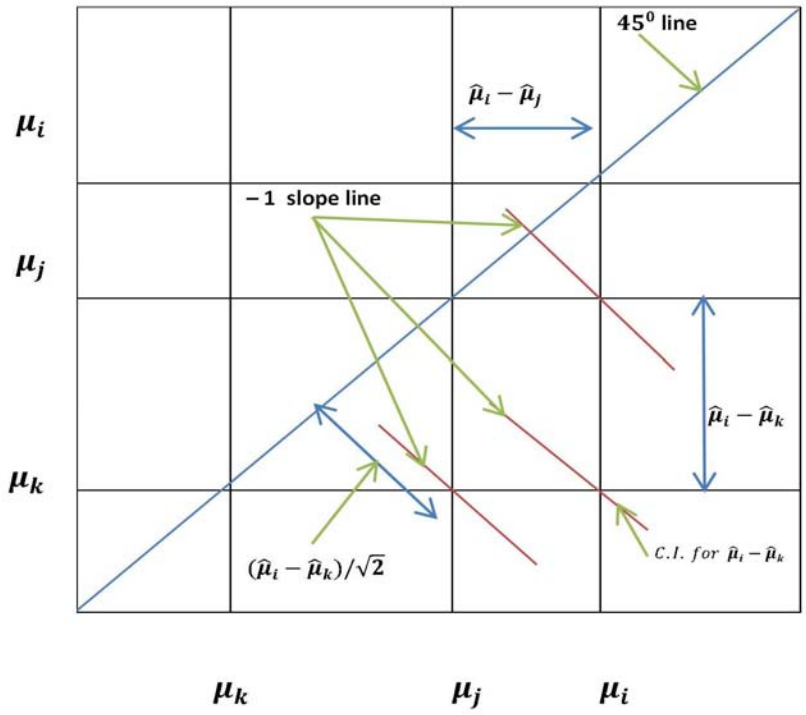

Figure 6.1: Diffogram (Mean-Mean Scatter Plot)

for the difference between two least-squares means $(\hat{\mu_i}, \hat{\mu_j})$ is represented by the length of '-1 slope' line centered at the intersection of $(\hat{\mu_i}, \hat{\mu_j})$, the same for $(\hat{\mu_i}, \hat{\mu_k})$. Hence, if the difference between the two least-squares means is significant, the corresponding line will not cross the 45^0 line, and vice versa. Generally, the Tukey-Kramer's method adjusts the '-1 slope' line for rotation and multiplicity; as such, all estimates are on the link scale Littell *et al.* (2006).

For the least - squares means for each factor level compared against the average of all levels (Analysis of Means), the graphical display is different from that of the diffogram. For the 'Analysis of Means', the x-axis represents factor levels and the y-axis represents the least-squares means on the link scale. Hence, the average of the least-squares means is given by the horizontal line in the center of the graph, and the magnitude of the difference of the factor level least-squares means from the average least-squares means is represented by the vertical lines from the horizontal lines. The lower decision limit (LDL) and the upper decision limit (UDL) are represented by the dashed horizontal steps plots on both sides of

the horizontal line. The decision here is that if the least-squares mean of the i^{th} level is significantly different from the average, the corresponding vertical line crosses one of the decision limits, and vice versa. Both the pair comparison of factor level least-squares means and analysis of means is discussed in detail in SAS PROC GLIMMIX User's Guide (2008), and in SAS for mixed models by Littell *et al* (2006). Generally, SAS PROC GLIMMIX procedure is used in the study to calculate parameter estimates and other diagnostics for the generalized linear mixed model.

6.1 Results

We present the solution for the fixed effects in Table 6.3, the solution for the interaction of the fixed effects in Table 6.4, and the covariance parameter estimates are presented in Table 6.2. Note that in this study, the variance component is chosen for the covariate structure because we are interested in estimating the variance among the random effect (PSU).

We used the -2 reg log pseudo-likelihood and the generalized chi-square over it's degree of freedom, to assess the model fit and consistency check of the model goodness of fit. The likelihood is proceeded by the word "pseudo", to indicate that it is computed from a pseudo-likelihood and not from the true likelihood used for data from simple random sampling design. The covariance parameter matrix as well as the asymptotic covariance parameter estimates of the covariance parameter estimates was also used to assess the model adequacy. We also used the least-squares means and the analysis of means to make inferences about the model parameter estimates. That is to say, we used the least-squares means and the analysis of means to assess whether there is a significant or insignificant difference between the effects in the model.

When the Scatterthwaite-based (degree of freedom) method was used, the standard errors were found to be the same as those in Table 6.3 and Table 6.4 for the containment method degree of freedom. Additionally, when the model was adjusted for uncertainty in estimating G and R, the same standard errors were not different from those presented in Table 6.3 and Table 6.4 respectively. Since the model fitted is a random intercept model with an overall intercept estimate of 2.8836 (Table 6.3), which is adjusted fairly by a small standard error

estimate of 0.1865 (Table 6.4), this implies that the model fitted the data well.

Table 6.1: Model Fit Statistics

Criterion	Statistics
-2 Res Log Pseudo-Likelihood	9648.33
Generalized Chi-Square	1873.79
Generalized Chi-Square / DF	0.97

Table 6.1 presents the assessment of the model fit using the log pseudo-likelihood and the generalized chi-square test. The minus twice the residual log pseudo-likelihood of the model fit is 9648.33, whereas the generalized chi-square is 1873.79. The ratio of the generalized chi-square statistics divided by it's degree of freedom is given by $\frac{1873.79}{1922} = 0.97$. This ratio measures the residual variability in the margin distribution of the data. Since $\phi = 1$; the ratio 0.97 is close to 1, this indicates that the variability in the data has been properly modelled and hence there was no residual over-dispersion. In other words, it indicates that there is no lack of fit when the random effect was included in the model (Schabenberger, 2005).

6.2 Covariance Parameter Estimates

We present the covariance parameter estimates in Table 6.2. From Table 6.2, the variance of the random PSU effect on the logit scale is estimated as $\hat{\sigma_u}^2 = 0.1293$, and standard error $\hat{\sigma_u} = 0.1865$. The same variance estimate is obtained when the PSU is nested within location, by industry and by number of employees. As discussed earlier, the estimates of the average logit between the models effects and their predictions on the logit scale of the data can be interpreted using the least-squares means, and the analysis of means which are presented in Section 6.3.

6.3 Interpretation of Results

In this section we present the interpretation of results based on the least-squares means and the analysis of means which are presented in graphical form. The coefficients for fixed effects are interpreted in the same manner as the logistic regression model discussed in Section

Table 6.2: Covariance Parameter Estimates

Covariance Parameter Estimates			
Cov Parm	Subject	Estimate	Std Error
Intercept	PSU	0.1293	0.1865
Intercept	PSU (Location*Industry/No.Employees)	0.1293	0.1865
Asymptotic Covariance Matrix of Covariance Parameter Estimates			
Cov Parm	Subject	CovP1	
Intercept	PSU	0.03477	

4.2. The summary of all pairwise comparisons of the least-squares means, and the analysis of means, are given in Figures 6.2 to 6.9. The diffogram, as discussed in the Section 6.2, displays a line for each comparison and the axes of the plots represent the scale of the least-squares means. The confidence limit for the least-squares means difference is presented by the length of the line, which is adjusted for the rotation as well as possible multiplicity. The 45^0 line is referred to as the reference line of the plot. Thus, the lines cross the 45^0 line if two least-squares means are insignificantly different.

For the analysis of means, the dashed horizontal step plots in the graph represent the upper and lower decision limits, as determined at the 95% decision limit. Hence, if the level is significantly different from the average, then the corresponding vertical line crosses the decision limit. Discussions on the least-squares means is presented in detail in the GLIMMIX Procedure of the SAS/STAT User's Guide (2008), and in Littell *et al*. (2006). Note that all the constrasts are on the logit scale, as discussed in Section 6.3, and are given as $\text{logit}\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right)$.

Figure 6.2 displays the comparison of least-squares means for business location (state) adjusted for multiplicity. It can be noted from Figure 6.2, that the lines centered at the interactions of business locations: 5-Aweil and 7- Rumbek; 5-Aweil and 1-Malakal; 5-Aweil and 9-Torit; 5-Aweil and 8-Yambio; 5-Aweil and 2-Bor; 5-Aweil and 10-Juba; 6-Wau and 7-Rumbek; 6-Wau and 1-Malakal; 6-Wau and 9-Torit; 6-Wau and 8-Yambio; 6-Wau and 2-Bor;

Table 6.3: Solutions for Fixed Affects for GLMM

Effect	**Estimate**	**Standard Error**	**t-Value**	**Pr** $>\|t\|$
Intercept	2.8836	0.8224	3.51	0.0005
State(Ref=Juba)				
Malakal	-1.0899	0.3943	-2.76	0.0058
Bor	0.7942	0.7739	1.03	0.3049
Bentiu	-2.4052	0.5019	-4.79	$<$ .0001
Kuajok	-1.9435	0.453	-4.29	$<$ .0001
Aweil	-3.5608	0.5591	-6.37	$<$ .0001
Wau	-2.4285	0.4017	-6.05	$<$ .0001
Rumbek	-2.0727	0.3801	-5.45	$<$ .0001
Yambio	-0.4386	0.447	-0.98	0.3266
Torit	-0.216	0.4813	-0.45	0.6536
Ownership(Ref=Sole Proprietorship)				
Partnership	0.08973	0.1684	0.53	0.5942
Company	0.05376	0.2547	0.21	0.8329
Others	0.3439	0.3076	1.12	0.2638
Internet(Ref=No)				
Yes	-1.9276	0.7236	-2.66	0.0078
Government Support(Ref=No)				
Yes	0.00598	0.3006	0.02	0.9841
Cash flow Problem(Ref=No)				
Yes	-0.1285	0.1255	-1.02	0.3059
Financial Loss due to shock(Ref=No)				
Yes	0.1274	0.1338	0.95	0.3413
Business Size(Ref=Small Enterprise)				
Micro enterprise	0.1916	0.2459	0.78	0.4361
Business Age	-0.01559	0.1831	-0.09	0.9322
Outstanding Loan(Ref=No)				
Yes	1.3897	0.8764	1.59	0.113
Startup Capital	-1.0451	0.3737	-2.8	0.0052
Gender(Ref=Female)				
Male	0.1116	0.3568	0.31	0.7544
Education Level(Ref=V. training)				
No schooling	0.2387	0.353	0.68	0.499
Primary school	-0.00678	0.3666	-0.02	0.9853
Secondary school	-0.0547	0.3415	-0.16	0.8728
University Degree	0.3293	0.3846	0.86	0.3919
Stakeholders(Ref=Foreign)				
Local	-0.6422	0.4962	-1.29	0.1957
Type of Industry(Ref=Other Services)				
MEMC	1.2511	0.4151	3.01	0.0026
TT	0.632	0.278	2.27	0.0232
APSA	0.7601	0.395	1.92	0.0545
ESHS	0.7335	0.3386	2.17	0.0304

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

Table 6.4: Solutions for Interaction of the Fixed Affects for GLMM

Effect	**Estimate**	**Standard Error**	**t-Value**	**Pr** $>\|t\|$
Internet*Stakeholders(Ref=No and Foreign)				
Yes and Local	1.7294	0.7376	2.34	0.0191
Loan*Education(Ref=No and V.training)				
Yes and No schooling	-2.1093	0.9108	-2.32	0.0207
Yes and Primary school	-1.5382	0.9235	-1.67	0.0959
Yes and Secondary school	-1.2555	0.9044	-1.39	0.1652
Yes and university degree	-2.0406	0.9594	-2.13	0.0335
State*Gender(Ref=Juba and Female)				
Malakal and Male	-1.1469	0.5011	-2.29	0.0222
Bor and Male	-1.8838	0.8548	-2.2	0.0277
Bentiu and Male	-0.03837	0.5851	-0.07	0.9477
Kuajok and Male	-0.9497	0.5439	-1.75	0.081
Aweil and Male	0.5956	0.6349	0.94	0.3483
Wau and Male	-0.5097	0.498	-1.02	0.3062
Rumbek and Male	-0.09341	0.5053	-0.18	0.8534
Yambio and Male	0.24	0.6626	0.36	0.7172
Torit and Male	-0.5653	0.6312	-0.9	0.3706

V= Vocational, MEMC=Mining, Energy, Manufacturing and Construction, TT=Trade and Transportation, APSA=Administration, Professional and Scientific Activity, ESHS=Education, Social, and Health services.

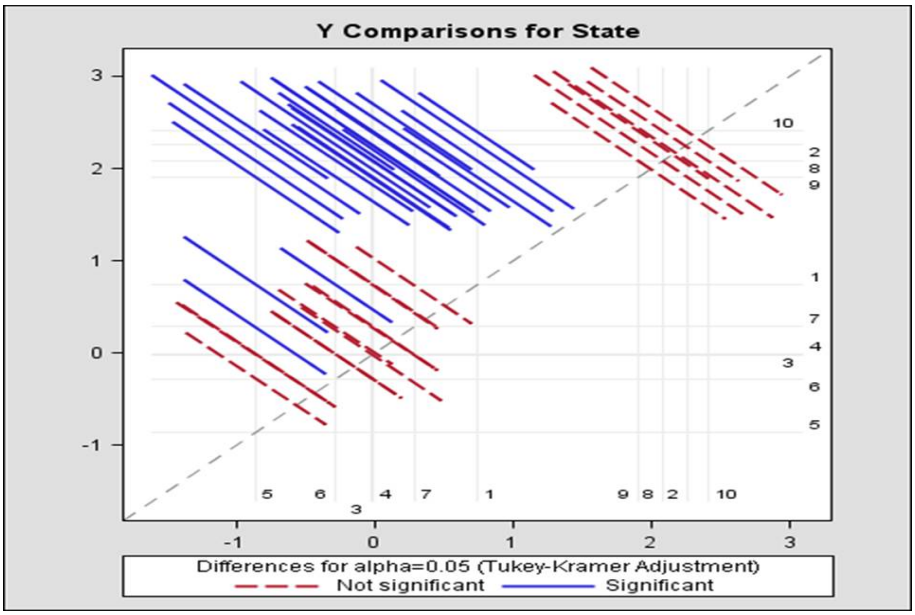

Figure 6.2: Diffogram for State Interaction effect

6-Wau and 10-Juba are significantly different. This implies that business success in Aweil is significantly different from those in Juba, Bor, Yambio, Torit, Malakal, and Rumbek. Note that, there are no statistical significant differences for the other locations not mentioned.

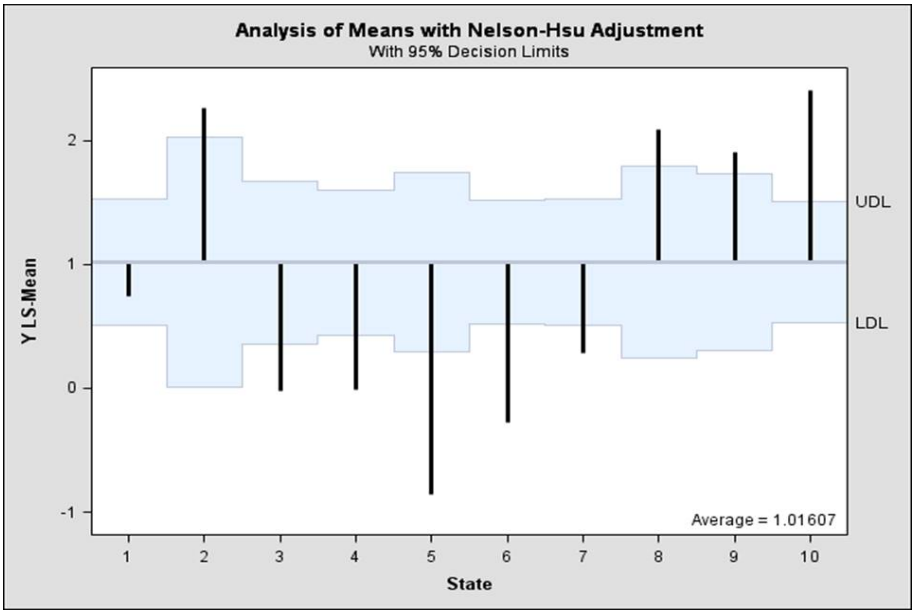

Figure 6.3: Analysis of Means for State Interaction effect

Figure 6.3 displays the analysis of means for business location (state) interaction effect on business success. The average business location effect on the logit scale is 1.01607 (Figure 6.3). Hence, it can be noted that the vertical lines corresponding to business locations: Bor; Bentiu; Kuajok; Aweil; Wau; Rumbek; Yambio; Torit; and Juba crossed the 95% decision limit. This implies that business success in these locations are significantly different from the average. The averages for Juba and Aweil are the most extreme on the upper and lower decision limits respectively. That is to say, the least-squares means for Juba are greater than the average, whereas the least-squares means for Aweil are less than the average.

As displayed in Figure 6.4, the pair comparison of least-squares means of the levels of business location by gender interaction effect on business success, are represented by the lines centered at the interactions: 52-Aweil female entrepreneurs and 71-Rumbek male entrepreneurs; 52-Aweil female entrepreneurs and 11-Malakal male entrepreneurs; 52-Aweil female entrepreneurs and 42- Kuajok female entrepreneurs; 52-Aweil female entrepreneurs and 12-Malakal female entrepreneurs; 52-Aweil female entrepreneurs and 21- Bor male en-

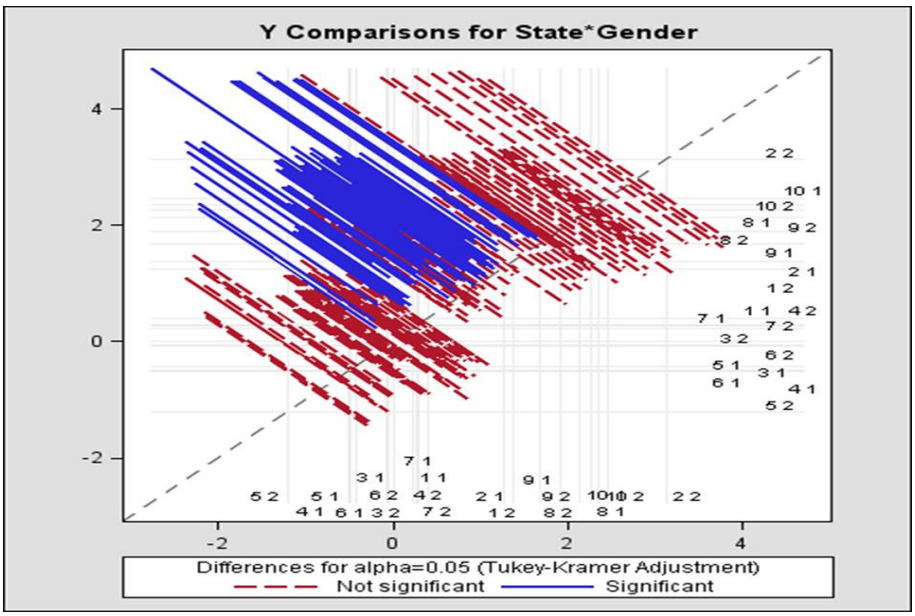

Figure 6.4: Diffogram for State by Gender Interaction Effect

trepreneurs; 52-Aweil female entrepreneurs and 91-Torit male entrepreneurs; 52-Aweil female entrepreneurs and 82-Yambio female entrepreneurs; 52-Aweil female entrepreneurs and 92- Torit female entrepreneurs; 52-Aweil female entrepreneurs and 81-Yambio male entrepreneurs; 52-Aweil female entrepreneurs and 10 2-Juba female entrepreneurs; 52-Aweil female entrepreneurs and 10 1-Juba male entrepreneurs; 52-Aweil female entrepreneurs and 22-Bor female entrepreneurs. These interactions have significantly different least-squares means. The same interpretation applies to 41- Kuajok male entrepreneurs, 51-Aweil male entrepreneurs, and 61-Wau male entrepreneurs, interacting with the above mentioned locations by gender, which also have significantly different least-squares means. Business success is not significantly different for 52-Aweil female entrepreneurs and 51-Aweil male entrepreneurs, 52-Aweil female entrepreneurs and 31-Bentiu male entrepreneurs, and 52-Aweil female entrepreneurs and 61-Wau male entrepreneurs, as explained by least-squares means differences given by the lines of their interactions (Figure 6.4).

The analysis of means for business location by gender interaction effect is given in Figure 6.5. The average business location by gender interaction effect on logit scale is 0.98687 (Figure 6.5). The differences of means for: 11-Malakal male entrepreneurs; 22-Bor female entrepreneurs; 31-Bentiu male entrepreneurs; 41-Kuajok male entrepreneurs; 51- Aweil male

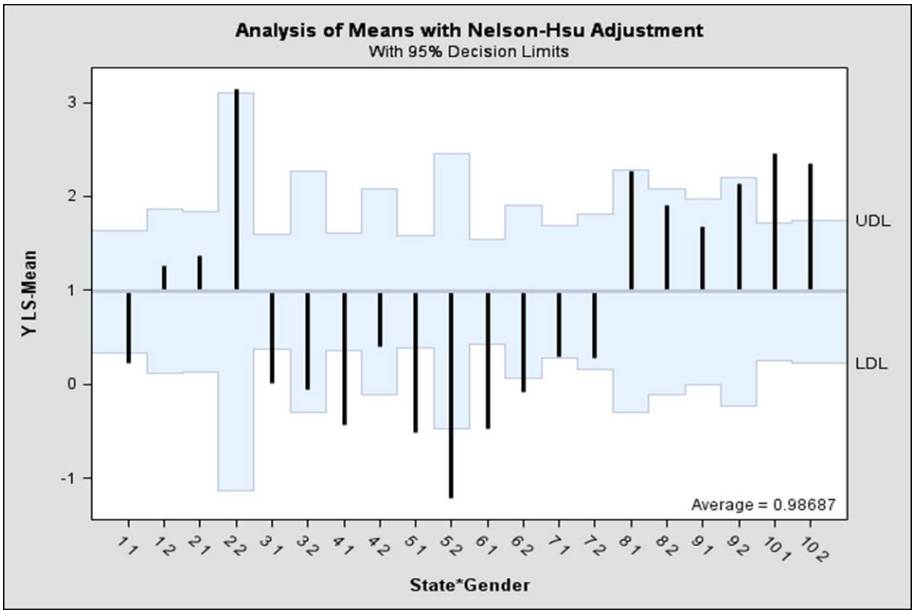

Figure 6.5: Analysis of Means for State by Gender Interaction Effect

entrepreneurs; 52-Aweil female entrepreneurs; 61-Wau male entrepreneurs; 62-Wau female entrepreneurs; 10 1-Juba male entrepreneurs; 10 1-Juba male entrepreneurs; and 10 2-Juba female entrepreneurs are significantly different from the average as displayed by the vertical lines that cross the 95% decision limits in Figure 6.5. It can be noted that 22-Bor female entrepreneurs and 52-Aweil female entrepreneurs have the most extreme averages, on the upper and lower decision limits respectively. For Bor female entrepreneurs, the least-squares means is greater than the average, whereas for Aweil female entrepreneurs, the least-squares means is less than the average.

Figure 6.6 displays the pair comparison of least-squares means of the level of business use of internet by stakeholders, on business success. The pair comparisons for business use of internet, and stakeholder interactions on business success are represented by the lines centered at the interactions of: 12-businesses that have used the internet for business communication who also have foreign stakeholders, and 11-businesses that have used the internet and who have local stakeholders; 12-businesses that have used the internet for business communication who also have foreign stakeholders, and 21- businesses that have not used the internet for business communication and who have local stakeholders; 12-businesses that have used the internet for business communication who also have foreign stakeholders,

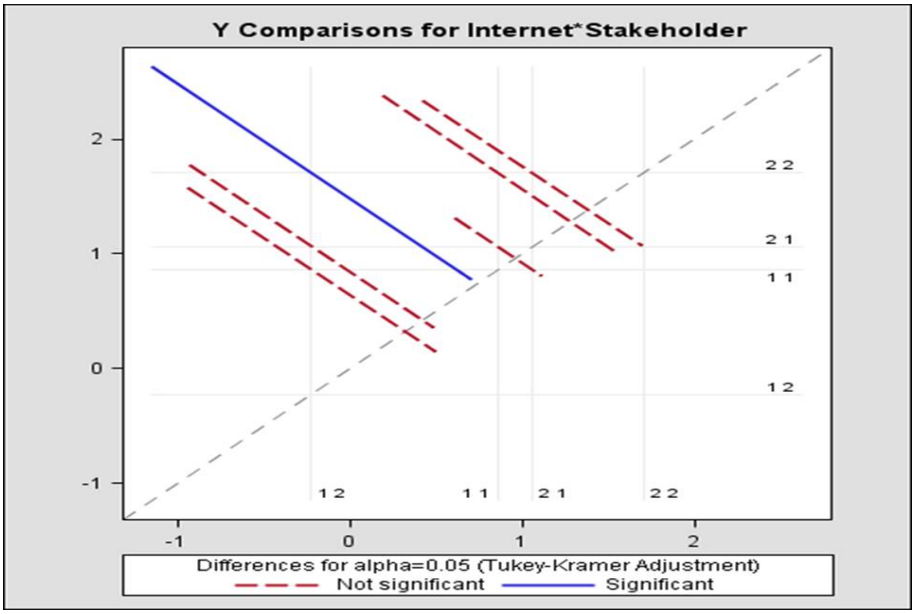

Figure 6.6: Diffogram for Internet and Stakeholders Interaction Effect

and 22- businesses, that have not used the internet and who have foreign stakeholders. These categories have significantly different least-squares means. That is to say, the least squares means of businesses that have used the internet and who have foreign stakeholders, is significantly different from businesses that have used the internet and who also have local stakeholders, businesses that did not use the internet and who have local stakeholders, and businesses that did not use internet and who have foreign stakeholders.

Figure 6.7 displays the analysis of means for business use of internet and stakeholders interaction effects on business success. The average on the logit scale is 0.9589. Figure 6.7 indicates no significantly different pairwise comparison of least-squares means for internet and stakeholders interaction effect. It can be noted from Figure 6.7 that there is no decision limit for the comparison. However, when using unadjusted values for inference, it can be noticed that difference levels of 12-Businesses that have used the internet and who also have foreign stakeholders, and 22-Businesses that did not used the internet and who have foreign stakeholders are significantly different from the average. The least-square means for businesses that have used the internet and who also have foreign stakeholders is below the average, whereas, businesses that did not use the internet and who have foreign stakeholders is above the average.

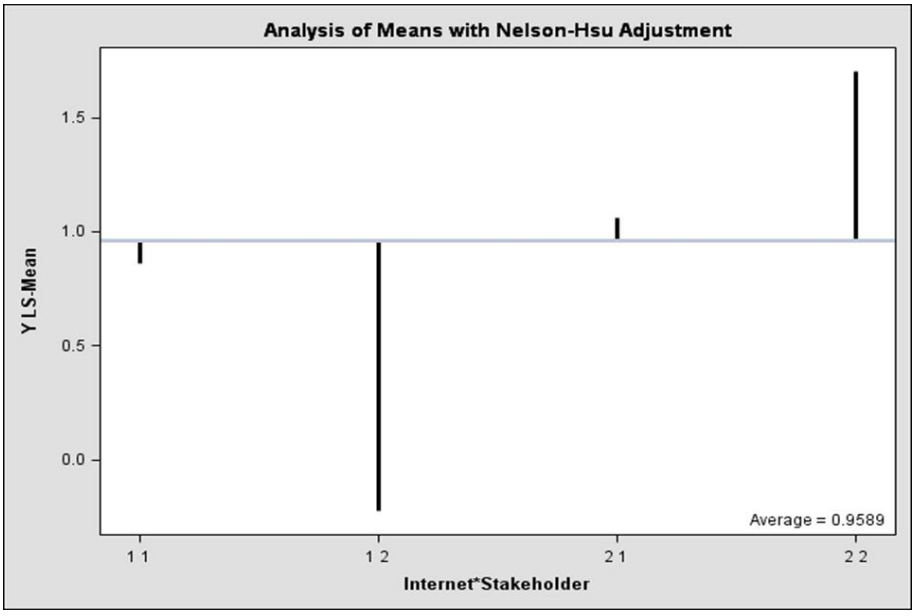

Figure 6.7: Analysis of Means for Internet and Stakeholders Interaction Effect

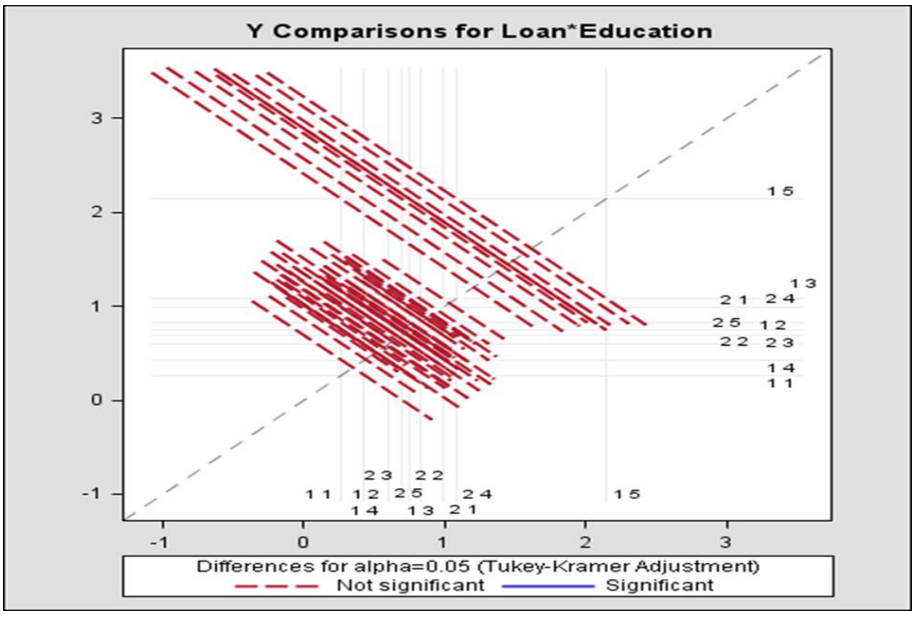

Figure 6.8: Diffogram for Outstanding Loan and Education Interaction Effect

The pair comparison of the least squares means, of the levels of businesses with outstanding loans and education level interaction effect, on business success, is presented in Figure 6.8. The pair comparisons of businesses with outstanding loans and education level interaction effect, are not significantly different between all levels. That is to say, entrepreneurs with no qualification and who have outstanding loan, entrepreneurs who completed primary school and who have outstanding loans, entrepreneurs who completed secondary school and have outstanding loans, entrepreneurs who have university degrees and have outstanding loans, and entrepreneurs who completed vocational training and have outstanding loans, have no significant differences. This implies that business success is not significantly different between all these levels.

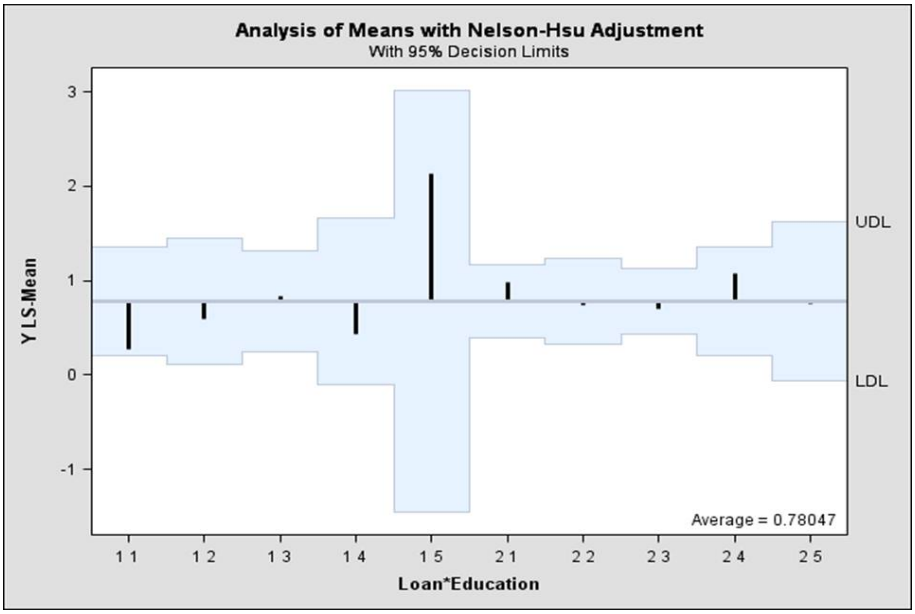

Figure 6.9: Analysis of Means for Outstanding Loan and Education Interaction Effect

Figure 6.9, displays analysis of means for businesses with outstanding loans and education level interaction effect, on business success. The average on the logit scale is 0.78047. As discussed in Figure 6.8, there are no significant differences of pairwise comparisons of least-squares means for businesses with outstanding loans and education level interaction effects. This can be noticed from Figure 6.9 where all the levels of businesses with outstanding loans and education level interaction effectd did not cross the 95% decision limit.

Chapter 7

Correspondence Analysis

Correspondence analysis is a multivariate technique uses for descriptive / exploratory analysis of simple two-way or multi-way contingency tables, containing some measure of correspondence between rows and columns (Greenacre and Blasius, 2006). Generally correspondence analysis is similar to the principal components analysis, the only difference is that the correspondence analysis is for categorical variables, whereas principal components analysis is for continuous variables. The technique was developed simultaneously in different parts of the world. For instance it has been popular in France because of the work of Jean Paul Benzecri and Pierre Bourdieu (Clausen, 1998). According to Clausen (1998), graphical interpretation will further the use of correspondence analysis. On the other hand, by using correspondence analysis, the relationship among categorical variables in large tables can be described in a summary form. Generally, the use of correspondence analysis has recently increased in many different fields, including economics, marketing, and business research. For instance, correspondence analysis can be used to graphically display a relationship between business location and type of business industry. One can examine for instance, how operations of different types of business industries differ from location to location.

Correspondence analysis is useful when analyzing associations between rows and columns in a contingency table (Hardle and Simar, 2003). Hard and Simar (2003), reported that the strength of correspondence analysis is in its capability of revealing the structure of a complex matrix, using a simpler matrix in a low dimension without losing essential information. The goal is to convert numerical information from a contingency table into a two-way dimensional

graphical display. It is generally, applicable to categorical data, that is to say it allows a researcher/analyst to visualize relationships among categories of categorical variables for large or complex data sets (Hardle and Simar, 2003). Generally, there are three steps involve in calculating categorical profiles in correspondence analysis. First, the row and column profiles are calculated. Note that the average row profile is the weighted means of the row profiles, whereas the average column profile is the weighted means of the column profiles. Second, the distance between points are calculated. This can be done by calculating the chi-square distance, which is interpreted as the weighted Euclidean distance between two points. The third step is to find the n-dimension space that best fits the data and to adjust the axes. The general idea is to minimize the distance between the axes and the points, by rotating the axes. This simply means that we want to minimize $\sum rd^2$, where r=row mass, and d^2= square distance from the point to the axis. Thus, the maximum number of dimensions is the minimum number of rows and columns. Hence, the smallest number of dimensions that captures most of the information will be selected based on the percent variance of the eigenvalues. A simple correspondence analysis can be applied if the contingency table is a 2×2 matrix (categorical Variable $Q = 2$), whereas a multiple correspondence analysis is applicable if the contingency table involves more than 2×2 matrix (categorical Variable $Q > 2$).

Definition and Notations

Let $\mathbf{N}$ be a $I \times J$ original matrix representing a contingency table of categorical variables with I row ($i = 1, 2, \ldots I$) and J column ($j = 1, 2, \ldots J$) having elements n_{ij}.

Then the row sum is denoted by $n_{i.}$ and is obtain as $n_{i.} = \sum_j n_{ij}$, and the column sum is denoted by $n_{.j}$ which is obtain as $\sum_i n_{ij}$. Thus, the grand total is denoted by n and is obtain as $n = \sum_i \sum_j n_{ij}$. Recall that, the first step in calculating categorical profiles for the correspondence analysis is to calculate the row and the column profiles. The profile of each row i is obtain by dividing the rows of the original Matrix $\mathbf{N}$ by their respective row sum, that is $i = n_{ij}/n_{i.}$, where $j = 1, 2, \ldots, J$. And the row mass r_i is obtained as row sum $n_{i.}$ of

the original matrix $\mathbf{N}$ divided by the grand total n, that is

$$r_i = \frac{n_{i.}}{n}$$

where the vector of row masses is denoted by r. Thus, the average row profile ($\bar{r}$) is obtain by the column sums divided by the original matrix $\mathbf{N}$ that is

$$\bar{r} = \frac{n_{.j}}{N}, \quad j = 1, 2, \ldots, J$$

The profile of each column j is obtain by dividing the columns of the original table $\mathbf{N}$ by their respective column sum, that is $j = n_{ij}/n_{.j}$, where $i = 1, 2, \ldots, I$. The column mass c_j is then obtain by column sums $n_{.j}$ of the original matrix $\mathbf{N}$ divided by the grand total n, that is

$$c_j = \frac{n_{.j}}{n}$$

where the vector of column masses is denoted by c. Hence, the average column profile ($\bar{c}$) is obtain by the row sum divided by the original matrix $\mathbf{N}$ that is

$$\bar{c} = \frac{n_{i.}}{N}, \quad i = 1, 2, \ldots, I$$

Correspondence Matrix

The correspondence matrix denoted by $\mathbf{P}$, can be obtain by dividing the original matrix $\mathbf{N}$ by the grand total n, that is

$$\mathbf{P} = \frac{\mathbf{N}}{n}$$

The centroid of the set of row points is the vector of columns masses (c), whereas the centroid of the set of column points is the vector of the row masses (r). These are the average row and column profiles respectively.

Generally, to perform the analysis with respect to the center of gravity, the correspondence matrix $\mathbf{P}$ is centered symmetrically by rows and columns, that is $\mathbf{P} - \mathbf{rc}^T$ so the $\mathbf{P}$ is correspondence to the average profiles of both sets of points. The solution to finding representation of both sets of points can be obtained by the singular value decomposition of the matrix of the standardized residuals.

The standardized residuals: consists of $I \times J$ matrix given by

$$\mathbf{S} = \mathbf{D_r}^{-1/2}(\mathbf{P} - \mathbf{rc}^T)\mathbf{D_c}^{-1/2} \tag{7.1}$$

with elements $s_{ij} = \frac{p_{ij} - r_i cj}{\sqrt{r_i c_j}}$. According to Greenacre and Blasius (2006), the sum of the squared elements of the matrix of the standardized residuals, that is

$$\sum_i \sum_j s_{ij}^2 = \text{trace}(\mathbf{SS}^T)$$

is the total inertia which is the amount that quantifies the total variance in the cross-table. Hence, the total inertia is given as

$$\text{total inertia} = \frac{\chi^2}{n}$$

The Singular Value Decomposition (SVD) of $I \times J$ matrix $\mathbf{S}$ according to Greenacre and Blasius (2006) is reveal by

$$\mathbf{S} = \mathbf{U\Gamma V}^T$$

where Γ is the diagonal matrix of positive number in descending order $\gamma_1 \geq \gamma_2 \geq \ldots \geq \gamma_s > 0$, which are the singular values. s is the rank of $\mathbf{S}$, and the matrices $\mathbf{U}$ and $\mathbf{V}$ are orthonormal. That is the column of the matrix $\mathbf{U}$ are the left singular vectors given as $\mathbf{U}^T\mathbf{V} = \mathbf{I}$ $(u_1, u_2, \ldots, u_k)$ whereas, the column of $\mathbf{V}$ are the right singular vectors given by $\mathbf{V}^T\mathbf{U} = \mathbf{I}$ $(v_1, v_2, \ldots, v_k)$. Note that, the singular values which are the square root of the eigenvalues can be used to describe the contribution of each dimension to the total variance.

Recall that the second step in calculating the categorical profile in correspondence analysis is to calculate the distance between points. Therefore, the distances between the two points can be measured by the chi-square distance which can be interpreted as weighted Euclidean distance. The chi-square depict the distances between profile points. In this study we use the chi-square distance to make significance dependence inference test for distances between points for the correspondence analysis. This is because the chi-square distance satisfies the equivalence principle. Thus, the chi-square statistic for the contingency table is then given by

$$\chi^2 = n \sum_i \sum_j \frac{(p_{ij} - r_i c_j)^2}{r_i c_j}$$

Recall that the third step in calculating the categorical profile in the correspondence analysis is to find the n-dimension space that best fit the data and adjust the axes. According to Greenacre and Blasius 2006, this can be done using the correspondence map which can be obtained from the results of the SVD. Thus, the principal and the standard coordinates can be calculated for the row and the columns categories as follows:

- principal coordinates of rows: $\mathbf{F} = \mathbf{D}_r^{-1/2}\mathbf{U}\Gamma$
- standard coordinates of rows: $\mathbf{A} = \mathbf{D}_r^{-1/2}\mathbf{U}$
- principal coordinates of columns: $\mathbf{G} = \mathbf{D}_c^{-1/2}\mathbf{U}\mathbf{V}\Gamma$
- standard coordinates of columns: $\mathbf{B} = \mathbf{D}_c^{-1/2}\mathbf{U}\mathbf{V}$

7.1 Multiple Correspondence Analysis

Multiple correspondence analysis is use to quantifies nominal or categorial data by assigning numerical values to the objects (cases) and categories so that the objects within the same category are close together and object in different categories are far apart (Roux and Rouanet, 2010). Generally the multiple correspondence analysis is applicable when the categorial variables in the continency table exceed two.

Let $\mathbf{Z}$ be an indicator matrix, having binary entries which represents the data with n categorical variables and m observations. Then the transpose of $\mathbf{Z}$ is given by $\mathbf{Z}^T\mathbf{Z}$. Therefore the Burt matrix $\mathbf{B}$ is then given by $\mathbf{B} = \mathbf{Z}^T Z$. Where $\mathbf{B}$ is a symmetric matrix that contains all the pairwise cross-tabulation of the categorical variables including the cross tabulations of each variable with itself. The coordinates in a maximum of k-dimensions of all categories is obtain through the Burt matrix (Greenacre and Blasius, 2006). Results from multiple correspondence analysis can be displayed by a biplot, or plot of joint categories. The biplot generally, reconstract data into a joint map of rows and columns. Refer to Greenacre and Blauis 2006 for further details on the multiple correspondence analysis.

7.2 Interpretation

"The common practice of interpreting, in the usual correspondence analysis display, distances between row and column points is controversial" (John and Robert, 2000). Generally, it is preferable to summarize the row and column coordinates in a single plot. From the plot results in correspondence analysis, the similarities of profiles between two variables are reflected in the distance between their two points (chi-square distance). This implies that if two points are close to each other, it can be concluded that their profiles are similar. For instance, if two rows have similar profiles, then their distributions are similar across the columns, whereas, if the two points are far apart, then they have different profiles. It should be noted that one can only interpret the distances between row points, and distances between column points, but not the distance between row point and column point (Bendixen, 1996). The centroid represents the average profile. Hence, if points are located away from the centroid or the origin, then they are distinctly different from the average profile.

In multiple correspondence analysis, the Cronbach's alpha can also be used to measure internal consistency or reliability among the variables. Thus, a higher value of the Cronbach's alpha coefficient implies that there is a relatively high internal consistency among variables. Refer to Greenacre and Blasius (2006) for further details concerning interrelation of correspondence analysis results. Generally, we use SPSS to perform correspondence analysis. The results of the analysis are presented in Section 7.3.

7.3 Results

We present the join plot of the categorical points of the response variable and the explanatory variables in Figure 7.1. Hence, we use the chi-square distance to check if there is a significant dependency between the variables. The interpretation of the relationship between variables is based on the point distance between the two variables in the row and the column, and on their contribution to the chi-square distances. Note that the distance between two variables is homogenous if they have similar profiles or similar response patterns.

From Figure 7.1 it can be observed that business locations Malakal, Rumbek, Yambio,

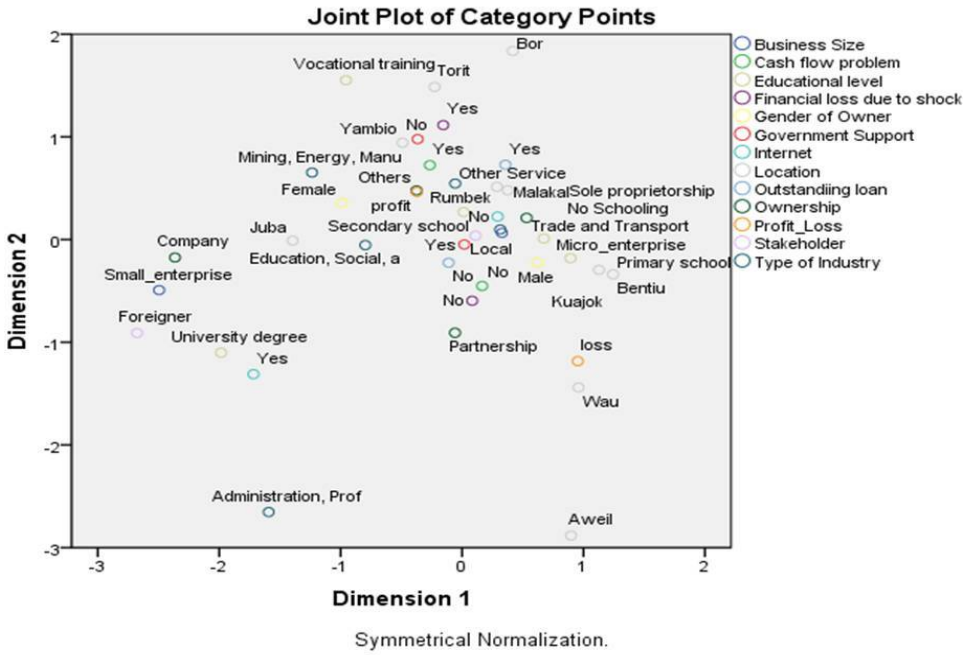

Figure 7.1: Join Plot of Categorical Points

Torit, Bor, and Juba have similar profiles, indicating that the rate of business success is high in these locations. Likewise, sole proprietorships and other types of business ownership have similar profiles, suggesting that the rate of success is high for sole proprietorships and other types of business ownership, compared to companies and partnership types of business ownership. The profiles of businesses who are internet users is different from those who are not internet users. The profiles of businesses that received support from the government and those that did not receive support are similar, implying that the rate of success is the same for both businesses. The profiles of businesses with cash flow problems and businesses with no cash flow problems are different, implying that the rate of success is different for businesses with / without cash flow problems. Similarly, businesses with financial loss due to shock and those with no financial loss due to shock have different profiles. Likewise, profiles of small-enterprises are different from those of micro-enterprises, suggesting that the rate of success among small-enterprise are different from micro-enterprise. Businesses with outstanding loans and business without outstanding loans also have similar profile. The rate of success is higher for female entrepreneurs compared to male entrepreneurs. Moreover, the

success rate is high for entrepreneurs who have completed secondary school, as to compared to those who have undergone vocational training, completed primary school and those with university degrees. Likewise, the rate of success is high for businesses with local stakeholders, as compared to those with foreign stakeholders. Businesses investing in mining, energy, manufacturing, construction, trade, transport, and other services have high rates of business success as compared to businesses investing in administrative, professional, and scientific services.

Chapter 8

Conclusion

8.1 Conclusions

The objective of the study was to identify factors that determine business success. The findings of the study highlights the factors for successful business investment in the country. The factors that determines business success were identied using four statistical models. These models are: the generalized linear model; the survey logistic regression model; the generalized linear mixed model; and correspondence analysis. First, a special case of the generalized linear model called the logistic regression model, was fitted to the data. A main effect model was fitted to the data and then a two-way interaction effects was allowed in the models. The likelihood ratio, the Score, and the Wald test statistics were used to make inferences about the null hypothesis. The Hosmer-Lemeshow goodness-of-fit test was used to assess the model fit. Model checking on the choice of the link function, influencial observations, and predictive accuracy power of the model was also done. All the model diagnostics favours the appropriateness of the selected fit. Accordingly, inferences were done based on the model results. However, since the logistic regression model is only appropriate for data from simple random sampling, and the data used in this study is from stratified random sampling, the selected model was refitted to the data using the survey logistic regression model which accounts for complexity of survey design. Since the business survey also includes random effects (PSUs), the selected model was refitted again using the generalized linear mixed model which also accounts for complexity of survey design. Both the survey logistic regression model and the generalized linear mixed model fitted the data well. Generally, all the four statistical models

used in this study indicated that the following are found to be significant factors to business success: state (business location); use of internet; government support to business; cash flow problem; business outstanding loan; gender; startup capital; type of industry sector and; the interaction terms state by gender; internet by stakeholders; and outstanding loans by education level. To confirm the results from the model, we further performed an exploratory analysis using multiple correspondence analysis.

The study findings suggest that micro and small enterprises can gain more profit if the business environment in which they operate is improved especially in certain locations in South Sudan. The need for innovation is crucial for business success, for instance the introduction of new business ideas, use of technology is very important for business operation. Government support to business is also significantly crucial. Cash flow problems were also found to be significant factors for business success. If businesses run out of available liquid capital and are not able to purchase the needed inputs for operation, liquidity (cash flow) problems may cause them to fail. Hence, given the liquidity problems experienced by micro-enterprises, policies that encourage micro-finance activities should be developed. Such policies may bridge the gap between micro-enterprises' lack of collateral for loans, and banks' fear of loan defaults. Female entrepreneurs should be encouraged to compete with their male counterparts in business management. Such policies on gender equity should focus on training female entrepreneurs in the skills required to run a successful business. Lastly, decisions on which type of industry sector an investor may wish to invest in are also crucial for business success. Feasibility studies have to done before an investor may decide to invest in a particular industry sector.

The correspondence analysis results also confirms the results obtained from the three models. The correspondence analysis makes the visual interpretation easier. Some of the categories are grouped around the business success and some are grouped around business loss. For instance, business locations Juba, Yambio, Rumbek, Malakal, entrepreneurs who have completed secondary school, business with cash flow problem, industries investing in mining, energy, manufacturing, and construction, female entrepreneurs, other type of business ownership, other type of industries activities, local stakeholders, business who have

received support from the government as well as those that did not received support from government contributed significantly to business success. Whereas, business locations, Bentiu, Kuajok, Wau and Aweil, partnership type of business ownership, have most of the business loss.

One of the shortcomings of this study is that the number of explanatory variables are not exhaustive. Nevertheless, extra care was taken to include the main explanatory variables based on the literature. The second short coming was that some variables classification was much more than what is in the study. We merged some groups to allow the interaction effects have adequate number of observations. This is mainly done on business industry type, ownership structure, use of internet, government support and business stakeholders. The future direction of this study is to use a decision tree type of analysis to group the factors as a complement of correspondence analysis.

Bibliography

[1] Agresti, A., Booth, G.J., Hobert, P.J, and Caffo, B.,(2000). Random - Effects Modelling of Categorical Response Data: *Sociological Methodology Volume* 30, *Issue* 1, *pages* 27 − 80

[2] Agresti, A., (2002). Categorical Data Analysis, Second Edition: *Wiley Series in Probability and Statistics*.

[3] Abor, J., and Quartey, P., (2010). Issues in SME Development in Ghana and South Africa: *International Research Journal of Finance and Economics*, *issue* 39 (2010).

[4] Anthony B. (2002) Performing Logistic Regression on Survey Data with the New SURVEYLOGISTIC Procedure: *SAS Institute Inc.*, *Cary*, *North Carolina*, *USA*. *Paper* 258-27.

[5] Archera, K.J, Lemeshow, S. and 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) 44504464.

[6] Bartelsman, E., Haltiwange, J. and Scarpetta, S. (2005). Measuring and Analyzing Cross-Country Differences in Firm Dynamics: *A selection from a published volume from the National Bureau of Economic Research*, *Title*: *Producer Dynamics*: *New Evidence from Micro Data*.

[7] Bendixen, M. (1996). A Practical Guide to the Use of Correspondence Analysis in Marketing Research: *Marketing Research On-Line Vol.One*, 1996.

[8] Bewley, H. Forth, J. and Robinson, C. (2010). Evaluation Methodology: Measurement of Drivers of Business Success and Failure: *Institute of Economic and Social Research*, *Report to the Department for Business Innovation and Skills*.

[9] Bray, J. (2007). The Role of Private Sector Development in Post-Conflict Economic Recovery: *UNDP*, *New York*.

[10] Breslow, N.E., and Clayton, D.G., (1993). Approximation Inference in Generalized Linear Mixed Models: *Journal of the American Statistical Association.*

[11] Buckland, S.T., Burnham, K.P., and Augustin. (1997). Model Selection: An Integral Part of Inference: *Biometrics*, **53**(**2**), $603 - 618$.

[12] Burnham, K.P., and Anderson, D.R., (2002). Model selection and Multimodel Inference: A practical Information-Theoretic Approach. *Second Edition*: *Spinger-Verlag New York Inc.*

[13] Chamber, R.L. and Skinner C.J. (2003). Analysis of Survey Data: *John Wiley and Sons ltd.*

[14] Chrysostome, E., and Rosson, P. (2004). The Internet and SMEs Internationalization: Promise and Illusions: *Paper presented at Conference of ASAC June* $5 - 8$, 2004 *Quebec.*

[15] Clausen, S.E. (1998). Applied Correspondence Analysis: An Introduction: *Series*: *Quantitative Application in the Social Science.*

[16] Coad, A. Blasco, A.S. and Teruel, M. (2010). Does Firm Performance Improve with Age?: *Social Science Research Network.*

[17] Cook, R.D., and Weisberg, s., (1994). Plots for Generalized Linear Mixed Models: *Computational Statistics and Data Analysis*, *Volume* 17, *pp.* $303 - 315$.

[18] Cramer J.S., (1999). Predictive Performance of Binary Logit Model in Unbalanced Sample: *Journal of Royal Statistical Society*, *volume* 48, *issue* 14, *pages* $85 - 94$.

[19] Davidsson, P., Kirchhoff, B., Hatemi A., and Gustavsson, H. (2002). Empirical Analysis of Business Growth Factors Using Swedish Data: *Journal of Small Business Management*, *Volume* 40, *Issue* 4, *Pages* 332-349, *october* 2002.

[20] Dibrell, C., Davis, P.S, and Craig, J. (2008). Fueling Innovation through Information Technology in SMEs: *Journal of Small Business Management*, *Volume* 46, *Issue* 2. *Page* 203-218, *April* 2008.

[21] Dobson, A.J, (2002). An Introduction to Generalized Linear Models, Second Edition: *A CRC Press Company*, *Boca Raton*, *London*, *New York*, *Washington*, *D.C.*

[22] Duijn, M.A.J.V., Gile, K.J., and Handcock, M.S., (2008). A framework for Comparison of Maximum Pseudo-Likelihood and Maximum-Likelihood of exponential family random

Graph Model. *Working Paper No.* 74 *Center for Statistics and the Social Sciences University of Washington.*

[23] Dunteman, G.H. (2006). An Introduction to Generalized Linear Models: *Quantitative Applications in Social Science*, *SAGE Publications.*

[24] Gilmour, A.R., Anderson, R.D., and Rae, A.L., (1985). The Analysis of Binomial Data by a Generalized Linear Mixed Model: *Biometric* (1985), 72, 3, *pp.* $593 - 9$.

[25] Greenacre M. and Blasius J. (2006) Multiple Correspondence Analysis and Related Methods: *Taylor and Francis Group.*

[26] Grunert, K.G, (1994). The Concept of Key Success Factors: Theory and Methods.

[27] Gupta V.K. Turban D.B. Wasti S.A. and Sikdar A. (2009). The Role of Gender Stereotypes in Perceptions of Entrepreneurs and Intentions to become an Entrepreneur: *Entrepreneurship Theory and Practice*, *Volume* 33, *Issue* 2. *Page* 397-417.

[28] Hardle, W. and Simar, L. (2003) Applied Multivariate Statistical Analysis: *Springer-Verlag Berlin Heidlberg New York.*

[29] Heeringa, S.G, West, B.T, and Berglund P.A. (2010). Applied Survey Data Analysis: *Taylor and Francis Group*, *LLC.*

[30] Hoffman. D. L. and Franke: G. R. (1986). Correspondence Analysis: Graphical Representation of Categorical Data in Marketing Research, *Journal of Marketing Research*, *Vol. XXIII*, (*August*). $p213 - 227$.

[31] Hosmer,D.W, and Lemeshow, S.(1989). Applied Logistic Regression, Second Edition: *John Wiley and Sons Ltd.*

[32] Hosmer, D.W, and Lemeshow, S. (2000). Applied Logistic Regression, Second Edition: *John Wiley and Sons LLC.*

[33] Hsu, J.C., and Peruggia, M., (1994). Graphical Representations of Tukey's Multiple Comparison Method: *Journal of Computational and Graphical Statistics*, *Vol.* 3, *No.* 2 (*Jun.*, 1994), *pp.* $143 - 161$.

[34] Hutcheson, Graeme, (2011). Generalized Linear Models: *SAGE Dictionary of Quantitative Management Research.*

[35] Indarti, N. (2004). Business Location and Success: A case Study of Internet Cafe Busi-

ness in Indonesia: *Gadjah Mada International Journal of Business*, *Volume* 2.

[36] Jackson, D.L., (2001). Sample Size and Number of Parameter Estimates in Maximum Likelihood Confirmatory Factor Analysis: A Monte Carlo Investigation. *Structural Equation Modeling*: *A multidisciplinary Journal*, $8:2, 205-223$.

[37] Jiang, J. (2007). Linear and Generalized Linear Mixed Models and their Applications: *New York*, 2007.

[38] John, M. and Robert, J. (2000). Correspondence Analysis of Two Mode Network Data: *Social Network* 22 (2000) 65- 75.

[39] Kantarelis, D. (2007). Theories of the Firm, 3^{th} *Edition*.

[40] Kepler E. and Shane S. (2007). Are Male and Female Entrepreneurs Really That Different?: *Small Business Research Summary*.

[41] Korn, E.L., and Graubard, B.I., (1995). Examples of Differing Weighted and unweighted Estimates from a sample survey: *The American Statistician*, *Volume* 49, *N*0. 3, *pp*-$291-295$.

[42] Kristiansen, S, Furuholt,Bjorn, Wahid and Fathul. (2003). Internet cafe entrepreneurs: pioneers in information dissemination in Indonesia: *International Journal of Entrepreneurship and Innovation*, *Volume* 4, *Number* 4. *P.P.* 251-263(13).

[43] Lee, Y., Nelder, J.A., and Pawitan, Y., (2006). Generalized Linear Model with Random Effects: A Unified Analysis via H-Likelihood: *Chapman and Hall / CRC*, *Taylor and Francis Group*.

[44] Leeuw, J. de., and Meijer E. (2008). Handbook of Multilevel Analysis: *Springer Science + Business Media*, *LLC* (2008).

[45] Lehtonen, R. and Pahkinen E.J. (1995). Practical Methods for Design and Analysis of Complex Surveys: *John Wiley and Sons Ltd*, *Baffins Lane*, *Chichester*, *West Sussex PO*19 1*UD*, *England*.

[46] Lindsey, J.K,(1997). Applying Generalized Linear Models: *Springer-Verlag New York*.

[47] Lindner, A. (2005). SME Statistics: Towards more Systematic Statistical Measurement of SME behaviour: *Expert Group Meeting on Industrial Statistics*, *Working Paper on SMEs and Entrepreneurship New York September* 2005.

[48] Littell, C.R., Milliken, A.G., Stroup, W.W., Woifinger, D.R., and Schabenberger, O., (2006). SAS System for Mixed Models: *Second Edition*, *Cary*, *NC*, *SAS Institute Inc.*

[49] Longenecter, J.G, Moore, C.W, Petty, J.W. and Palich, L.E. (2006). Small Business Management: An Entrepreneurial Emphasis. *USA.*

[50] Loderer, C.F. and Waelchli, U. (2010). Firm Age and Performance: *Social Science Research Network.*

[51] Marsh, H.W., and Balla, J., (1994). Goodness of fit in confirmatory factor analysis: the effects of sample size and model parsimony: *Quality* and *Quantity* 28: $185 - 217$, 1994.

[52] Mayers, R.H, Douglas C. and Vining G. (2002) Generalized Linear Models with Application in Egineering and the Science.

[53] McCullagh, P, and Nelder J.A. (1989). Generalized Linear Models, Second Edition: *Monographs on Statistics and Applied Probability* 37, *Chapman and Hall Ltd.*

[54] McCulloch, C. E., and Searle, S. R. (2001). Generalized Linear, and Mixed Models: *Wiley Series in Probability and Statistics* 2001.

[55] Menard, S. (2002). Applied Logistic Regression Analysis, Second Edition: *SAGE Publictions inc.*

[56] Mead, D. and Liedholm, C. (1998). The dynamics of micro and small enterprises in developing countries: *World Development Volume* 26 *Issue* 1.

[57] National Bureau of Statistic (NBS), (2010). Poverty in Southern Sudan Estimates from National Baseline Household Survey (NBHS).

[58] Olsson, U, (2002). Generalized Linear Models, An Applied Approach: *Sweden.*

[59] Pan, Z., and Lin D.Y., (2005). Goodness-of-fit method for Generalized Linear Mixed Models: *Biometric*, *volume* 61, *issue* 4, *may* (2005).

[60] Papadaki, E. and Chami, B. (2007). Growth Determinants of Micro-Business in Canada: *Small Business Policy Branch Industry Canada.*

[61] Pendergast, J. F., Gange, S. J., Newton, M. A., Lindstrom, M. J., Palta, M., and Fisher, M. R., (1996). A Survey of Methods for Analyzing Clustered Binary Response Data: *International Statistical Review*, $Vol.64$, $No.1$ $(Apr., 1996)$, $pp.89 - 118$.

[62] Ricketts, M. (2002). The Economics of Business Enterprises: An Introduction to Eco-

nomic Organization and the theory of the firm.

[63] Robb, A. and Fairlie, R.W. (2008). Determinants of Business Success: An Examination of Asian-Owned Businesses in the United States (February 2008): *The Australian National University Centre for Economic Policy Research*, *Discussion paper No.*569.

[64] Rogoff, E.G, Lee, M.S, and Suh, D.C. (2004). Attributions by Entrepreneurs and Experts of Factors that cause and Impede business success: *Journal of Small Business Management*, *Volume* 42, *issue* 4, *Pages* 364 – 376, *October* 2004.

[65] Rogers, M. (2004). Network, Firm Size and Innovation: *Small Business Economics*, 22:141-153, 2004.

[66] Roux, B.L. and Rouanet, H. (2010). Multiple Correspondence Analysis: *Quantitative Application in SocialScience.*

[67] Sarkar, S.K, Midi, H, and Rana, Sohel. (2011). Detecting of Outliers and Influential observation in Binary Logistic Regression: An Empirical Study: *Journal of Applied Science.*

[68] Saridakis, G, Mole, K, and Hay, G. (2007). Do Liquidity Constraints in the First year of Trading Reduce the Likelihood of Firm Growth and Survival? Evidence from England: *Institute for Small Business and Entrepreneurship*, *Glasgow*, *Scotland.*

[69] Satchell, S., and Xia, W., (2006). Analytic Models of Roc Curve: Applications to Credit Rating Model Validation: *Quantitative Finance Research Center*, *University of Technology Sydney*, *Research paper NO.* 181.

[70] SAS/STAT 9.2 User's Guide 2008: The SURVEYLOGISTIC Procedure: *SAS*, *Insitute Inc.*, *Cary*, *NC. USA.*

[71] SAS/STAT *Version* 8 User's Guide: Introduction to Survey Sampling and Analysis Procedures, (1999): *SAS*, *Insitute Inc.*, *Cary*, *NC. USA.*

[72] SAS/STAT 9.2 User's Guide 2008: The GLIMMIX Procedure: *SAS*, *Insitute Inc.*, *Cary*, *NC. USA.*

[73] Schabenberger, O., 2005. Introducing The GLIMMIX Procedure for Generalized Linear Mixed Models: *SUGI*, 30, *Proceedings.*

[74] Seiler, C. (2010). Dynamic Modelling of Business Surveys: *Ifo Institute for Economic Research at University of Munich*, *Working Paper No.*93.

[75] Spulber, D.F. (2009). The Theory of the Firm, Microeconomics with Endogenous, Entreprenuers, Firms Markets and Organization.

[76] Stokes, D. and Wilson, N (2010). Small Business Management and Entrepreneurship.

[77] Storey, D.J. (1994). Understanding Small Business Sector: *International Thomson Business press.*

[78] Varian, H. (2010). Intermediate Microeconomic: A modern Approach. *W.W. Norton* and *Company.*

[79] Vittinghof, E. Glidden, D.V, Shiboski, S.C and McCulloch, C.E. (2005). Regression Method in Biostatistics: Linear, Logistic, Survival and Repeated Measures Models: *Spinger*, *NewYork.*

[80] Wolfinger, R., and O'Connell, M., (1993). Generalized Linear Mixed Models; A Pseudo-Likelihood Approach: *Journal of Statistical Computation and Simulation.*

[81] Zelterman, D, (2002), Advanced Log-Linear Models using SAS: *SAS Institute Inc*, *Cary NC*, *USA.*

Appendix A

Generalized Linear Models SAS Procedures

We used the PROC LOGISTIC and the PROC GENMOD of the SAS system to select and fit the generalized linear model discussed and fitted in Chapter 4. The logit, probit, and the complementary log-log link functions were used.

A.1 Main-Effect Model

The main-effect model was fitted using the PRO LOGISTIC as follows:

ods html;

Proc logistic descending data=data;

class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 /param=ref;

model y= X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X3*X13 X9*X12 X1*X11/link=logit alpha=0.05 lackfit;

run;

ods html close;

Where, Y=Profit/Loss, X1=State (Business Location), X2=Ownership Structure, X3=Technology (Internet), X4=Government Support, X5=Liquidity Problem, X6=Financial loss due to shock, X7=Firm Size, X8=Firm Age, X9=Outstanding Loan, X10=Startup Capital, X11=Gender, X12=Education Level, X13=Stakeholders, X14=Industry Type. The option 'descending'=model the probability of Y=1. 'Lackfit' request the Hosmer-Lemeshow goodness of fit test for un-grouped binary response data.

A.2 Model Fitting using PROC GENMOD

The PROC GENMOD procedure was used to do further diagnostics, such as check for over-dispersion, calculation of the predicted probabilities, linear predictor statistics, and the residuals. This was implemented as follows:

ods html;

proc genmod descending data=data;

class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 /param=ref;

model y= X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X3*X13 X9*X12 X1*X11/dist=bin link=logit alpha=0.05 aggregate scale=deviance scale=pearson converge=1e-20 obstats type3;

run;

ods html close;

Where, 'aggregate'=specifies the subpoplution on which the pearson and the deviance are calculated. 'scale'=specifies the scale parameter for overdispered model, usually genuine for binomial or poisson distribution. 'converge'=sets the convergence criterion. 'obstats'=specifies an additional statistics including, residuals, predicted values, linear predictors and the dfbetas statistics. 'type3'=requests statistics for type3 contrast.

A.3 Plots using PROC LOGISTIC

1- Plots in PROC LOGISTIC can be done using the output statement, or by directly specifying the plot options in the PROC LOGISTIC statement or the model statement. The following are some of the plots done directly using the PROC LOGISTIC Statement and the model statement.

ods html;

ods graphics on;

Proc logistic descending data=data plot (only label)=(phat leverage dpc);

```
class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14;
model Y=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14/link=logit alpha=0.05
lackfit plcl outroc=rocl;
run;
ods graphics off;
ods html close;
```

Where, 'Phat leverage dpc'=plots the diagnostics for the leverage and influencial observations. 'outroc=rocl'= plots the *ROC* curve for the model predictive accuracy power.

2- Plot using the output statement in PROC LOGISTIC procedure:

```
ods html;
ods graphics on;
proc logistic descending data=data;
class X1 X2 X3 X4 X5 X6 X7 X9 X11 X12 X13 X14;
model Y=X1 X2 X3 X4 X5 X6 X7 X9 X11 X12 X13 X14/ link=logit alpha=0.05;
output out=sasuser p=pred xbeta=logit resdev=resdev;
run;
```

```
ods html;
ods graphics on;
proc gplot data=sasuser;
plot DevianceResidual*logit;
ods graphics off; ods html;
```

A.4 Plots using PROC GENMOD

The PROC GENMOD was used to plot the Cook's distance for influence diagnositics. The plot was done directly by specifying the plot option in the PROC GENMOD statement as follows:

```
ods html;
ods graphics on;
proc genmod descending data=data plots=cooksd;
class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14;
model Y=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14/dist=bin link=logit;
run;
ods graphics off;
ods html close;
```

The options 'plots=cooksd' plot the Cooks' distance for test of influential observations. whereas the 'plots=Predicted' was used to obtain the probability distribution of the predicted values for the logit, probit, and complementary log-log models respectively.

A.5 Checking Link Function Using PROC GENMOD

The choice of the link function was test using the PROC GENMOD as follows:

```
ods html;
proc genmod descending data=data;
model Y=lpred sqrtlpred/dist=bin link=logit;
run;
ods html close;
```

Where, 'lpred'=linear predictors. 'sqrtlpred'=Squared linear predictors.

Appendix B

Survey Logistic Model SAS Procedures

The survey logistic regression model discussed in Chapter 5, was fitted using the PROC SURVEYLOGISTIC. The same main effect model in Table 4.2 and Table 4.3 was refitted using the PROC SURVEYLOGISITIC procedure to account for the complexity of the survey design. The process was implemented as follows;

B.1 Model Fitting

The survey logistic regression model was fitted using the PROC SURVEYLOGISTIC as follows:

```
ods html;
proc surveylogistic data=data total=2000;
stratum stratum /list;
cluster PSU;
class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 / param=reference;
model Y(descending)=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X3*X13
X9*X12 X1*X11 /link=logit alpha=0.05;
weight samplingweight;
run;
ods html close;
```

Where, ‘stratum’=was classified according to business location, industry type, and number of employees as defined in Section 3.1. ‘PSU’=Primary Sampling Units ‘Total’=specifies an input data set that contains the stratum population totals. ‘list’=requests a summary of the stratification.

Appendix C

Generalized Linear Mixed Model SAS Procedures

The generalized linear mixed model (random intercept) discussed in Chapter 6, was fitted using the PROC GLIMMIX. The same main effect model in Table 4.2 and Table 4.3, was refitted using PROC GLIMMIX procedure to produce the the random intercept model. The procedure was implemented as follows;

ods html;

ods graphics on;

proc glimmix data=data asycov;

class class X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14;

model Y(descending)=X1 X2 X3 X4 X5 X6 X7 X8 X9 X10 X11 X12 X13 X14 X3*X13 X9*X12 X1*X11/dist=binary solution alpha=0.05 ddfm=kenwardroger ddfm=satterthwaite; DDFM=SATTERTHWAITE;

random intercept / subject=PSU type=vc;

lsmeans X1 X1*X12 X4*X6 X1*X11/ plots=diffplot adjust=tukey alpha=0.05;

lsmeans X1 X1*X12 X4*X6 X1*X11/ plots=anomplot adjust=nelson alpha=0.05;

ods graphics off;

ods html close;

Where, 'asycov' displays the asymptotic covariance matrix of the covariance parameter estimates, 'DDFM=KENWARDROGER'= is an approximation which involves adjustment of the estimated variance-covariance matrix of the fixed and random effect that accounts for

uncertainty that may exist when estimating **G** and **R** in the model. 'Type=vc' specifies the covariance structure, which in our case is the variance covariance structure.

When the 'DDFM=KENWARDROGER' is replaced with the 'DDFM=SATTERTHWAITE' in the model statement, the satterthwaite-based degree of freedom can be obtained without accounting for uncertainty when estimating **G** and **R**. These two options are included in the model statement after the solution is obtained. When both options were included in the model statement, we found that the estimates are not different from Table 6.3, Table 6.4, and Table 6.6 respectively. Note that Table 6.3 and Table 6.4 present the solution for the conditional binary response model, given the random PSU effects where the marginal covariance matrix is block-diagonals, and the observation from the PSU forms the blocks. The residual PL was used to estimate the variance estimates for the GLMM. Whereas the satterthwaite method and the containment method were used to calculate the degrees of freedom. The Newtom-Raphson with ridging optimization technique was used for the covariance parameter optimization. The objective function was obtained using the residual likelihood technique. Refer to SAS (9.2) User's Guide or Littel *et al.* (2006), for details on these methods.