STATISTICAL MODELLING AND ANALYSIS OF BUSINESS OWNERSHIP IN SOUTH AFRICA

by

LEPULANA SYLVESTER SEBOLA

DISSERTATION

Submitted in fulfillment of the requirement for the degree of

MASTERS OF SCIENCE

in

STATISTICS

in the

FACULTY OF SCIENCE AND AGRICULTURE

(School of Mathematical and Computer Sciences)

at the

UNIVERSITY OF LIMPOPO

SUPERVISOR: DR. KD MOLOI

CO-SUPERVISOR: PROF. A TESSERA

2023

Declaration

I Lepulana Sylvester Sebola declare that the dissertation hereby submitted to the University of Limpopo, for the degree of Master of Science in Statistics has been composed by myself and has not been submitted by me at this University or other institution for any degree or professional qualification. I further declare that this is my work in design and in execution, and that all material contained herein has been duly acknowledged.

Signature: *Mr L.S Sebola*
Date: *November 2022*

Abstract

Business owners play a vital role in the social and economic development of South Africa by providing employment and services to the country's citizens. South African infrastructure and community services are maintained by tax contributions collected from business owners. Furthermore, business owners also play the central part of circulating money in the county by providing services to the society in exchange of money. It is then evident that the success or failure of business owners directly affect the wealth of the country. In this study, the Quarterly Labour Force Survey (QLFS) and Survey of Employer and Survey Employers (SESE), both for 2017 from StatsSA was used, and the Generalised Linear Models (Multinomial Logistic Regression and Log-linear regression models) were applied to analyse and model business ownership in South Africa. The Chi-square statistic test from the descriptive statistics results showed that there is strong association between business ownership and the following categorical variables; gender, population group, marital status, age group, attended school, marital status, and province. The study utilised Multinomial Logistic Regression to identify factors affecting business ownership in South Africa. Gender, age group and attended school were three factors that are highly statistical significant with all their categorical levels having significant coefficients. Log-linear regression model was further used test if there was a significant interaction effect between four factors (own business, gender, population group and age group). The study found that only the 3-way effect interaction was significant, meaning that it had the high probability of improving the model than 4-way effect. Another objective of the study was to

analyse the accessibility of finance by business owners. The study applied the Log-linear model using 2017 SESE data and found that the black population group dominates in terms of financial accessibility, the female gender also had a greater chance of getting access to loans than the male counterpart. The study recommends research on business ownership using the post COVID-19 data to investigate the effect of the Corona Virus pandemic on business ownership in South Africa using statistical methods.

Key words: Business ownership, Generalised Linear Model, Multiple Logistic Regression, Log-linear regression, Covid-19

Dedication

To God, my parents and siblings who stood by me through thick and thin

Acknowledgements

I would like to express my heartiest thanks to God who made it possible for me to reach this stage even during the most difficult times of COVID-19. My heartfelt thanks to my supervisor and Head of Department Doctor K.D Moloi and my co-supervisor Professor A Tessera for their support and patience from the beginning until the end of my dissertation. Special thanks to the EDPSETA bursary for funding my masters of science studies. Many thanks to Miss Michelle Ntaba for her diligent proofreading on this dissertation. Not forgetting my parents and siblings for their infinite support, "***ke a leboga Ditlou!***". There would be no dissertation without the above-mentioned, may the All mighty God bless you all.

Contents

List of Figures

List of Tables

List of Abbreviations and Acronyms

LRM	Logistic Regression
GLMs	Generalised Linear Models
OR	Odds Ratio
PO	Proportional Odds
MRM	Multinomial Regression Model
MLRM	Multinomial Logistic Regression Model
MLE	Maximum Likelihood Estimate
CMs	Count Models
SADP	South African Development Plan
QLFS	Quarterly Labour Force Survey
SESE	Survey of Employers and Self-Employed
LLR	Log-Linear Regression
SAS	Statistical Analysis Software
RR	Relative Risk
BIC	Bayesian Information Criterion
NB	Negative Binomial
AIC	AKaike's Information Criterion
BIC	Bayesian Information Criterion
SC	Schwartz Criterion
CI	Confidence Interval

Chapter 1

Introduction and Background

1.1 Introduction and background

About 70% of business owners in South Africa started businesses due to unemployment, wanted an extra source of income and also for independence. South Africa is dominated by business owners from the Small,Medium and Micro Enterprises (SMMEs) sector of which the majority are black Africans between the ages of 35 and 40. 66% of employment in South Africa comes from SMMEs and only 12.5% of these jobs are for business owners themselves. In 2001 South Africa recorded almost 2.3 Million informal business owners and this number has decreased to 1.8 Million in 2017 due to the $2008/9$ global recession. Furthermore, over 50% of South African business owners have less than R1500 turnover and cannot retain or hire more employees (StatsSA, 2019).

South Africa is not unique from other developing countries in terms of challenges faced by business owners, and subsequently result in the increase of business failure rate. According to Mamman et al. (2019), the increase in the rate of business failure from the developing countries affects the local economic

stability as well as development growth.

1.2 Problem Statement

During the Apartheid era, Black South Africans were excluded from any meaningful participation in the economy Bhorat et al. (2018). This had a far reaching effect on all aspects of life in South Africa including business ownership. Black entrepreneurs lacked the skills and financial capacity to successfully finance and manage their businesses and were limited to small-scale trade in daily consumer products (Mudenda, 2013). After the political transition from apartheid to democracy in 1994, the South African government focused on the economic transformation by increasing the participation of black people in the economy with the introduction of Broad-Based Black Economic Empowerment (B-BBEE) in 2003. Although the formation of B-BBEE gave black South African citizens the opportunity to own businesses from different industries, concerns are still being raised by politicians and researchers on the racial, gender gap and other factors affecting the success business owners (Akinsomi et al., 2016). There is therefore a need to further investigate and identify other factors associated with business ownership using other methods. Hence, the present study intends to analyse business ownership by using statistical methods such Logistic regression and Log-linear regression.

1.3 Motivation of the study

The present study was motivated by the need to understand business ownership and associated factors affecting business ownership in South Africa. For the purpose of gaining more knowledge about business ownership, multiple Logistic Regression (LR) and count models will be utilised. Business ownership in

South Africa was not equitable for some categories of people, hence the government launched policies such as Broad-Based Black Economic Empowerment (B-BBEE) to redress inequalities around ownership (Irene, 2017).

One of the objectives of the B-BBEE Act was to achieve a substantial change in the racial and gender composition of business ownership and redress the economic marginalization of the disadvantaged groups under apartheid (Republic of South Africa, 2003). Some argue that, because of the policy, the participation of black South Africans in the economy has seen improvements. According to the Johannesburg Stock Exchange (2015), "As at end 2013, black South Africans hold at least 23% of the Top 100 companies listed on the Johannesburg Stock Exchange". This includes direct black ownership and black ownership through institutional investment schemes.

Several researchers doubt that the B-BBEE Act has significantly improved the meaningful participation of black South Africans in the economy of the country. Mudenda (2013) and Irene (2017), claim that the B-BBEE policy has not sufficiently reduced the ownership imbalances brought by the apartheid regime. Mudenda (2013) further states that the previously disadvantaged citizens are still not provided with equitable chances of participating in the wealth of the country through business ownership. Despite the effort made by the South African government, Irene (2017), Valdez and Romero (2018), Hikido (2018), Sixaba and Rogerson (2019) and Beresford (2020), argue that a significant racial and gender gap still exists in business ownership. Odeku and Rudolf (2019) in their found that most black South Africans do not have financial access, credit facilities, and subsequently fail to retain ownership of businesses.

With regard to gender, Seekings and Nattrass (2008), Ndhlovu and Spring (2009), Witbooi and Ukpere (2011) and Melton et al. (2019) argue that South

African black women are still underrepresented in business ownership compared to men. In the study conducted by Henning and Akoob (2017), 91% of female business owners reported that they have never attended any business training from the government or any private sector, and that might have inhibited the success of female-owned businesses. Hence, a lot has to be done by the South African government to ensure the equality of race and gender in business ownership.

The extent of business ownership by the different population groups and women varies by sector. World Bank (2017) reported that over 90% of business owners are from Small, Medium and Micro Enterprise (SMME) sector. The contribution of SMMEs towards the Gross Domestic Product (GDP) has been above 30%, and moreover, SMMEs generates over 75% of employment in South Africa StatsSA (2019). According to a report by Department of Trade and Industry (2019), 62% of formal small business owners in the early 2002 were white. Their share in 2017 fell to about 45%. In contrast, black South Africans consistently owned around 90% of informal businesses. A quarter of formal small business and just under half of informal small business was owned by women in 2017.

Several researchers have studied business ownership using non-statistical methods such as thematic analysis, exploratory methods, and case study analysis. This study will apply statistical methods like multiple Logistic Regression (LR) and count models as those methods have not been mostly used in studying business ownership. LR models are mostly resistant to over-fitting by the application of regularization techniques and further makes no assumptions about the distribution (Pohar et al., 2004). While, count models assume that the errors do not follow a Normal distribution, but Poisson distribution. In count models, the response variables are not modelled as the linear functions of regression coefficients, but only modelling the natural log of dependent variables (Zeileis et al.,

2008). Hence, the lack of thoroughness in the non-statistical methods and gap in the literature also motivated the recent study to use statistical methods to model and analyse business ownership.

This study intends to use statistical modelling to gain deeper understanding of business ownership and suggest ways of redressing the problem of inequitable business ownership in South Africa. Statistical modelling techniques such as Logistic Regression and Count Models will be used to analyse and identify factors that determine business ownership in South Africa. The study will use the 2017 Quarterly Labour Force Survey (QLFS) data and the 2017 Survey of Employers and the Self-employed (SESE) collected by Statistics South Africa (Stats SA).

1.3.1 Aim

The aim of the study is to model and analyse business ownership in South Africa using statistical models.

1.3.2 Objectives

The objectives of the study are to;

1) Utilise Logistic Regression (LR) model and Count Models (CMs) to model business ownership.
2) Analyse the accessibility of finance by business owners.
3) Perform a comparative study of LR model and CMs.
4) Make recommendation on which statistical method can be preferred to model business ownership.

1.4 Scientific contribution

The identification of factors that affect business ownership may help the South African government to come up with the necessary interventions to support South African business owners. This research can also be helpful to the South African Development Plan (SADP) for 2030 which aims to promote business ownership for job creation and economic growth in the country. The current business owners in South Africa will also be aware of challenges that may arise in the near future and they can do essential preparations and planning. By comparing Logistic Regression (LR) and Count Models(CMs), the study will also check the best method in addressing the problem and that will be a great contribution to the existing literature.

1.5 Structure of the research

The research structure is as follows; The first chapter (Chapter 1) consists of the background of the study, the problem statement, the motivation of the study as well as the scientific contribution of this research. Chapter 2 of the research consists of the reviewed literatures related to the study followed by Chapter 3 which states descriptive and statistical methodology of the study and how the data was analyzed. Chapter 4 is the analysis and the results followed by the discussions, conclusions and recommendations in Chapter 5.

Chapter 2

Literature Review

2.1 Introduction

This chapter consists of the following sections; introduction which is Section 2.1 followed by Section 2.2 which is the historical review of business ownership in South Africa. Section 2.3 looks at the contribution of business ownership in the economy, followed by Section 2.4 which is the review of the related literatures and lastly Section 2.5 will be the conclusion.

2.2 Business ownership history in South Africa

During the Apartheid era, business ownership in South Africa was mostly dominated by the white minorities. The government led by South African National Party (NP) which came to power in 1948 restricted most black owned business operations in certain parts or areas of the country. South African black business owners were not allowed to have large-scale businesses and they could

only trade on the daily consumable products, including coffee, breads, soaps and newspapers (Ponte et al., 2007).

In 1969 the well known organisation called National African Federated Chamber of Commerce (NAFCOC) was established with the aim of promoting fairness in business ownership for all South Africans, however, NAFCOC goals only became successful at the end of $1976s$. Policies initiated by NAFCOC gave black business owners opportunities to make businesses in several industries, such as banking, publishing, retailing and construction. In spite of all the initiated policies, there were still other restrictive measures against black business owners in shareholding (Maserumule, 2015). According to Ponte et al. (2007), in 1990, NAFCOC formed the black economic empowerment programs to continue creating equity in business ownership in South Africa

South Africa became a democratic country in 1994 and held the first democratic elections where every citizen had an opportunity to vote regardless of gender and race. The post-apartheid government led by African National Congress (ANC) allowed most black South Africans to start their own businesses and register their companies on the Johannesburg Stock Exchange (JSE). Presently, there is no discrimination in the South African financial institution and the socio-economic inequality is thus low (Sixaba and Rogerson, 2019). Black-owned businesses are now found in various economic sectors in South Africa and black South Africans now have access to proper education to acquire necessary ownership skills (Department of Trade and Industry, 2019).

2.3 Contribution of business ownership in the economy

The contribution of business ownership in the global and local economy have been studied by several researchers including Clover and Darroch (2005a), Frese et al. (2007), Kongolo (2010), Ladzani (2010), Abor and Quartey (2010), Dunlavy et al. (2017), Riaz and Batool (2018) and Mamman et al. (2019). The aforementioned also applied different methods and data sets in their studies, and the majority confirmed that business owners do not only contribute in the global economy, but also play a vital role in the development of the local communities where their businesses are based.

According to Maliranta and Nurmi (2019), Small and Medium Enterprises (SMEs) are currently leading with the number business owners around the world. It was reported by World Bank (2017) that 90% of worldwide businesses are from SMEs sector and about 40% of Gross Domestic Products (GDP) in most developing countries is contributed by those SMEs. Hence, the promotion of Business ownership should be a keen interest to most governments across the world and ensure a stable and consistent support towards business owners.

South African government have recognized the contribution of SMEs to the country's GDP and further initiated the Department of Small Business Development (DSBD) to monitor and provide necessary support to South African business owners. However, concerns have been raised by most business owners from various industries in South Africa about the dissatisfaction of the support provided by the DSBD (Colin, 2019). It is therefore a clear indication that more has to be done by South African government to support local business owners.

2.4 Related literature reviewed

Several scholars from different countries have also studied factors that affect business ownership. Despite the different methodologies which were applied by researchers including Seekings and Nattrass (2008), Ndhlovu and Spring (2009), Irene (2017),Henning and Akoob (2017), Sixaba and Rogerson (2019) and Odeku and Rudolf (2019), factors such as gender, age group, population group, financial constraints and education background were identified to be related with business ownership.

2.4.1 Gender

The role of gender in business ownership has been broadly studied around the world and also in South Africa. Phillips et al. (2014) show how the South African government support female business owners through the initiated policies such Black Business Council (BBC), Business Unity South Africa (BUSA), Business Leadership South Africa (BLSA) and Black Economic Empowerment (BEE) to improved the rate of female business ownership in the country. However, the results of the study show that female business owners are still not satisfied with the government support as it was found that 77% of female business owners participated in the study, have never received any form assistance from the government officials.

Mitchell (2004) conducted a study on motivation of business ownership using the case study of South Africa. A sample of 690 business owners was selected. Chi-square statistic test was performed to test if there is a significant difference between male and female business owners. The study indicates a Chi-square value of 13.32 with the corresponding P-value less that 0.01, which implies that there is a significant difference between male and female business owners in South Africa. Based on the fundings, male and female business owners in

South Africa experience different challenges in business. The study recommends the establishment of policies that will address specific challenges faced by female business owners in South Africa.

Kelley et al. (2011) found that between the year 2001 and 2008 there was an increase in the number of female business owners across the globe. However, the study indicates that the increase is not significant since on average, chances for females to start and own businesses are still twice less compared to the chances for the male counterpart. According the study, majority of women especially in African countries are underrepresented in business ownership because they are perceived by the community to be caregivers, therefore their duty is to remain at home and take care of children. Thus, most women are not motivated to start businesses compared to men.

South African businesswomen have been complaining about the financial exclusion by the government. Witbooi and Ukpere (2011) indicate that gender gap still exist across all population groups in South Africa when it comes to business ownership. The study emphasised that, although women are regarded as good debit payers across the world, majority in South Africa are still financially excluded and have less business opportunities compared to men. The study indicates that most women continue to occupy the traditional roles of an "African women", which amongst them includes, staying at home, bearing children, doing cheap labour in the households. Hence, the participation of female business owners in South Africa is low.

Brijlal et al. (2013) investigated the influence of gender on the success of business owners in Western Cape province, South Africa. 369 business owners were randomly interviewed of which 71% were men and 21% were women. Results from cross tabulation indicates that majority of women start owning businesses

at younger age compared to men. Moreover, male business owners were found to be more successful than female business owners. According to the study, the difference in the level of education between male and female business owners was not found to be significant, which implies that majority of female and male business owners have equal education level.

Mboko and Smith-Hunter (2010) conducted a case study design method in the qualitative research on female business owners from Zimbabwe. The main aim of the study was to analyze the development and investigate the surviving strategies of Zimbabwean female business owners. The study also focused on how the environment can affect or influence the success of female ownership behavior. The data was collected from female business owners working in the textile industry which is regarded as one of the most popular industries dominated by females in most African countries. It was discovered that over 60% of female business owners in the textile industries have formal qualifications and some are in teaching and nursing professions, however, almost all those female business owners lack business training and experience. The study found that almost 90% of those females started businesses due to the less salary they earned from their professions. Furthermore, the study reached the common conclusion with Fairlie and Robb (2009) and Giandrea et al. (2008) in terms of business performance that most female owned businesses perform worse than male owned businesses due to inability of women to outlast the pressure in the business world.

Leoni and Falk (2010) used binary probit model on cross sectional data to identify the factors associated with business ownership in Australia. The research outcomes show a 0.4% increase in probability of a man to become a business owner and a 0.24% increase probability for a female to become a business owner. According to the study, the is a huge gender gap between old business owners

(age≥ 35) than between young business owners (age< 35) .

Regression and decomposition techniques was applied by Fairlie and Robb (2009) to study the performance of male-owned and female-owned businesses using the United States (U.S.) Census Bureau data. Female-owned businesses were found to have lower survival rate than male-owned businesses. On average, the probability that a female-owned business and male-owned business closes was 24.4% and 21.6%, respectively. The other outcome from the study was that, female-owned businesses have lower financial turnover than the male-owned businesses.

Carter and Weeks (2002) conducted study on business ownership in Germany. One of the objectives in the study was to investigate the perception of women towards business ownership. The outcomes indicates that majority of females prefer to be business owners due the flexibility and more earnings.

2.4.2 Age

Chipeta et al. (2016) studied the influence of age on business ownership intentions using th case study of university students in Gauteng province, South Africa. The researcher used convenience sampling method to select 350 students. The study defines students with age between 18 and 24 as young adults and students aged 30 years old or more as old adults. Age was one of the factors identified by Principal Component Analysis (PCA) and the Analysis Of Variance (ANOVA) was used to test the significant difference between young and old adults. According to the findings, the mean scores of young and old students was 4.04 and 4.83, respectively. ANOVA showed that there is a significant difference between the mean scores. The study concludes that, old students are more likely to become business owners than young students.

According to Van Scheers (2010) findings on the role of ethnicity and culture in the development of business owners in South Africa (SA), it was found that 37.5% business owners are in the age-group $(30 - 39)$. Age was identified to be one of the significant factors affecting development business owners in South Africa.

International Finance Corporation (2019) investigated on the trends and challenges faced by Small, Medium and Micro Enterprises (SMMEs) owners in South Africa between 2008 and 2017. The descriptive results from the report indicates a fair increase in the number of business owners between the age of 25 and 34 since 2008. The report also shows that the rate of business ownership for individuals of age 35 and more is higher than for individuals between the age of 25 and 34. According to the report, the increase in the number of young South African business owners is not significant as the rate of business ownership in South Africa is still very low compared to other countries.

Giandrea et al. (2008) employed the multivariate analysis technique on Health and Retirement Survey (HRS) data to study business ownership. The data consists of over $12,500$ participants from United State of America. The study results shows age as one of the factors contributing to the business ownership rate. The study indicates an increase in number of business owners with age and the majority of old people aged $(51 - 60)$ had greater chances of owning businesses than young people. The study suggested that the increase was due the financial savings and investments of old people, which grant them better chances of starting businesses. Recommendation made in the study was that, retired workers should be encouraged to take interest in business ownership so that they can remain active and valuable in the economy.

Hatfield (2015) studied business ownership in United Kingdom (UK) and dis-

covered that the rate of business ownership increases with age. Results in the study indicate that, 20% of UK employees within the age group of $(55-56)$ are owning businesses compared to 5% of young employees within age group $(15-24)$. Among the study results, a high volatility in business ownership rate was discovered on age group $(15-24)$.

Van Praag (2003) performed the duration analysis on the survival and the success of young small business owners in the United State of America (USA). The data was obtained from the National Longitudinal Survey of Youth (NLSY). Survival analysis was used to study the survival of small business owners and the researcher firstly outlined the difference between Ordinal Least Square (OLS) regression and the survival model. It was stated that, the survival model has some features that are unique with respect to the problem that are they are applied for. From the estimated results, age was found to be the significant variable, which means that business owners that have started owning businesses at an older age are likely to survive and be successful in their businesses than young business owners. The study argued that, the success of old business owners is due to experience and financial stability.

2.4.3 Population group

Preisendoerfer et al. (2014) conducted a study on why there is lack of black business owners in Eastern cape, South Africa. Experts, including lecturers, government officials, consultants, managers and Chief Executive Officers (CEO) from different companies were interviewed. Almost all experts indicated that the apartheid historical background of South Africa contributes to the lack of black business owners. The second dominated reason from experts was lack of financial resources. The study concluded that majority of black South African are recovering from the apartheid sanctions and the inequality

is still a burden is South Africa.

According to Global Entrepreneurship Monitor (2019) report, between 2017 and 2019 there has been an increase in the number of white business owners and a slightly decrease in the number of black business owners. The report argues that the decrease in the number of black business owners is due the low survival rate of black-owned businesses especially in Small, Medium and Micro Enterprises (SMMEs) sector and the argument is also supported by SEDA (2021).

Research conducted by Fairlie (2004) investigated on the trend in ethics and racial business ownership in United State of America (USA) using the Current Population Survey (CPS) data. Results from trend analysis indicates that Whites and Asians have the highest rate of business ownership than other racial groups. Within the white population group, 7.4% were female business owners while 13.1% were found to be male business owners. According to the study, black business owners were dominated by all racial groups. Findings suggest that the high rate of white business ownership in the past 20 years is due to the increase in number of white males in the American labour force. Although, the gap between the number of white and black business owners has been found to be declining and the researcher argues that it was due to an increase in the number of educated black men with respect to white men.

Fairlie and Robb (2007) in the other study used the Characteristic of Business Owner (CBO) survey to investigate the factors affecting black-owned businesses. Non-linear decomposition techniques was used. It was discovered that majority of black-owned businesses have lower average sales and returns, and the employees hired are fewer. Hence, these businesses are more likely to close or stop operating compared to the white-owned businesses. The study outlined

that the above mentioned findings are caused by the influence of the family backgrounds since most black business owners are less likely to have someone in the family who owned a business in the past, compared to the white business owners. According to study, about 12.6% of black business owners acquired business experience from their family background compared to 23.3% of white businesses owners.

2.4.4 Education background

Global Entrepreneurship Monitor (2017) report shows how business ownership rate in South Africa increased with the level of education between 2001 and 2017. According to the report, in 2017, 4.3% of business owners in South Africa between the age of 18 and 64 did not have formal education, 22.4% had secondary education while 52% had post-secondary education. Since 2001, the rate of business ownership in South Africa remained 2.2%, which is significantly low compared to the majority of the developing countries. The report emphasised the significance of quality education and rightful business skills to promote young business owners in South Africa. According to the study, level of education is one of the factors affecting business ownership in South Africa. The report concluded that the quality of education in South Africa is poor compared to other countries and it was supported by World Economic Forum (2019) report which shows that South Africa is rated position 114 in terms of quality of education out of 137 countries.

Peters and Brijlal (2011) investigated the relationship between the level of education and the success of business owners in South Africa. The study surveyed 320 businesses which were randomly selected within Kwa-Zulu Natal (KZN) districts. One of hypothesises formulated in the study was that there is no relationship between the level of education and the success of business owners.

The Chi-square test results indicated a P-value less than 0.05, which implies that the level of education and success of business owners are not independent, hence, the hypothesis was rejected. According to the study, the more educated a business owner gets, the higher chances of being successful in business and majority of business owners apply the knowledge they obtained at school into business.

Preisendoerfer et al. (2014) performed an empirical study with the aim of analysing factors that affect business ownership rate in South Africa. About 354 individuals were randomly selected during the study and some of them were business owners. From the results, education was found to be a contributing factor to the rate of business ownership. Results shows that as an individual gets more educated, the probability of becoming a business owner also increases. The study results supported the recent study by Bhorat et al. (2018), which indicated that education and financial skills are significant factors that affect business ownership in South Africa.

Mudenda (2013) researched on the perception of black business owners in Durban, (South Africa) towards the Black Economic Empowerment (BEE). The results showed that education background led over 62% individuals to become business owners. Most of the participants indicated that the level of education they obtained, had an influence on their decision to become business owners and only few received support from BEE. Almost half of business owners indicated that their entries in to business was also motivated by their work professions such as in law and education.

Roodt (2005) conducted a study on the factors related with business ownership in South Africa and the skills required. The descriptive results indicated that 45% of business owners in the study had at least grade 12 or higher qualifica-

tions, it also showed that 95% of graduated business owners became successful in their businesses. Findings from the study reveal that majority of graduated business owners believed that knowledge from technical education is necessary for the growth of any business. The above results were too similar to Dotson et al. (2013) study, which also identified education as one of the factors associated with the success of a business owner.

Maliranta and Nurmi (2019) employed the longitudinal employer-employees data of 2011 from Finland to analyze business ownership, employees and business performance. Econometrics model was used and the outcomes suggested that the likelihood of a business to be more productive and successful depends on the education level of the manager or the owner. However, the above arguments were challenged by the research conducted by Iversen et al. (2010), which indicated that not every education provided to a business owner can contribute to the success of his of her business.

2.4.5 Financial background

Leshilo and Lethoko (2017) conducted an exploratory study in Limpopo province (South Africa) on challenges faced by young business owners. A sample of 50 youth entrepreneurs participated in the study and the majority were between the age of 26 and 29. According to the results, 8% participants received financial assistance from banks, 12% were funded through government schemes, 20% were assisted by their family members and 60% used their personal savings to start businesses. The study discovered that majority of young business owners find it difficult to access financial assistance from banks as they do not have credit history. It was recommended in the study that South African government should ease certain regulations and lending policies in the financial institutions so that majority of young business owners can qualify for financial

assistance.

Sitharam and Hoque (2016) conducted a cross-sectional study on factors that affect the success of business owners in South Africa.The data was collected from 74 business owners in Kwa-Zulu Natal, South Africa, through online questionnaires. The results of the study show that 72% of business owners believe that financial access is one of the factors affecting business ownership in South Africa. According to the study, lack of funding contributes to the high rate of start-up business failure in South Africa.

Investigation on the relationship between business ownership and the economic development was conducted by Carrère et al. (2015) using 23 Organization for Economic Cooperation and Development (OECD) countries. One of the research's objectives was to identify the impact of an economic development on business ownership. An interrelationship model consisting an equation that deals with the cause of change and the equation of consequences was developed in the study. Weighted least square was applied to estimate the equilibrium relation parameters. The study found that there is a positive relationship between high rate of business ownership and level of per capita income in Italy.

2.4.6 Location

Clover and Darroch (2005a) studied the factors affecting business ownership in Small, Medium and Micro Enterprises (SMMEs) in Kwa-Zulu Natal (KZN), South Africa. The stratification method of sampling was utilized in the population of 266 agribusinesses funded by $Ithala$ Development Finance, and 44 SMMEs were sampled from Four strata. The aim of the study was to identify factors affecting SMMEs owners in KZN. Principal Component Analyses (PCA) was used to reduce number of variables and also summarises on the data ob-

tained from 36 correlated variables. The outcomes of the research showed that, insufficient government support (with mean $= 3.86$), lack of access to startup capital (with the mean $= 3.45$) and higher number of crimes (with mean $= 3.00$) were the leading factors that inhibit the success and development of SMME owners. The study concluded that most factors are influenced the location of the SMMEs since most of them are situated in rural areas.

The research on the labour markets (business ownership, unemployment and employment) in Finland was conducted by Tervo (2008) using the longitudinal census file data of $22,2644$ random sampled individuals. Markov chain analysis was employed to analyze the transition between business ownership, unemployment and employment. The transition matrix was estimated by Maximum Likelihood Estimator (MLE). To compare the transition matrix for both urban and rural areas, homogeneity was tested using the Chi-square statistic. The results revealed a significant difference in the transition probability for urban and rural labour markets. The study also disclosed a 6% ergodic probability of being a business owner in the urban labour market and 9.5% ergodic probability in rural labour market. Hence, in Finland there are higher chances of becoming a business owner from rural labour market. It was concluded that, since the rural labour markets are struggling from unemployment and lower population, therefore there is a need for business ownership to be enforced in the rural areas for economic stability.

2.4.7 Other factors

Fatoki (2014) utilized the Principal Component Analyses (PCA) to investigate the challenges encountered by young business owners in South Africa. According to the descriptive statistics results, lack of savings was found to be the leading factor that affect the success business owners with the highest mean of

4.96 and lack of knowledge in business was the second leading factor with the mean of 4.90. Factor analysis results indicates the highest Cronbach's alpha value of 0.861 and 0.730 for lack of savings and lack of knowledge in business, respectively. The study concludes that financial constrains and lack of skills and knowledge in business are the leading factors affecting young business owners in South Africa. The study recommended that entrepreneurship skills should be prioritized in most the higher institution of learning.

Van Scheers (2010) performed a correlation analyses to study the role of ethnicity and culture in the development of business owners in South Africa (SA). About 300 business owners were sampled in the Tshwane and Johannesburg region, 73% of business owners were men and 27% were women. Findings show that retail industry have the highest percentage (48.1%) of business owners, while the construction industry had the lowest percent of 1.3%. The study failed to reject the null hypothesis that, there is a positive relationship between ethnicity and the successful business owners in SA.

The study undertaken by Musara and Gwaindepi (2014) focused on factors that affect business ownership activity in South Africa. A sample of 153 business owners was selected and provided with questionnaires of which 150 were completed. Findings show that the majority (about 67%) of business owners indicated that corruption was one of the factors that negatively affected them when they were starting businesses. There was a strong positive correlation (r= 0.891) between corruption and delays in business registration process, which means that corruption may be caused by the delays in the process of business registration.

The cross-sectional results from Pretorius et al. (2005) on business ownership showed that 90% of business owners have above two years of experience in

business while 8.7% were start-up business owners. Ownership experience was found to be significant across all genders. The study further concluded that start-up business owners require more support as they encounter many challenges in business than experienced business owners.

Temtime and Pansiri (2004) undertook a study in Botswana applying PCA and Varimax rotation in investigating the critical success factors hindering the operations and the development of SMEs. The primary data was collected from 250 sampled SMEs owners. The descriptive results have shown 68% of male business owners and 32% of female business owners in the private sector. 65% of business owners were operating in the retail and wholesale industry, 25% in the service industry while only 10% were operating in manufacturing industry. Findings from the study indicate that about 21% of SMEs owners had 4 to 5 years management experience and 73% had at least five years experience. The PCA shows that, marketing action factor, socio-economic factor and fixed assets investments factor had a higher contribution to the failure of SMEs in Botswana with the means of $3.46, 4.04$ and 3.75, respectively. Personal factor and competitive strategy factor had the Cronbach's alpha values greater than 0.5 which was taken as a cut-off value to measure the reliability of the statistical inference. Factors with the alpha greater than 0.05 were considered to be critical factors affecting the development and the performance of SMEs in Botswana.

Gill and Biger (2012) conducted a non-experimental design on the selected business owners in Canada and applied multiple regression analysis to identify factors that are related to business owners in Canada. Business ownership was found to have a strong negative correlation with lack of management skills, lack of finance and also challenges in the market. The strong negative correlation results means that the above mentioned factors are related to busi-

ness ownership. The above findings were found to be similar to the studies conducted by Keyser et al. (2000),Moy and Luk (2003) and Okpara and Wynn (2007).

Hatfield (2015) studied business ownership in Europe using European Social Survey data. The study reported that about 14% of European employees are business owners. According to the study, the rate of business ownership in Northern and Western European countries is lower than the rate of business ownership in Southern and Eastern European economies. The study outlined that the great recession which took place between 2007 and 2009 had a huge impact on the business ownership and unemployment rate in European countries.

Guerra and Patuelli (2016) conducted a study at Switzerland using $650,000$ business owner data from the Federal Statistics Office (FSO). The aim of the study was to investigate factors that motivated individuals to become business owners. Empirical analysis results showed that about 92% of employees remain in the employees category and only 19% switch to the category of business owners one year later. Multinomial logistics model was utilized and the model fit were assessed by AIC and BIC which were found to be 171.74 and 2023.60, respectively. The Chi-square ratio test reported a high improvement of significant model after including the subjective variables. The overall conclusion from the study was that, the working conditions (e.g. annoying boss and colleagues or rigid working hours) push many employees in Switzerland to become business owners.

2.5 Conclusion

The above reviewed literatures identified factors that are related to business ownership in South Africa and globally, amongst factors, age, gender, education, financial background, population group were included. There are other factors such as marital status were not given sufficient attention by several studies. Hence, this present study will add such factors to investigate their effect on business ownership in South Africa.

The literature reviewed shows a numerous number studies on business owners in South Africa and globally. However, there is still lack of statistical methods such as Generalised Linear Models (GLMs) employed, especially in South Africa. Hence, the present study will utilise some GLMs such as Logistic regression and Log-linear regression to analyse and model business ownership in South Africa.

Chapter 3

Methodology

3.1 Introduction

The following chapter will cover the collection method and data analysis tools that were used in the study. The section will also outline in details, the statistical methodologies applied and also the rational behind the application. Generalized Linear Models (GLM) will be reviewed followed by two statistical analysis methods (Logistics regression and count models) and the inferential statistics.

3.2 Data collection and analysis

The Quarterly Labour Force (QLFS) and Survey of Employers and the Self-employed (SESE) of 2017 secondary data will be used for the purpose of the study, both datasets are collected from StatsSA. QLFS and SESE data will be analyzed by Statistical Analysis Software (SAS) version 9.2 and the R-studio statistical software is used as a supporting analysis tool and the final scripts

will be attached in the Appendix section.

3.3 Statistical Methodologies

The statistical methodologies used in the study includes Logistic Regression model and count models. All the two methods falls within GLMs, therefore, it is necessary to firstly review GLMs and its components. For the purpose of the study, Both count models and Logistic regression model are applied separately with the same objective to analyse and identify factors that determine business ownership in South Africa.

3.3.1 Generalized Linear Models

The GLMs was proposed by Nelder and Wedderburn (1972) with the aim of expanding the general linear models in a way that the covariates together with factors are linearly related to the response variable. In general linear model, the errors in the response variable are assumed to follow a normal distribution with mean μ and variance σ^2. The general linear models assume a linear relationship between independent and the dependent variable. However, in Generalised Linear Model, the errors of the response variable are assumed to belong in the exponential family of distribution, namely, Binomial, Poisson, Gamma, normal, Beta, Tweedie, Inverse Gaussian, Chi-square, Rayleigh and Negative binomial, amongst others. Hence, in Generalized Linear Models, the response variables need not to be continuous or following normal distribution. There are three components in every GLM.

3.3.2 Components of GLM

- Random components.

  Random component is used to identify the response variable and its prob-

ability distribution. That is the probability distribution of the response variable (Y). In the case logistic regression, the distribution of (Y) will be Binomial where (Y) belongs to the exponential family distribution and has the probability density function of the following form;

$$f(y|\theta,\phi) = \exp\left[\frac{y(\theta) - b(\theta)}{\phi} + c(y,\phi)\right] \tag{3.1}$$

The following Table 3.1 shows some of the Generalized Linear Models with their respective distributions;

Table 3.1: **Examples models and their Distributions.**

Models	**Distributions**
Linear Regression	Normal
Logistic Regression	Binomial
Log-linear	Poisson
Poisson regression	Poisson
Multinomial	Multinomial

- Systematic components

  The systematic component specifies the explanatory variable that is used as predictor in the following model;

  $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$ where:

  β_0 is the intercept,

  β_s are the parameter estimators for the explanatory variables x_i,

  $i = (1, 2, \cdots, n)$.

- Link function.

  The Link function connects the expected value or the mean of random component with the systematic component. The general link function can be denoted by $g(\mu)$ and written in the form;

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{3.2}$$

Table 3.2 consists of some link functions together with their random variables and models.

Table 3.2: **Examples models and their link functions.**

Model	Link
Linear Regression	Identity
Logistic Regression	Logit
Log-linear	Log
Poisson regression	Log
Multinomial	Generalized Logit

3.3.3 Exponential family of distribution

According to Nelder and Wedderburn (1972), the probability distribution of random variable (Y) falls within a class of exponential family, if its Probability Mass Function (PMF) or Probability Density function (PDF) can be expressed as;

$$f(y|\theta,\phi) = \exp\left[\frac{y(\theta) - b(\theta)}{\phi} + c(y,\phi)\right], \tag{3.3}$$

Where a random variable (Y) depends on the natural parameters (θ) and the dispersion parameter (ϕ).
$b(\theta)$ is the cumulant function which is utilized in the derivation of the mean and the variance.

By letting μ_i being the expected value of y_i, the GLM can be formed through the introduction of a monotonic link function that connects μ_i with a set of independent variables $(x_1, x_2, \cdots, x_p)$ in the from;

$$g(\mu_i) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} \tag{3.4}$$

For instance, in the case of Bernoulli distribution with the following random variable;

$$Y = \begin{cases} 1, & \text{with } (\pi) \text{ as the probability of success} \\ 0, & \text{with } (1-\pi) \text{ as the probability of failure} \end{cases}$$

Therefore, the PMF is given by;

$$\begin{aligned} f(y,\pi) &= \pi^y(1-\pi)^{1-y} \\ &= \exp\left[y\text{Log}(\pi) + (1-y)\text{Log}(1-\pi)\right] \\ &= \exp\left[y\text{Log}\left(\frac{\pi}{1-\pi}\right) + \text{Log}(1-\pi)\right] \end{aligned} \tag{3.5}$$

Equation 3.3 is expressed in the form of Equation 3.2 or exponential family where;

$\theta = \text{Log}\left(\frac{\pi}{1-\pi}\right)$ for $\pi = \pi(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)}$

$b(\theta) = -\text{Log}(1-\pi(\theta) = \text{Log}(1-\exp(\theta))$

$c(\theta)$ = $\phi = 1$

Note that the first derivative of $b(\theta)$ can be expressed as;

$b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \pi = E(Y)$

And also $b''(\theta) = \frac{\exp(\theta)}{(1+\exp(\theta))^2} = \pi(1-\pi) = Var(Y)$

The below table shows other distributions families that can be expressed as Equation 3.2.

Table 3.3: **Exponential Families Expression of functions** $a(.)$**,** $b(.)$ **and** $c(.)$**.**

Family	$a(\phi)$	$b(\theta)$	$C(y,\phi)$	**Variance**	**Terms**
Gaussian	ϕ	$\frac{\theta^2}{2}$	$\frac{1}{2}\left[\frac{y^2}{\theta} + \text{Log}_e(2\pi\theta)\right]$	σ^2	1
Binomial	$\frac{1}{n}$	$\text{Log}_e 1 + e^\theta$	$\text{Log}_e\binom{n}{ny}$	$\mu(1-\frac{\mu}{n})$	1
Poisson	1	e^θ	$-\text{Log}_e y!$	μ	1
Gamma	ϕ	$-log_e - \theta$	$\phi^{-2}\text{Log}_e\frac{y}{\phi} - \text{Log}_e y - \text{Log}_e\Gamma(\phi^{-1})$	$n\theta^2$	$\frac{1}{\phi}$
Inverse-Gaussian	ϕ	$-\sqrt{2\theta}$	$-\frac{1}{2}\left[\text{Log}_e(\pi\phi y^3) + \frac{1}{\phi y}\right]$	$\frac{\mu}{\theta}$	1

3.3.4 Logistic Regression

A special case of GLM when the response variable has only two categorical outcomes, for instance (Success= 1 and Failure = 0) can be appropriately analyzed by Logistic Regression (LR) (Barber and Thompson, 2004a). In order to apply LR in this study, the variable "own business" is treated as a response variable (Y) having two categories (Yes $= 1$) for an individual who owns a business and (No $= 2$) for an individual who does not own a business. The distribution of the response variable (Y) is used in the study which has the probability of successes/own business(Yes $= 1$) denoted by (π) and the probability of failure/not own business denoted by ($1-\pi$). If we have P number of explanatory variables, it implies that the random variable will follow a binomial distribution with two parameters, (n, π). Since the value of success probability varies with respective to the observation value (x), therefore, $\pi(x)$ can be used instead of just π. The linear probability model for binary logistic regression can then be denoted by;

$$\pi(x) = \alpha + \beta(x) \tag{3.6}$$

where α is the intercept of $\pi(x)$, β is the parameter estimate and x is the independent variable. As shown from Table3.1, the logistic regression model has a Logit-link function written as follows;

$$logit(Y) = \text{Log}\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x \tag{3.7}$$

From Equation 3.6, the link function can be written as the Logit transformation;

$$\pi(x) = \left[\frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}\right]$$

3.3.5 Multiple Logistic Regression

Logistic Regression (LR) model can be further extended to Multiple Logistic Regression (MLR) model when there are more than one explanatory variables to predict a single binary response variable (Agresti, 2018). MLR is then valid for this study since there are more than one independent variables and a single dependent variable "own business" that has dichotomous outcomes which are (2) for 'No' and (1) for 'Yes'.
Consider MLR model that has p predictors of x, $\pi(x) = P(Y = 1)$ can be modeled as follows;

$$Logit\left[\pi(x)\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p \tag{3.8}$$

and $\pi(x)$ can be formulated as a Logit transformation;

$$\pi(x) = \left[\frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_p x_p)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3+\cdots+\beta_p x_p)}\right]$$

The effects of predictors (x_{i_s}) are represented by the parameters (β_i^s) from the Logit that $Y = 1$.

Parameter estimation of Logistic regression

Maximum Likelihood Estimation (MLE) is employed in this study to estimate the parameters. MLE is used to observe the maximum solution from the likelihood function and it can only derive the non-linear equation Silva and Tenreyro (2010).
Equation 3.3 can be used to form a likelihood function as follows ;

$$\begin{aligned} L(y,\theta) &= \prod_{i=1}^{n} \exp\left(\frac{y(\theta) - b(\theta)}{\phi} + c(y,\phi)\right) \\ &= \exp\left[\sum_{i=1}^{n}\left(\frac{y(\theta) - b(\theta)}{\phi} + c(y,\phi)\right)\right], \end{aligned} \tag{3.9}$$

The log of the likelihood function in Equation 3.9 can be expressed as;

$$l(y,\theta) = \log L(y,\theta) = \sum_{i=1}^{n}\left(\frac{y(\theta)-b(\theta)}{\phi} + c(y,\phi)\right), \tag{3.10}$$

Equation 3.10 is differentiated with respect to β_j where; $(j = 0, 1, 2, \cdots, p)$ in order to solve the parameters and set each derivative to zero and the first derivative of $l(y,\theta)$ is called Fisher's score function denoted by;

$$u(\theta) = \frac{\partial \log L(\boldsymbol{y},\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \tag{3.11}$$

From Chain Rule method of differentiation, we will have;

$$\frac{\partial \log L(y,\theta)}{\partial \theta} = \frac{\partial l_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}$$

and all factors, β_{j_s} are obtained using the following system of equations;

$$\frac{\partial l}{\partial \beta_j} = \sum_{i=1}^{n}\left(\frac{y_i-\mu_i}{a_i(\phi)}x_{ij}\right) = 0, \tag{3.12}$$

Newton-Raphson or Fisher's scoring algorithm which are also available in the statistical software packages can be used to solve the system of equations presented in Equation 3.12.

Confidence interval

Various scientific reports have given a great focus on the Confidence Interval (CI) estimation due to its proven power that it has on estimating parameters compared to other hypothesis testing. Standard errors can be expanded around a point estimate to form a confidence interval. Margin of errors are provided by allowing the negative and the positive sings on both sides of the point estimate.

For instance, the CI for β is found by testing the null hypothesis, $H_0 : \beta = \beta_0$, then the interval of β_0 such that Chi-square statistical test is less or equal to $\chi^2 = Z^2_{\frac{\alpha}{2}}$. By using the Wald statistic;

$$\left[\frac{\hat{\beta} - \beta_0}{SE}\right]^2 \leq Z^2_{\frac{\alpha}{2}}, \tag{3.13}$$

Therefore, the CI will be $\hat{\beta} \pm Z_{\frac{\alpha}{2}}$,
Considering the Logistic regression case, by fixing $x = x_0$, the 95% CI of $Logit\pi(x)$ is given by;

$$(\hat{\alpha} + \hat{\beta_{x_0}} \pm 1.96SE), \tag{3.14}$$

Where $SE = \sqrt{Var(\hat{\alpha} + \hat{\beta_{x_0}})} = \sqrt{Var(\alpha) + x_0 Var(\hat{\alpha}) + 2x_0 Cov(\hat{\alpha}; \hat{\beta})}$

Diagnostic, Model Checking and Goodness of Fit

The study will apply model diagnostic to check if the model selected is valid and also fit the data. Model diagnostic is also helpful in identifying the independent variables that are mostly significant.

- **Log-likelihood Goodness of Fit test**
  The Log-likelihood ratio statistic is used in the study to observe or test how good the model is. Log-likelihood ratio statistic is used to compare the null model (model with only intercept) and the final model (model with one or more explanatory variables) in order to observe if the final model has significantly improved to fit data. The hypothesis can be formulated as follows;
  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$ (there is no improvement of the full model over the null model)
  H_1 : At least one of the coefficients $\beta_1, \beta_2, \cdots \beta_p \neq 0$ (There is an improvement of the full model over the null model)

In the above hypothesis, the null (H_0) is rejected under 5% level of significant, if $p - value$ is less than 0.05 and conclude that there is an improvement of a model with one or more explanatory variables over the model with only intercept.

- **Chi-square Goodness of Fit test**

  The Chi-Square Goodness of fit is used to test the consistency of the observed data with the model fitted, for a large sample (n), the Chi-square distribution can be approximated by the Pearson chi-square statistic that was proposed by Karl Pearson in 1900. The Pearson chi-square can be formulated as;

$$\chi^2 = \Sigma \frac{(n_{ij} - u_{ij})}{u_{ij}} \tag{3.15}$$

  The hypothesis for Chi-square test can be formulated as follows;
  $H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$ (There is no significant difference between the observed and the expected value)
  H_0 : At least one of the coefficients $\beta_1, \beta_2, \cdots \beta_p \neq 0$ (There is a significant difference between the observed and the expected value)
  The null hypothesis (H_0) is rejected if $p - value$ is less than 0.05 and conclude that there is a significant difference between the observed and the expected value.

- **Hosmer-Lemeshow**

  The Hosmer-Lemeshow (HL) is also a Goodness of Fit test and have been utilized by several scholars specifically for Logistic Regression (LR) models. HL test if the probability of the predicted values are the same as the observed events of the population subgroups. In HL, the Pearson Chi-square is used to perform comparisons between the expected counts and

the observed counts. For a smaller $p-value$ (≤ 0.05) implies the lack of fit, while the larger $p-value$ means that there is no enough evidence of poor fitting Hosmer Jr et al. (2013). The hypothesis testing for HL Goodness of Fit can be formulated as follows;

$$H_O : E(Y) = \frac{\exp \mathbf{X}'\beta}{1+\exp \mathbf{X}'\beta}$$
$$H_1 : E(Y) \neq \frac{\exp \mathbf{X}'\beta}{1+\exp \mathbf{X}'\beta}$$

Odds Ratios

Odds Ratio (OR) have been largely used in the interpretation and measurements of the association between covariates in the model. For a Logistic regression with single independent variable that has a dichotomous outcomes ($x = 1$ and $x = 0$), then the response with $x = 1$ will have the odds outcome of $\pi_1/1-\pi_1$ while the response with $x = 0$ will have the odds outcome of $\pi_0/1-\pi_0$. Hence, the OR can be presented in the form;

$$OR = \frac{\pi_1/[1-\pi_1]}{\pi_0/[1-\pi_0]},$$

By substituting the Logistic regression model with binary response, the OR can now be formulated as follows;

$$\begin{aligned} OR &= \left[\frac{\exp(\alpha+\beta_1)/\frac{1}{1-\exp(\alpha+\beta_1)}}{\frac{\exp(\alpha)}{1+\exp(\alpha)}/\frac{1}{1+\exp(\alpha)}}\right] \\ &= \left[\frac{\exp(\alpha+\beta_1)}{\exp(\alpha)}\right] \\ &= \exp(\alpha+\beta_1-\alpha) = \exp(\beta_1), \end{aligned} \tag{3.16}$$

The above OR measures the likelihood or unlikelihood of a particular outcome to happen in those with $x = 1$ than those with $x = 0$. The idea of OR can be extended to Multiple Logistic Regression (MLR) where there are more than one explanatory variable which are categorical in nature.

Model selection techniques

Selection of a suitable model becomes more difficult and challenging as the number of independent variables increases since there will be a high interaction between predictors. The main objective in model selection is to select the most complex model that fits the data well and such model should also be easily interpreted without over-fitting the data (Agresti, 2003). There are several model selections that have been used by scholars, however, this study will only focus on Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) also known as Schwarz information criterion. Researchers including Wang and Liu (2006), Liddle (2007), Hansen (2007), Acquah and Carlo (2010), Penny (2012), Lefort et al. (2017), among others preferred the AIC due to its better estimation. AIC checks the closenesses of the fitted model value with the true values and if the fitted model is very close to the true value through the expectations or the mean, then such model can be selected to predict the reality.

Given a maximized Log-likelihood (L) with a model having P parameters, the AIC can be formulated as follows;

$$AIC = -2(L - P) \tag{3.17}$$

The model with a minimum value of AIC is said to be good, and fits better (Lefort et al., 2017). The common alternative of AIC is called BIC which is closely related with AIC. In both AIC and BIC, the chances of selecting the best model become high as the number of independent variables increases. However, AIC tends to out-best BIC when it comes to selecting the overall best model regardless of the sample size (Acquah and Carlo, 2010).

3.3.6 Count Models

Count Models (CMs) are applicable to analyze the count data that shows number of occurrences of any event within a fixed period of time Nelder and Wedderburn (1972). Count variable was frequently applied by researchers from behavioral sciences to count the number of events occurred in a fixed period of time. According to Cameron and Trivedi (2013), the count variable can only cater discrete values that are non-negative (e.g number of children a couple has, number of accidents in an hour, number of goals scored for each match in a season, e.t.c). Poisson regression model is one of the popular count models used by scholars such as Ridout et al. (2001), Famoye and Singh (2006), Silva and Tenreyro (2010), Zou and Donner (2013), Liao et al. (2016), and Chen et al. (2019) among others.

Poisson Regression

The Generalized Linear Model (GLM) with count data approximate a Poisson distribution as a random component. A Poisson distribution has only one mode and it is skewed to the right having only one parameter (μ) which is greater zero (Agresti, 2003). The variance of a response variable $Var(y)$ should be equal to the expected value or the mean of the response variable $E(y)$. Thus, the larger value of $E(y)$ will imply a greater variability in the response variable y. Poisson distribution function can be written as;

$$p(y) = \frac{\exp(-\mu).\mu^y}{y!} \quad for \;\; y = 0, 1, 2, 3, \cdots \tag{3.18}$$

Poisson has natural parameter called Log-mean that has a Log-link function as shown from Table3.2. Poisson log-linear regression is a special case of GLM that has a random variable that is Poisson distributed and have a Log-link function. Hence, the Log-linear model that has P number of explanatory variables (x_i) estimated by parameter (β_i), where $i = 1, 2, 3, \cdots, p$ can be expressed

in the form;

$$Log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p \tag{3.19}$$

The exponential form of mean μ can be written as

$\mu = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_p x_p)$

From Equation 3.18, if $\beta_i > 0$ then, $\exp(\beta_i) > 1$ and $E(Y)$ increases as x_i increases.

Log-Linear Regression Models

Log-Linear Regression(LLR) models is used to model count data that is represented as cell counts in a contingency table. According to Agresti (2018), contingency tables can be predicted by LLR models by specifying the how the magnitude of each frequency cell count depends on the levels of categorical variables. Hence, the data pattern from either two-way or three-way contingency tables can be described by LLR analysis. What makes LLR models unique from other GLMs is that all variables in LLR are treated equally and therefore there is no distinction between independent and dependent variables.

Log-Linear Regression for Two-way contingency tables

Consider a Two-way($a \times b$) contingency tables with r rows and c columns, then a table will have a cell with the observed frequency count (Y_{ij}). A marginal frequency of the i^{th} row and the j^{th} column, respectively will be denoted as;

$Y_{i+} = \sum_{j=1}^{c} Y_{ij}$,

$Y_{+j} = \sum_{i=1}^{r} Y_{ij}$,

And the sample of observations is denoted as;

$n = Y_{++} = \sum_{i=1}^{r} \sum_{j=1}^{c} Y_{ij}$

By letting π_{ij} be the probability that an observation comes from cell i, j, we can also denote π_{+j} and π_{i+} as marginal probabilities while π_{++} is equal to 1. If columns and rows are statistically independent, it means that the ob-

servation probability is equal to the product of the marginal probabilities i.e $\pi(ij) = \pi_{i+} \times \pi_{+j}$. The expected value of cell i, j can be further denoted by $\mu_{ij} = E(Y_{ij}) = n\pi_{ij}$. Both the column and row marginal expected values are expressed as follows;

$$\mu_{i+} = \sum_{i=1}^{r} n\pi_{ij} = n\pi_{i+}, \tag{3.20}$$

$$\mu_{+j} = \sum_{j=1}^{c} n\pi_{ij} = n\pi_{+j}, \tag{3.21}$$

where μ_{ij} can be written as;

$$\mu_{ij} = (\pi_{i+}) \times (\frac{\pi_{+j}}{n}), \tag{3.22}$$

By taking the Logarithm from both sides of the Equation 3.21, we then have;

$$\eta_{ij} = log_e(\mu_{ij}) = log_e(\mu_{i+}) + log_e(\mu_{+j}) - log_e(n), \tag{3.23}$$

Equation 3.23 can be re-parameterized to form a Log-linear model for the two-way statistical independent in the form;

$$\eta_{ij} = \mu + \alpha_i + \beta_j, \tag{3.24}$$

where;
$\sum \alpha_i = 0$ and $\sum \beta_j = 0$
The parameters from Equation 3.24 can be solved as follows;

$$\frac{\eta_{++}}{rc}$$
$$\alpha_i = \frac{\eta_{i+}}{c} - \mu$$
$$\beta_j = \frac{\eta_{+j}}{r} - \mu$$

In other cases where variables are not statistically independent, the Log-linear

model will have an additional parameter (γ_{ij}) for the interaction effects and can be written as;

$$\eta_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}, \tag{3.25}$$

Where α_i is the row effect, β_j is the column effect and γ_{ij} is the interaction effect for column and row variables.

Log-Linear Regression for Three-way contingency tables

The Three-way contingency table with three variables (A,B,C) will have observation cell counts (Y_{ijk}) and the Log-linear model can be expressed as;

$$\eta_{ijk} = \mu + \alpha_{A(i)} + \alpha_{B(j)} + \alpha_{C(k)} + \alpha_{AB(ij)} + \alpha_{BC(jk)} + \alpha_{AC(ik)} + \alpha_{ABC(ijk)} \tag{3.26}$$

Where; $\sum \alpha_{1(+)} = \sum \alpha_{12(i+)} = \sum \alpha_{123(ij+)} = 0$
Parameters of the model in Equation 3.25 can also be solved as follows;

$$\mu = \frac{\eta_{+++}}{abc}$$
$$\alpha_{A(i)} = \frac{\eta_{i++}}{bc} - \mu$$
$$\alpha_{B(j)} = \frac{\eta_{+j+}}{ac} - \mu \; \alpha_{AB(ij)} = \frac{\eta_{ij+}}{c} - \mu - \alpha_{A(j)} - \alpha_{B(j)}$$
$$\alpha_{ABC(ijk)} = \eta_{ijk} - \mu - \alpha_{A(i)} - \alpha_{B(j)} - \alpha_{C(k)} + \alpha_{AB(ij)} - \alpha_{BC(jk)} - \alpha_{AC(ik)}$$

Inference for Log-Linear regression

Chi-square Goodness-of-fit is used to test how good the model fits by finding the significant difference between the observed values and the expected values (Barber and Thompson, 2004a). The following are the Deviance (G^2) and the Chi-square(χ^2) test for the three-way contingency tables.

$$G^2 = 2 \sum n_{ijk} log \left(\frac{n_{ijk}}{\hat{\mu}_{ijk}} \right) \tag{3.27}$$

$$\chi^2 = \Sigma \frac{(n_{ijk} - \hat{\mu}_{ijk})}{\hat{\mu}_{ijk}} \tag{3.28}$$

The degrees of freedom(DF) for G^2 test which is the Deviance of model is found by subtracting the number of parameter from the number of cell counts. Hence, the more the model becomes complex, the lower the DF.

Over/Under Dispersion

There are several cases where the assumptions of Poisson regression are not met. One of the common cases is when the mean of a response variable y is greater that the variance of the response outcome, that is called Over-dispersion. The other case is when the variance is now greater that the mean of y, and this one is called Under-dispersion (Chen et al., 2019). Over-dispersion causes the the estimates of the standard errors to be lower and that can also results in the overestimation of the significance. The above mentioned cases are tested by Pearson Chi-square Dispersion (PCD) statistic in SAS, if $PCD = 1$, it implies that the variance is equal to the mean and that is called equi-dispersion. Hence, neither over nor under dispersion exist and Poisson regression will fit better. If the PCD statistic is not equal to one, the Negative Binomial can fit better since there is neither over dispersion nor under dispersion (Atkins et al., 2013)

Negative Binomial regression

The Negative Binomial (NB) regression is also a count model that approximate the negative binomial distribution having two parameter (k, μ), given by;

$$f(y; k, \mu) = \frac{\Gamma(y+k)}{\Gamma(k).\Gamma(y+1)} \cdot \left[\frac{k}{\mu+k}\right]^k . \left[1 - \frac{k}{\mu+k}\right]^y \quad for\ y = 0, 1, 2, \cdots \tag{3.29}$$

The link function NB regression is given by: $g(\mu) = log(\mu)$
The systematic component for NB regression can expressed as;
$g(\mu) = log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

Which implies that $\mu = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

From Equation 3.12, $E(y) = \mu$ and $Var(y) = \mu + \frac{\mu^2}{k}$ where $\frac{1}{k}$ is called the dispersion parameter, as $\frac{1}{k}$ approaches 0, the $Var(y)$ approaches μ. Negative Binomial (NB) regression is utilized as the alternative method of Poisson regression if over/under dispersion exists. Over/under dispersion occurs when there is higher or lower variability in the data set than it is expected under the assumed distribution (Barber and Thompson, 2004b). NB has a variance which is a function of mean and has another parameter which is called dispersion parameter (Atkins et al., 2013). The variance converges to the same values as the mean when the dispersion parameter gets larger and the NB turns to Poisson.

Parameter Estimation for count models

The Maximum Likelihood estimation (MLE) was used to estimate coefficients in both Poisson and Negative binomial regression models. The study started with the MLE for Poisson regression by finding the following likelihood function and its logarithm ;

$$ln[L(\boldsymbol{Y}, \boldsymbol{\beta})] = \sum_{i=1}^{p} - \exp(\boldsymbol{X_i \beta}) - ln y_i! \tag{3.30}$$

The next step from Equation 3.30 is to take the derivative of the Logarithm with respect to each coefficient and equate the results to zero, thus; $\sum_{i=1}^{p}[y_i - \exp(\boldsymbol{x_i \beta})]^{\mathbf{x_i}} = 0$. There will not be any closed form of solution due to the non-linear equation that will results from the derivative function. The regression coefficient set that will maximize the log-function must be derived from the iteration algorithm method. The distribution of MLE is said to be multivariate normal with $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\beta V_{\hat{\beta}}})$, where $\boldsymbol{V_{\hat{\beta}}} = (\sum_{i=1}^{p} \mu_i x_i x_i')^{-1}$.

Model section techniques for count models

From both Poisson and Negative binomial regression, AKaike's Information Criterion (AIC) was used to compare models that are related and can also measure how good the model managed to predict the data (Nelder and Wedderburn, 1972). A model with a minimum AIC among others is regarded as a good model. The other alternate method of model selection that is commonly used is the Schwartz Criterion (SC) known as Bayesian Information Criterion (BIC).

Count model Diagnostics

Pearson Chi-square tests as well as the Deviance statistic tests are used to measure the overall performance of the count models. Pearson Chi-square statistics test can be expressed as follows;

$$P = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{3.31}$$

While the Deviance or G-statistic can be formulated as;

$$D = \sum_{i=1}^{n} \left[y_i ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \tag{3.32}$$

Pearson Chi-square and Deviance statistic test are approximately Chi-square distributed with the degrees of freedom $(n-k)$. The rejection of the test implies that there is lack-of-fit and failing to reject test means that there is no evidence of lack-of-fit (Agresti, 2018).

After fitting the count models, Pearson residuals are utilized in the study to test for the Goodness-of-fit. Pearson residuals are computed as $\chi^2 = \sum_{i=1}^{k} r_i^2$, where;

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{y}_i}} = \frac{(observed - expected\ counts)}{\sqrt{expected\ counts}} \tag{3.33}$$

3.4 Conclusion

The purpose of utilizing MLR in this study is to model the response variable against several independent variables. The selection criteria is used to deduce the final model that fits the data and with the smallest AIC or BIC value. Maximum Likelihood Estimator(MLE) is used to estimate the parameter of the explanatory variables and to identify the significant factors associated with the response variable. The most significant factors identified in MLR will be considered in Log-Linear Regression (LLR) analysis to investigate if interaction effects exist between the factors and also identify the significant interaction.

Chapter 4

Results

4.1 Introduction

This chapter consists of two statistical methodologies under Generalized Linear Models (GLMs) namely; Multiple Logistic Regression(MLR) and the Log-Linear Regression (LLR) which belongs to count models. The descriptive statistics, Logistic and Log-linear regression model results are displayed and interpreted from Statistical Analysis Software Version 9.2 (SAS) outputs supported by R-studio statistical software. The datasets used in the analysis includes 2017 Quarterly Labour Force Survey (QLFS) and 2017 Survey of Employers and the Self-employed (SESE), all obtained from Statistic South Africa (StatsSA). MLR was utilised on QLFS while LLR was used on both QLFS and SESE datasets.

4.2 Insights of data

The QLFS is a sample survey conducted within South African households. The survey is conducted every quarter from South African individuals between the

age of 15 and 64 and the survey is based on labour market activities. The SESE is a survey also conducted by StatsSA in the informal sector at the end of Third quarter, each year. The rationale behind SESE focusing on the South African informal sector is that the sector has the largest economic activities in the country. During SESE, business owners from QLFS are identified and followed up in the form of interviews to investigate more about their business characteristics which include the financial state of their businesses as it is one of the interests of the study.

4.3 Descriptive statistics using QLFS

The Cross tabulation, summary statistics, pie charts and testing for association between variables were performed under the descriptive statistics to describe the relationship between two categorical variables using SAS FREQ PROC. The results for other results are displayed in the Appendices section.
In this study, the QLFS data that was used consists of four quarters of the year 2017 with the population of $39,421$. The data contained the following variable; Own business, gender, age group, population group, province, geographical area, education status, attended school, paid work and marital status.

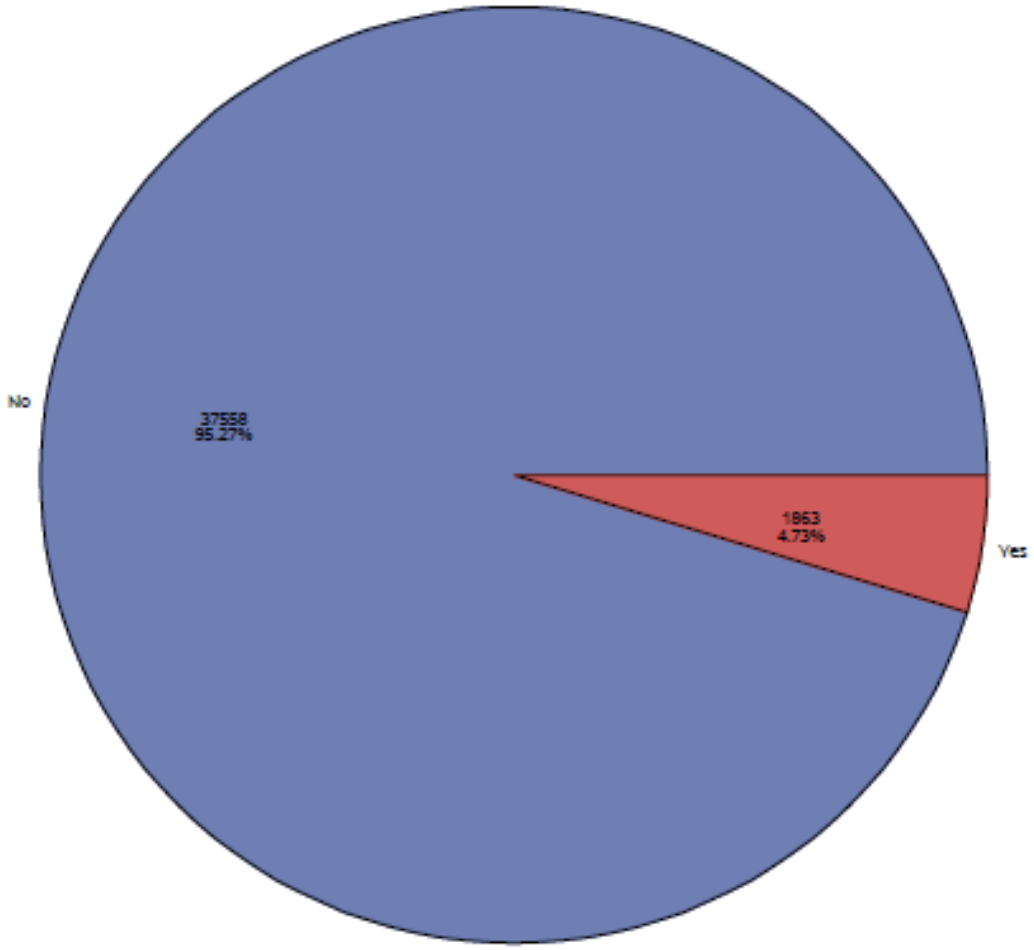

Figure 4.1: A Pie chart depicting business ownership distribution from the data.

Figure 4.1 above shows that the number of business owners were found to be far less than non-business owners. Only $1,863$ participant own business, that is less than 5% of the total survey while about 95% do not own businesses.

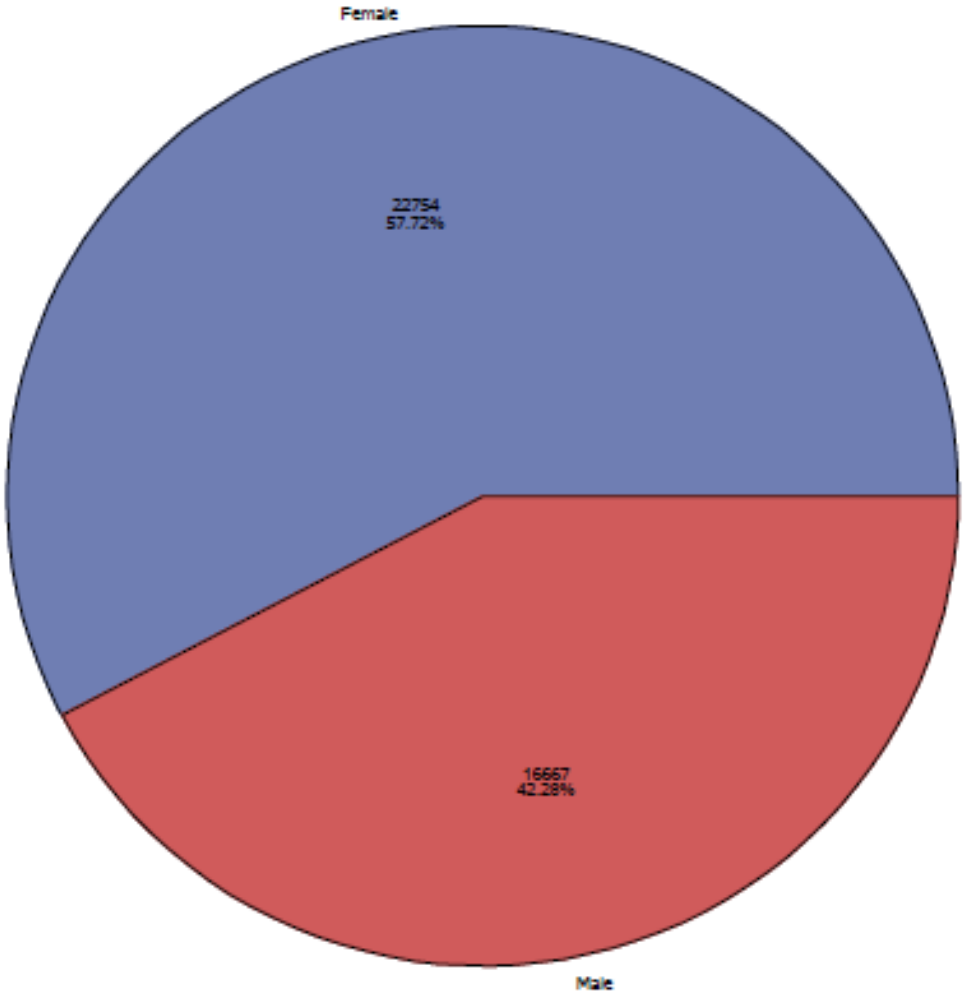

Figure 4.2: A Pie chart depicting gender distributed from the data.

Figure 4.2 shows that the female participants were more than males participants with about 57% while males contributed 42% to the survey.

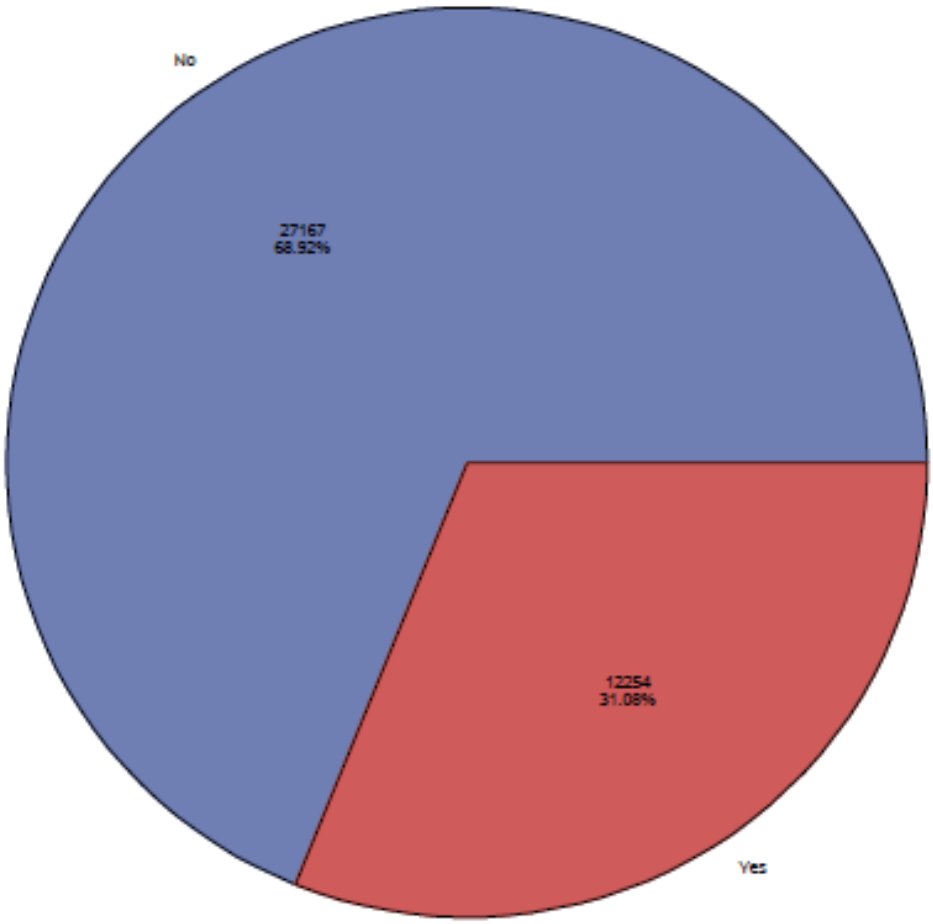

Figure 4.3: A Pie chart depicting paid work distribution from the data.

After the participants being asked whether they are paid workers or not, the above Figure 4.3 shows that about 69% of were not paid workers while 31% were paid workers. Hence, there were less paid workers compared to paid workers.

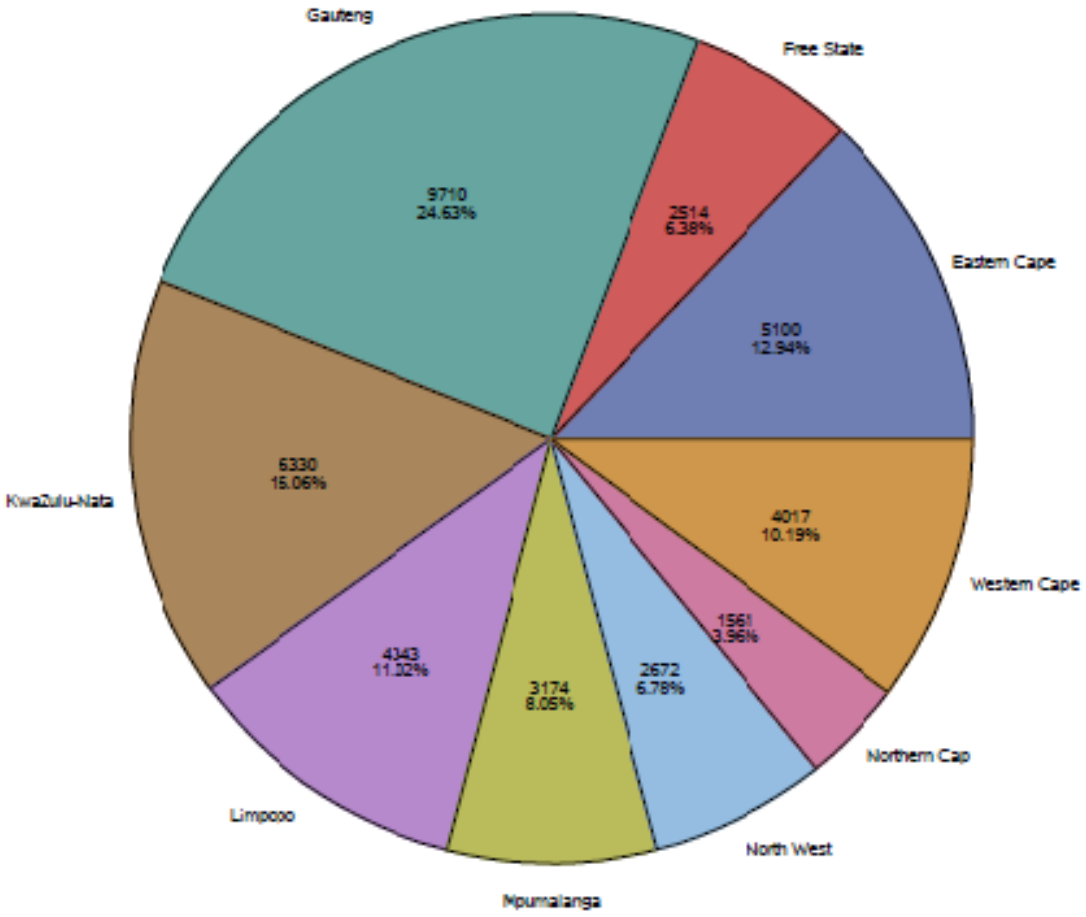

Figure 4.4: A Pie chart depicting province distribution from the data.

Out of nine South African provinces, Figure 4.4 above shows that about 25% of participants were from Gauteng province and had the highest proportion compared to the rest of the provinces. Gauteng was followed by Kwa-Zulu Natal with about 16% while the two provinces which had the lowest proportion were Northern Cape and Free State with about 4% and 6.38%, respectively.

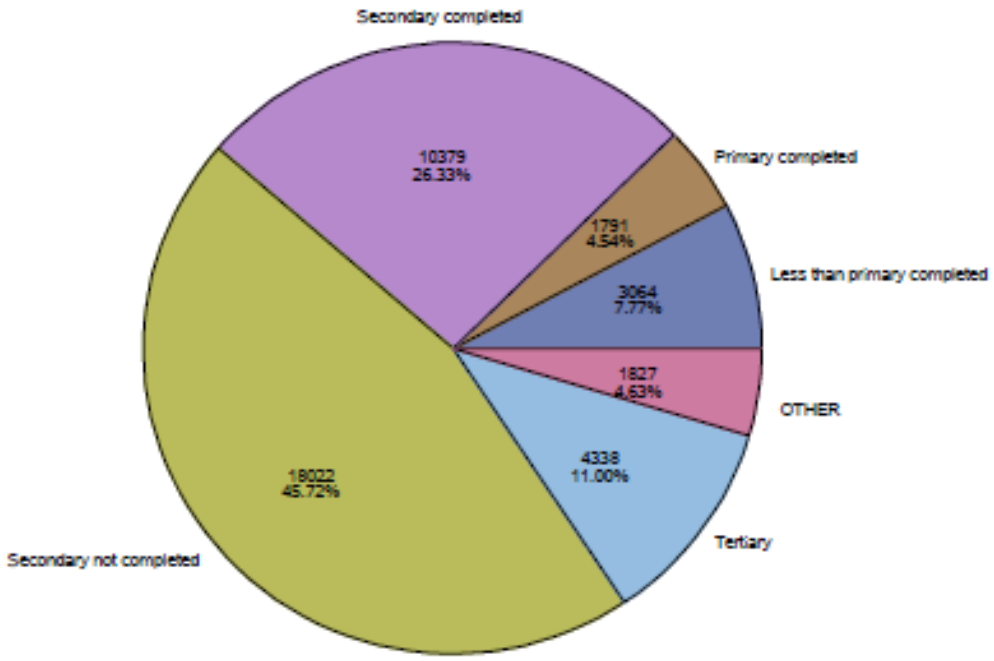

Figure 4.5: A Pie chart depicting education status distribution from the data.

The above pie chart in Figure 4.5 presents the highest proportion of participants who did not complete secondary education with about 45% followed by those who completed secondary education with 26% of the survey. The category with lowest proportion was for the participant who completed primary education.

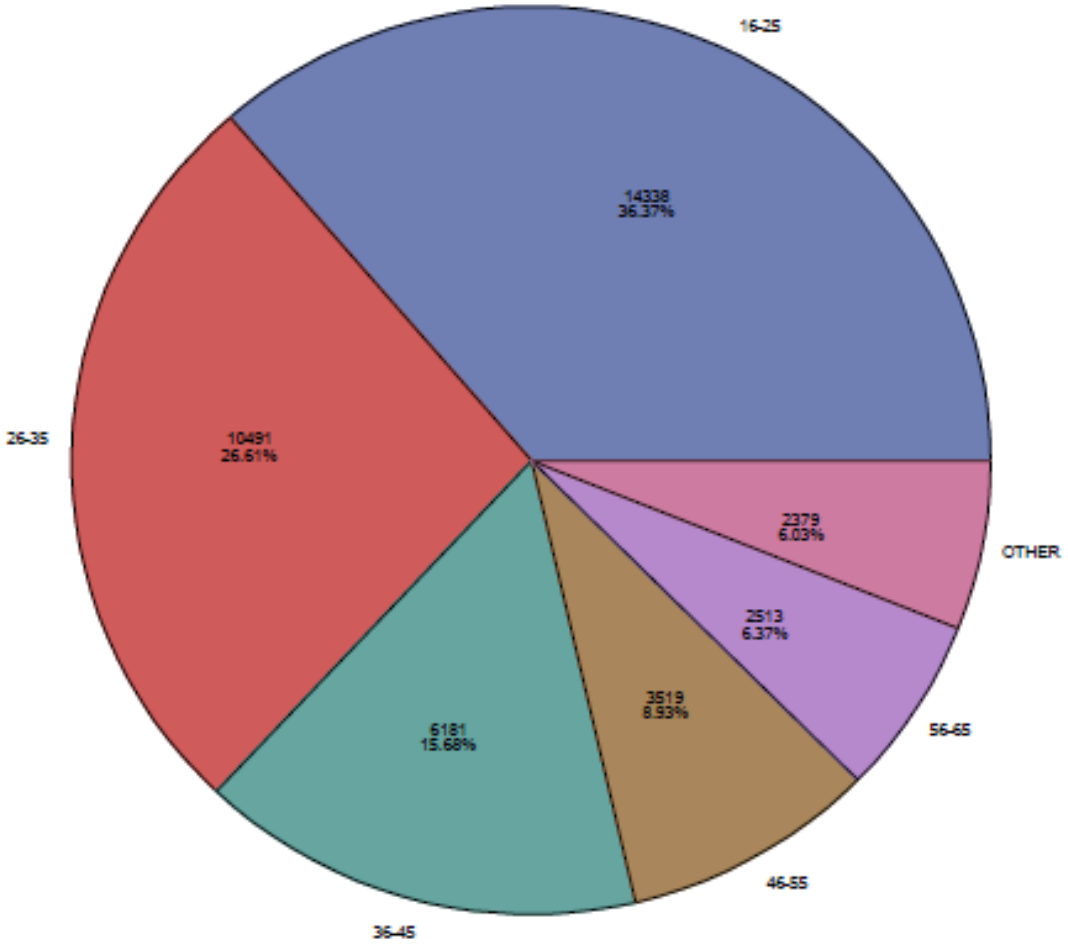

Figure 4.6: A Pie chart depicting age group distribution from the data.

Most participants according to Figure 4.6 from the survey were between the ages of 16 and 25 with about 36%, followed by the age group of $26-35$ with 26.61%. The lowest proportion were for participants who are either over the age of 65 or bellow 16.

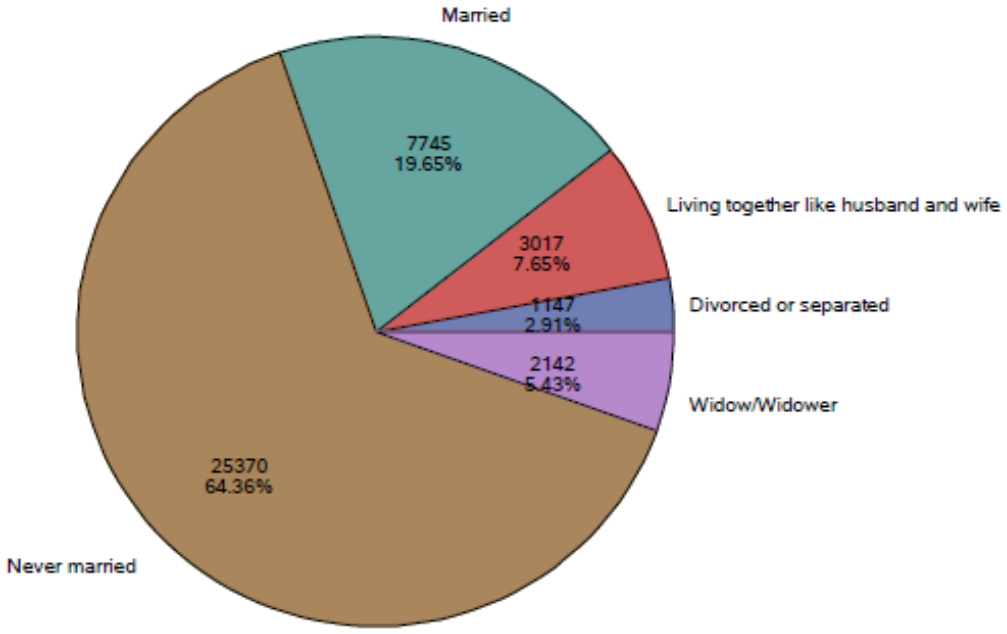

Figure 4.7: A Pie chart depicting marital status distribution from the data.

From the above Figure 4.7 chart, majority of participants were never married having the highest proportion of about 64% followed by the married category with 19.65%. Participants who have divorced or separated have the lowest participation in the survey with the value 2.91%.

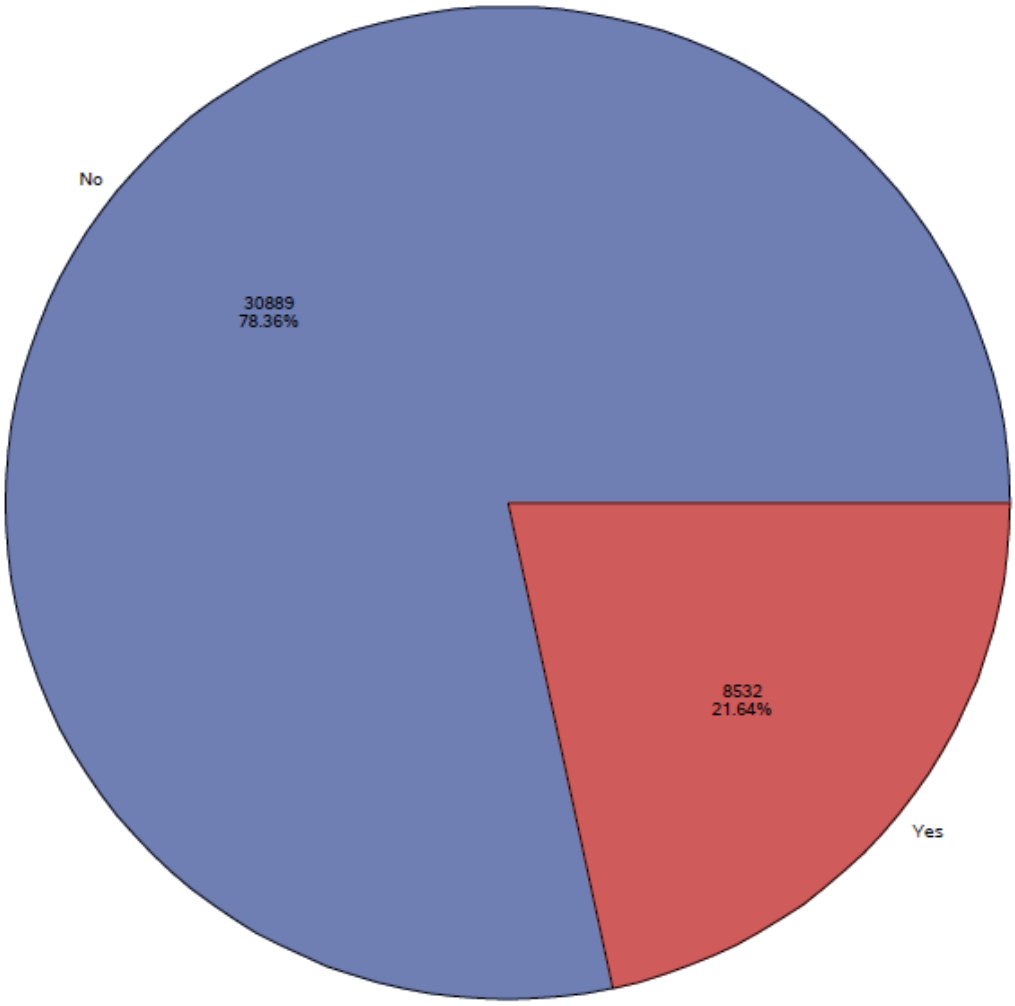

Figure 4.8: A Pie chart depicting the number of participants who attended school.

Figure 4.8 shows about 22% of participants attended a formal school compared to 78% which are those who did not attended school.

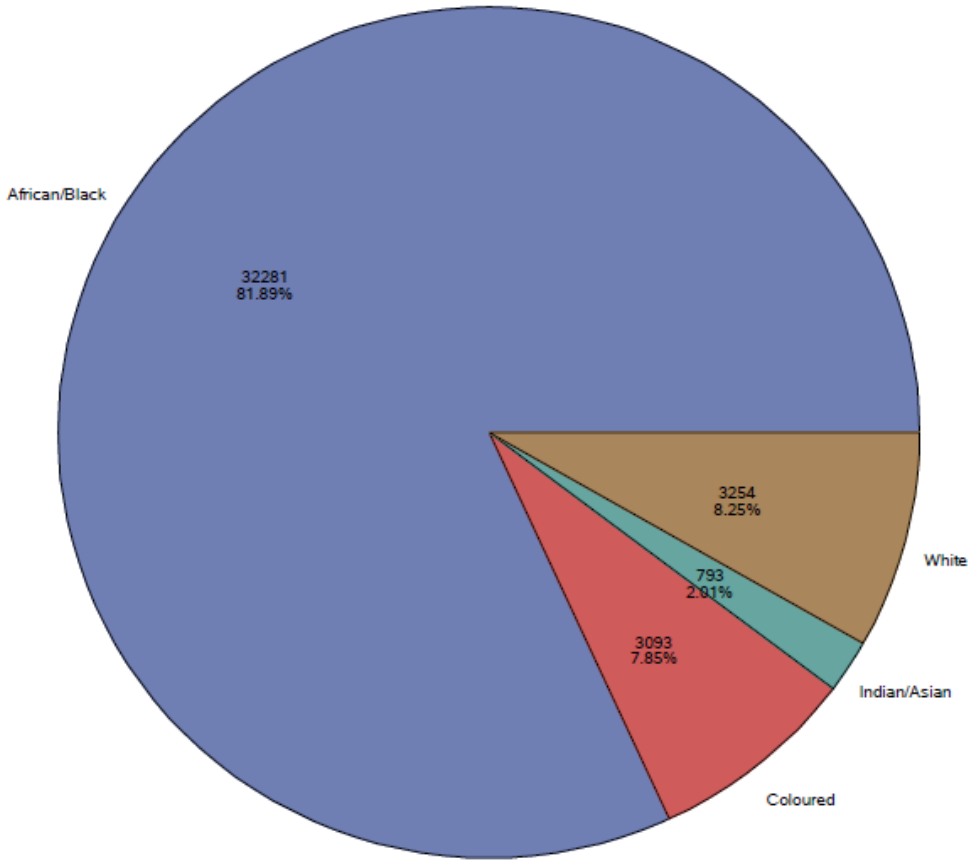

Figure 4.9: A Pie chart depicting population group distribution from the data.

African black participants largely dominated with the proportion of about 82 followed by Whites, Coloured and Indian or Asian with 8.25%, 7.85% and 2.01%, respectively.

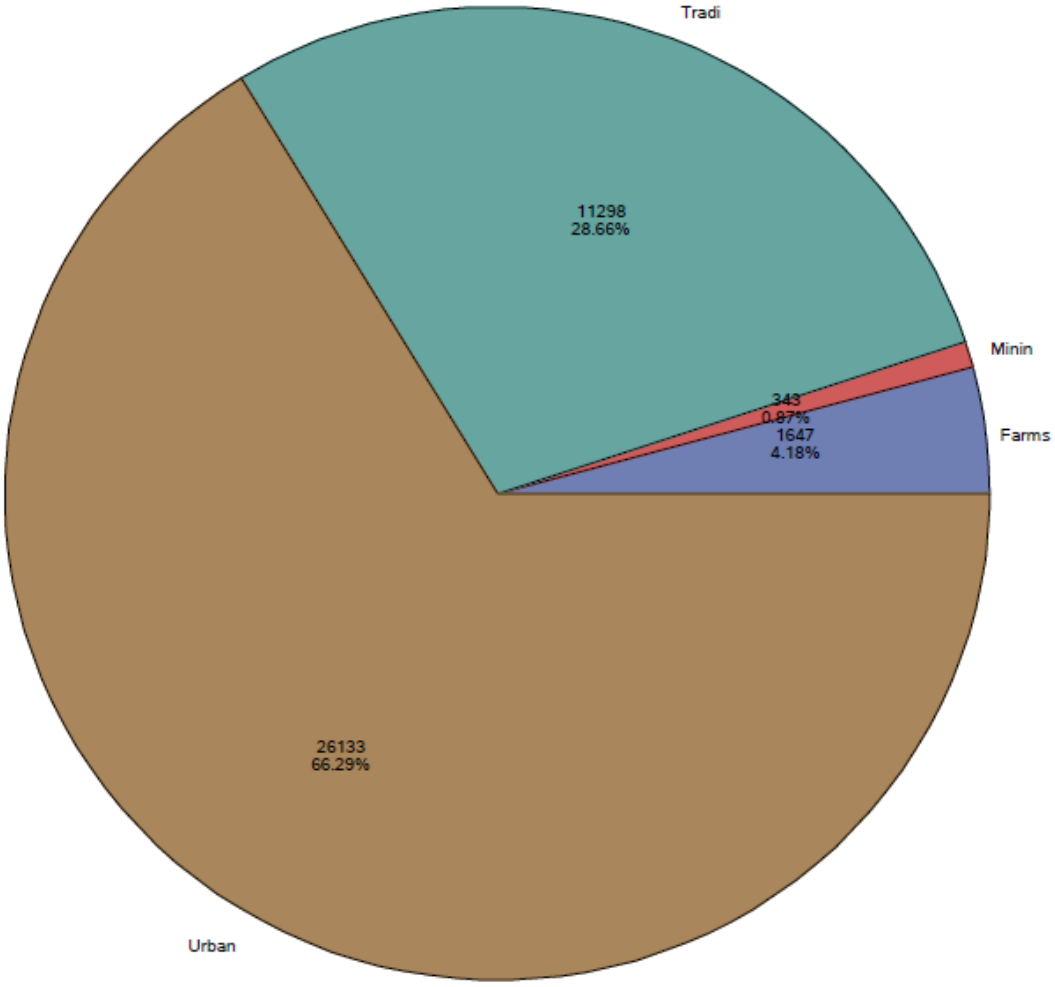

Figure 4.10: A Pie chart depicting geographical type distribution from the data.

In terms of the geographical type of location where of participants live, the above pie chart shows that the four categories were highly dominated by Urban area participants with 66.29%. The second highest category were those who come from traditional communities or areas with 28.66% followed by participants from farm areas (4.18%) and the least proportion was those coming from mining areas with less than a percent.

4.3.1 Association among variables

The study utilised the SAS PROC FREQ to compute and display the cross-tables and the statistic tests for the binary variable ”own business” against other categorical variables namely; gender, age group, population group, province, marital status, educational status, attended school, geographical type and paid work. The Chi-square test of independence from SAS CHISQ function was used to test if there is a significant association amongst the variables.

The formulated null hypothesis for statistical association test is that, there is no association between two or more categorical variables and alternative states that the significant association exists. If the P-value of the Chi-square statistic is less than the significant Alpha (set to be 0.05), the null hypothesis is rejected under the 5% level of significance. The rejection of null hypothesis indicates the statistical significant probability of association between the selected the categorical variables. For the purpose of the study, all Chi-square tests including, Likelihood ratio and Mantel-Haenszel were performed.

Table 4.1: Contingency table for business ownership versus gender.

Own Business	**Gender**		
	Male	**Female**	**Total**
Yes	1027	836	1863
No	15640	21918	37558
Total	16667.00 42.28	22754 57.72	39421 100.00

Table 4.2: Statistical test for “own business” versus “gender”.

Statistic	**DF**	**Value**	**P-value**
Chi-Square	1	132.2384	< 0.0001
Likelihood Ratio Chi-Square	1	130.2912	< 0.0001
Continuity Adj. Chi-Square	1	131.6864	< 0.0001
Mantel-Haenszel Chi-Square	1	132.2350	< 0.0001
Phi Coefficient		0.0579	
Contingency Coefficient		0.0578	
Cramer’s V		0.0579	

Table4.1 display the contingency table to outline the distribution of the variable “own business” that has two responses (yes and no) against two categories of gender (male and female). According to the results, there are more males owning business owners than females.

Table4.2 shows the Chi-square test of association between “own business” and “gender”, and The P-values for all Chi-square statistic tests are less 0.05, which implies the rejection of the null hypothesis under 5% level of significance. Hence, there is a statistical significant association between variables “own business” and “gender”. The Phi Coefficient, Contingency Coefficient and Cramer’s V vales are close to zero which implies a negligible positive relationship between own business variable and gender.

Table 4.3: Odds ratio and Relative risk for “own business” versus “gender”.

Statistic	**Value**		95% CI
Odds Ratio	1.7216	1.5678	1.8905
Relative Risk (Column 1)	1.3238	1.2685	1.3815
Relative Risk (Column 2)	0.7689	0.7307	0.8092

Table4.3 depicts results of Odds Ratio (OR) and Relative Risk (RR) for variable "Gender" on business ownership. From the table, OR $= 1.7216$ means that males have 72.16% more chances of owning business than the Female counterpart. The confidence interval limits for the OR is $1.5678 - 1.8905$ and the estimated OR falls between the interval. Hence, there is a 95% confidence that the true OR falls between the interval. From the table, Relative Risk Column 1 (RRC1) is for Male while the Relative Risk Column 2 (RRC2) is form Female. RRC1 $= 1.3238$ shows that Males have 32.38% more risks of owning business than their females counterpart. RRC2 $= 0.7689$ indicates that there is 76.89% less Risks of owning business than males.

Table 4.4: Contingency table for business ownership versus population group.

Own business	Population group				
	African	**Coloured**	**Indian**	**White**	**Total**
Yes	1529	44	38	252	1863
No	30752	3049	755	3002	37558
Total	32281	3093	793	3254	39421
	81.89	7.85	2.01	8.25	100.00

Table 4.5: Statistical test for "own business" versus "Population group".

Statistic	**DF**	**Value**	**P-value**
Chi-Square	3	140.8188	< 0.0001
Likelihood Ratio Chi-Square	3	157.8426	< 0.0001
Mantel-Haenszel	1	27.4886	< 0.0001
Phi Coefficient		0.0598	
Contingency Coefficient		0.0597	
Cramer's V		0.0598	

Table4.4 contingency table shows that there are more Africans owning busi-

nesses than other population groups, followed by Whites, and the least population group is Indians category. Table4.5 shows the Chi-square association test between categorical variables, “Own business” and “Population group”. The P-values of Chi-Square, Likelihood Ratio Chi-Square and Mantel-Haenszel are less than 0.05, which implies that the null hypothesis is rejected under 5% level of significance. Hence, there is a statistical significant association between “own business” and “population group”. The values of The Phi Coefficient, Contingency Coefficient and Cramer’s V are close to zero, implying a weak positive relationship between “own business” and “population group”.

Table 4.6: Contingency table for business ownership versus marital status.

Own business	Marital status					
	Married	**Living Together**	**Widow/Widower**	**Divorced**	Never married	**Total**
Yes	561	159	112	112	919	1863
No	7184	2858	2030	1035	24451	37558
Total	7745	3017	2142	1147	25370	39421
	19.65	7.65	5.43	2.91	64.36	100.00

Table 4.7: Statistical test for “own business” Versus “marital status”.

Statistic	**DF**	**Value**	**P-value**
Chi-Square	4	245.4946	$<$ 0.0001
Likelihood Ratio Chi-Square	4	221.9136	$<$ 0.0001
Mantel-Haenszel Chi-Square	1	169.1688	$<$ 0.0001
Phi Coefficient		0.0789	
Contingency Coefficient		0.0787	
Cramer’s V		0.0789	

According to the results in Table4.6, people who never got married are dominating the number of business owners than other statuses and the divorced people have the least number of business owners. The results in Table4.7, the

Chi-square statistical association was tested between the categorical variables "Own business" and "marital status". The P-value for all Ch-square test performed were also found to be less than 0.05 which implies an evidence of statistical association between business ownership and marital status. The Phi Coefficient, Contingency Coefficient and Cramer's V values are all about 0.078, that is close to zero, implying a weak relationship between business ownership and marital status.

Table 4.8: Contingency table for business ownership versus age group.

Own business	**age group**							
	16-25	**26-35**	**36-45**	**46-55**	**56-65**	**66-75**	**over 75**	**Total**
Yes	155	594	546	329	185	42	12	1863
No	14183	9897	5635	3190	2328	1528	797	37558
Total	14338	10491	6181	3519	2513	1570	809	39421

Table 4.9: Statistical test for "own business" versus "Age group".

Statistic	**DF**	**Value**	**P-value**
Chi-Square	6	914.4790	< 0.0001
Likelihood Ratio Chi-Square	6	1021.0014	< 0.0001
Mantel-Haenszel Chi-Square	1	251.9917	< 0.0001
Phi Coefficient		0.1523	
Contingency Coefficient		0.1506	
Cramer's V		0.1523	

The age groups (26-35) and (36-45) were dominating other age groups with number of business owners above 500 according to Table4.8, and the minority of business owners were of age group over 70. Table4.9 shows an evident of statistical significant association between "own business" and "age group" as the P-value of all the Chi-Square test used are less than 0.05, therefore, the

null hypothesis is rejected under 5% level of significant. The values of Phi Coefficient, Contingency Coefficient and Cramer's V are all between 0.1 and 0.4, which indicates a slightly weak positive relationship between "Own business" and "Age group".

Table 4.10: Contingency table for business ownership versus education status.

Own business	Education Status							
	NS	**LPC**	**PC**	**LSC**	**SC**	**T**	**O**	**Total**
Yes	75	200	90	696	450	325	27	1863
No	1425	2864	1701	17326	9929	4013	300	37558
Total	1500 3.81	3064 7.77	1791 4.54	18022 45.72	10379 26.33	4338 11.00	327 0.83	39421 100.00

Note: NS : No Schooling, LPC: Less than primary completed, PC: Primary Completed, LSC: Less than Secondary Completed, SC:Secondary Completed, T:Tertiary,O:Others.

Table 4.11: Statistical test for "own business" versus "education status".

Statistic	**DF**	**Value**	**P-Value**
Chi-Square	6	138.8457	< 0.0001
Likelihood Ratio Chi-Square	6	126.3291	< 0.0001
Mantel-Haenszel Chi-Square	1	4.3995	0.0359
Phi Coefficient		0.0593	
Contingency Coefficient		0.0592	
Cramer's V		0.0593	

Table4.10 shows that majority of business owners have less than secondary education level, closely followed by those who completed secondary and those with tertiary education. Table4.11 presents association between the categorical variables "own Business" and "education status" was tested using three Chi-square statistics. The P-values for all tests are less than 0.05, which implies that the association is statistically significant. The values of Phi Coefficient, Contingency Coefficient and Cramer's V are about 0.059 which implies a weak positive relationship between "own Business" and "education status".

Table 4.12: Contingency table for Business ownership versus province.

Own business	Province									
	WC	**EC**	**NC**	**FS**	**KZN**	**NW**	**GP**	**MP**	**L**	**Total**
Yes	175	177	29	116	259	103	568	171	265	1863
No	3842	4923	1532	2398	6071	2569	9142	3003	4078	37558
Total	4017 10.19	5100 12.94	1561 3.96	2514 6.38	6330 16.06	2672 6.78	9710 24.63	3174 8.05	4343 11.02	39421 100.00

Note: L : Limpopo, GP: Gauteng, FS: Free State, NC: Northern Cape, WC:Western Cape, NW:North West, EC:Eastern Cape, KZN: Kwa-Zulu Natal, MP: Mpumalanga.

Table 4.13: Statistical test for “own business” versus “province”.

Statistic	**DF**	**Value**	**P-value**
Chi-Square	8	106.3923	< 0.0001
Likelihood Ratio Chi-Square	8	113.4790	< 0.0001
Mantel-Haenszel Chi-Square	1	61.5974	< 0.0001
Phi Coefficient		0.0520	
Contingency Coefficient		0.0519	
Cramer’s V		0.0520	

The province with the majority of business owners was found to be Gauteng with above 500, while Northern Cape province had the least number of business owners than the all the provinces in South Africa. Table4.13 shows the P-values of all the Chi-square tests less than 0.05, indicating a statistical significant association between the two categorical variables, ”Own business” and ”Province”. The values of Phi Coefficient, Contingency Coefficient and Cramer’s V are about 0.059 which implies a weak positive relationship between “own pusiness” and “province”.

4.4 Application of Multiple Logistic Regression using QLFS

The SAS 9.4 PROC LOGISTIC system was utilised on QLFS 2017 dataset to perform the model selection criterion, for testing the global null hypothesis, Goodness-of-fit tests, and estimates of the predictor variables using the Maximum Likelihood estimation (MLE).The considered model has the intercept, binary response variable "own business" with the explanatory variables namely; gender, age group, population group, province, geographical area, education and marital status.

4.4.1 Model selection

The backward elimination method was performed to arrive in the final model that has the lowest Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) known as Schwarz Criterion (SC). Table4.13 shows the values of $AIC = 15010.7$ and $AIC = 13328.789$ for the intercept only and the intercepts with covariates, respectively. The value of $SC = 15019$ for the intercept only $SC = 13611.997$ for the intercept and the covariates.

Table 4.14: Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	15010.700	13328.789
SC	15019.282	13611.997
-2 Log L	15008.700	13262.789

Table 4.15: Testing for the global null hypothesis.

Test	**Chi-Square**	**DF**	**P-value**
Likelihood Ratio	1745.9106	32	< 0.0001
Score	1485.4950	32	< 0.0001
Wald	1068.9468	32	< 0.0001

4.4.2 Model Diagnostic and Goodness-Of-Fit results

Table 4.16: Deviance and Pearson Goodness-of-Fit Statistics.

Criterion	**Value**	**DF**	**Value**	**P-value**
Deviance	3685.2592	6382	0.5774	1.0000
Pearson	7971.8260	6382	1.2491	< 0.0001

Table 4.17: Lemeshow Goodness-of-Fit test

Chi_square	**DF**	**P-value**
5.65900	8.00	0.69

Table4.15 shows the Likelihood ratio, score and the Wald test with the P-value less than 0.05. Hence, the null hypothesis is rejected under 5% level of significant, which implies that there is an improvement of a model with all explanatory variables over the model with only intercept. Pearson statistics and the Hosmer and Lemeshow Goodness-of-Fit test results are shown on Table4.16 . The P-value of Deviance statistic and the P-value for Hosmer and Lemeshow is 1.000 and 0.69, respectively, which are all greater than 0.05. Hence there is no enough evidence of poor fitting within the model.

4.4.3 Maximum Likelihood Results

The backward elimination method was used to come up with the final model and the results of the final model is displayed in Table4.18 to Table4.20. Cate-

gories of selected variables are shown in the table together with the estimates, standard errors, Wald Chi-square values as well as the P-values of each category.

Table 4.18: Effects Analysis.

Effect	**DF**	Wald Chi-Square	**P-value**
Gender	1	211.7921	< 0.0001
Marital_status	4	33.5873	< 0.0001
Geo_type	3	49.6051	< 0.0001
Province	8	78.4123	< 0.0001
Population	3	95.3887	< 0.0001
Education	6	26.6387	0.0002
Age_group	6	306.4777	< 0.0001
Attended	1	108.9961	< 0.0001

Table 4.19: a) Maximum of Likelihood estimates.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	P-value
Intercept		1	-4.7173	0.1626	842.1420	$<$ 0.0001•
Gender	**Female**	1	-0.3877	0.0266	211.7921	$<$ 0.0001•
Marital_status	**Divorced**	1	0.0809	0.0889	0.8283	0.3628
Marital_status	**Living together**	1	-0.0300	0.0773	0.1510	0.6976
Marital_status	**Married**	1	0.0640	0.0509	1.5795	0.2088
Marital_status	**Never married**	1	-0.2704	0.0509	28.2038	$<$ 0.0001
Geo_type	**Farms**	1	-0.5614	0.1484	14.3155	0.0002•
Geo_type	**Mining**	1	-0.4638	0.2484	3.4855	0.0619
Geo_type	**Tradition**	1	0.5224	0.0999	27.3317	$<$ 0.0001
Province	**Eastern Cape**	1	-0.2654	0.0771	11.8533	0.0006•
Province	**Free State**	1	0.0684	0.0935	0.5352	0.4644
Province	**Gauteng**	1	0.1960	0.0555	12.4547	0.0004•
Province	**KwaZulu-Nata**	1	0.0266	0.0677	0.1548	0.6940
Province	**Limpopo**	1	0.4154	0.0716	33.7041	$<$ 0.0001•
Province	**Mpumalanga**	1	0.2458	0.0791	9.6617	0.0019•
Province	**North West**	1	-0.2023	0.0971	4.3420	0.0372•
Province	**Northern Cap**	1	-0.6550	0.1719	14.5250	0.0001•

"•" Implies Non-significance.

Table 4.20: b).Maximum of Likelihood estimates.

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	P-value
Population	**African/Black**	1	0.1311	0.0692	3.5865	0.0583
Population	**Coloured**	1	-0.9382	0.1308	51.4414	$<$ 0.0001•
Population	**Indian/Asian**	1	0.1341	0.1379	0.9454	0.3309
Education	**Less than primary completed**	1	0.0511	0.0779	0.4308	0.5116
Education	**No schooling**	1	-0.0679	0.1141	0.3545	0.5516
Education	**Other**	1	0.4500	0.1825	6.0767	0.0137•
Education	**Primary completed**	1	-0.0324	0.1030	0.0988	0.7533
Education	**Secondary completed**	1	-0.2667	0.0611	19.0760	$<$ 0.0001•
Education	**Secondary not completed**	1	-0.1461	0.0542	7.2783	0.0070•
Age_Gr	**16-25**	1	-0.5676	0.0969	34.3259	$<$ 0.0001•
Age_Gr	**26-35**	1	0.4503	0.0709	40.3180	$<$ 0.0001•
Age_Gr	**36-45**	1	0.8289	0.0690	144.1253	$<$ 0.0001•
Age_Gr	**46-55**	1	0.7813	0.0737	112.3432	$<$ 0.0001•
Age_Gr	**56-65**	1	0.4860	0.0850	32.6729	$<$ 0.0001•
Age_Gr	**66-75**	1	-0.6695	0.1469	20.7828	$<$ 0.0001•
Attend_school	**No**	1	1.1277	0.1080	108.9961	$<$ 0.0001•

"•" Implies Non-significance.

Table4.19 and Table4.20 display the final model after the elimination procedure and variables including gender, population group, marital status, education status (Education_status), geographic type of location (Geo_Type), province and age group were found to be significant, as their statistical P-values were less than 0.05 and only some categories or levels within other variables were not significant. All categories of variables with P-values more than 0.05 are considered to be non-significant. Only three response level under marital status variable namely; divorced, living together married were not found to be significant.

Table 4.21: The association between the observed responses and the predicted probabilities.

Percent Concordant	70.6	**Somers' D**	0.477
Percent Discordant	22.9	**Gamma**	0.51
Percent Tied	6.5	**Tau-a**	0.043
Pairs	69970554	**c**	0.739

The above Table4.21 shows the percent concordance and disconcordance, the value of $c = 0.766$ is called the index of concordance which is used to test if the model is able to predict the response variable "own business". The Tau-a measure how best the model it is when compared to the random chances, and the values has to be between 0 and 0.5 according to Agresti (2003).
The Reliable Operating Characteristics (ROC) Curve on Figure 4.11 below is used to measure how accurate the model can distinguish between the two categories of the response variable, (own business = Yes) and (own business = No). The results show an Area Under Curve (AUC) $= 0.76$, which implies that there are about 76% probabilities that the model can predict between the individuals who own businesses and those who do not own businesses.

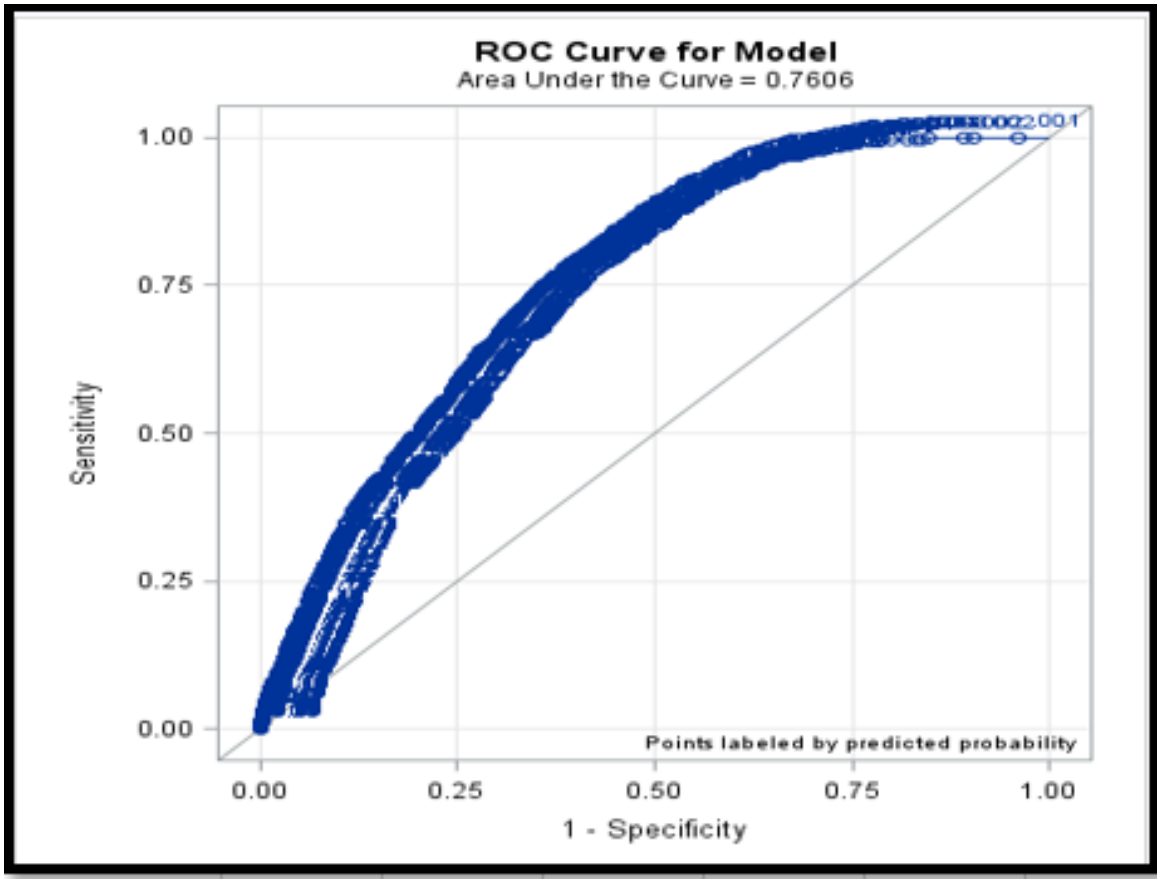

Figure 4.11: ROC curve analysis.

4.5 Application of Log-linear regression using QLFS

The Log-linear regression was applied in two different datasets collected from StatsSA. The first data was the 2017 secondary QLFS, the second data is the 2017 SESE.

From the QLFS that was also used in Section 4.3, the following categorical variables namely; own business, gender, population and age group were considered in the analysis. On this application of Log-linear regression analysis, the Log-linear analysis assumptions were tested under cell counts and residu-

als subsection, and the other tests include, the Goodness-of-Fit test, the K-way order effects as well as the estimation of the parameters. A saturated model was selected which contains all the interaction effects. Log-linear analysis is used to perform the model selection to identify the best model that fits the data and also to predict the associated factors.

4.5.1 Cell counts and residuals

One of the assumptions in the Log-linear models is the non-zero frequencies, as according to Agresti (2018), the expected cell count frequencies in the Log-linear analysis should be greater than 1 and table should not have more than 20% of the expected frequencies that are less than 5. The below Table4.22 and Table4.23 proves the above assumption.

Table 4.22: Part 1: Cell counts and residuals.

				Observed	Expected		
Own business	**gender**	**Population group**	**Age**	**count**	**count**	**Residuals**	**Std. Residuals**
Yes	textMale	textAfricans/Black	16-25	101.00	86.87	14.12	1.51
			26-35	295.00	286.62	8.37	0.49
			36-45	261.00	271.45	-10.45	-0.63
			46-55	143.00	154.12	-11.12	-0.89
			56-65	64.00	68.51	-4.51	-0.54
			66-75	13.00	9.19	3.80	1.25
		coloured	16-25	3.00	3.52	-0.53	-0.28
			26-35	10.00	7.99	2.00	0.70
			36-45	8.00	6.99	1.01	0.38
			46-55	4.00	4.85	-0.85	-0.38
			56-65	1.00	3.12	-2.12	-1.20
			66-75	1.00	0.51.	0.49	0.68
		Indian/Asian	16-25	5.00	3.38	1.61	0.87
			26-35	12.00	11.25	0.75	0.22.
			36-45	9.00	7.13	1.87	0.70
			46-55	3.00	4.90.	-1.90*	-0.86
			56-65	1.00	2.38	-1.38	-0.89
			66-75	0.00	0.89	-0.89	-0.94
		White	16-25	3.00	7.26	-4.26	-1.58
			26-35	15.00	20.38	-5.38	-1.19
			36-45	23.00	22.09	0.90	0.19
			46-55	23.00	21.77	1.23	0.26
			56-65	17.00	11.36	5.64	1.67
			66-75	7.00	5.01	1.98	0.88
	Female		16-25	39.00	48.76	-9.76	-1.39
			26-35	224.00	230.80	-6.81	-0.44
		African/Black	36-45	181.00	189.14	-8.13	-0.59
			46-55	119.00	103.42	15.58?	1.53
			56-65	71.00	60.59	10.40E	1.33
			66-75	10.00	11.50	-1.51	-0.44
		Coloured	16-25	2.00	1.11	0.88	0.84
			26-35	3.00	5.34	-2.34	-1.01
			66-45	9.00	4.88	4.12	1.86
			46-55	3.00	2.95	0.05	0.03
			56-65	0.00	2.21	-2.21	-1.49
			66-75	0.00	0.51	-0.51	-0.71
		indian/Asian	16-25	0.00	0.35	-0.35	-0.59
			26-35	2.00	2.37	-0.37	-0.24
			66-45	3.00	2.27	0.72	0.48
			46-55	2.00	1.56	0.43	0.34
			56-65	1.00	1.04	-0.04	-0.04
			66-75	0.00	0.40	-0.40	-0.63
		White	16-25	2.00	3.72	-1.73	-0.89
			26-35	33.00	29.22	3.77	0.69
			36-45	52.00	42.03	9.96	1.53
			46-55	32.00	35.41	-3.41	-0.57
			56-65	30.00	35.75	-5.75	-0.96
			66-75	11.00	13.96	-2.96	-0.79

Table 4.23: Part 2: Cell counts and residuals.

own Business	gender	Population group		Observed count	Expected count	Res iduals	Std. Residuals
No	Make	African/Black	16-25	5955.00	5969.08	-14.08	-0.18
			26-35	3377.00	3385.38	-8.38	-0.14
			66-45	2046.00	2035.56	10.44	0.23
			46-55	1123.00	1111.87	11.12	0.33
			56-65	623.00	618.52	4.48	0.18£
			66-75	246.00	249.81	-3.81	-0.24
		Coloured	16-25	583.00	582.48	0.52	0.02
			26-35	225.00	227.00	-2.00	-0.13.
			66-45	125.00	126.01	-1.01	-0.09
			46-55	85.00	84.14	0.85	0.09
			56-65	70.00	67.86	2.13	0.26
			66-75	33.00	33.48	-0.48	-0.08
		Indian/Asian	16-25	124.00	125.61	-1.61	-0.14
			26-35	71.00	71.75	-0.75	-0.09
			66-45	27.00	28.87	-1.87	-0.34
			46-55	21.00	19.09	1.90	0.43
			56-65	13.00	11.62	1.38	0.40
			66-75	14.00	13.10	0.89	0.24
		White	16-25	275.00	270.73	4.27	0.26
			26-35	136.00	130.62	5.38	0.47
			66-45	89.00	89.90	-0.90	-0.09
			46-55	84.00	85.22	-1.23	-0.13.
			56-65				
			66-75	72.00	73.98	-1.99	-0.23
	Female	African/Black	16-25	6241.00	6231.27	9.72	0.12.
			26-35	5077.00	5070.19	6.81	0.09
			66-45	2646.00	2637.85	8.14	0.16
			46-55	1372.00	1387.58	-15.58	-0.41
			56-65	1007.00	1017.37	-10.37	-0.32
			66-75	583.00	581.48	1.51	0.06.
		Coloured	16-25	613.00	613.89£	-0.90	-0.03
			26-35	509.00	506.66	2.34	0.10
			66-45	290.00	294.11	-4.11	-0.24
			46-55	171.00	171.05	-0.05	0.00
			56-65	163.00	160.76	2.24	0.17
			66-75	112.00	111.48	0.51	0.05
		Indian/Asian	16-25	110.00	109.65	0.35	0.03.
			26-35	128.00	127.62	0.37	0.03.
			66-45	77.00	77.72	-0.72	-0.08
			46-55	51.00	51.43£	-0.43£	-0.06
			56-65	43.00	42.95	0.05	0.00
			66-75	50.00	49.59	0.40	0.05
		White	16-25	282.00	280.27	1.73	0.10.
			26-35	374.00	377.77	-3.77	-0.19
			66-45	335.00	344.96	-9.96	-0.53
			46-55	283.00	279.58	3.41	0.20
			56-65	359.00	353.26	5.74	0.30
			66-75	418.00	415.04	2.95	0.14

4.5.2 Goodness-of-Fit Test

Table 4.24: Pearson and Deviance Goodness-Of-fit.

	Chi-Square	df	P-Value
Likelihood Ratio	48.6	35	0.063
Pearson	44.101	35	0.139

Table4.25 shows the Goodness-of-Fit test using both Likelihood and Pearson Chi-square which are based on the final model after the backward elimination procedure. The P-value for both statistic tests is less than 5%, which is the level of significance, therefore,the null hypothesis that the final model fits the data is not rejected. Hence, the final model fits the data well.

4.5.3 K-way and High order effects

Table 4.25: K-way and High order effects.

K		df	Likelihood Ratio		Pearson		
			Chi-Square	P-Value	Chi-Square	P-Value	Iterations
K-way and Higher Order	1	95	120967.99	0.00	284925.88	0.00	0
	2	85	4168.75	0.00	5393.08	0.00	2
	3	53	216.77	0.00	213.02	0.00	4
	4	15	18.22	0.25	17.67	0.28	3
K-way Effects	1	10	116799.24	0.00	279532.80	0.00	0
	2	32	3951.98	0.00	5180.06	0.00	0
	3	38	198.55	0.00	195.35	0.00	0
	4	15	18.22	0.25	17.67	0.28	0

Table4.25 shows two K-way tests, the first fours rows test the null hypothesis that the K-way and high order effects are zero while the remaining four rows is used to test null hypothesis that the K-way effects are zero. Likelihood and Pearson Chi-square are utilised to test the statistically significance of K-way and high order effects, results indicates the non-significance difference of 4-way

(k=4) effects with P-value $= 0.28$ while all other effects (k=1,2,3) are significant with P-value < 0.05.

4.5.4 Parameter estimation

The purpose of the parameter estimation is to investigate the significant interaction between Four factors, namely; own business(B), gender(G), population group(P) and age(A). The results on Table4.26 to Table4.27 show the coefficient estimate, Z-values, P-values and the confidence interval. According to the results, the main effect has significant coefficients, which implies that all factors have a separate significant effects on the model. Under the interaction effects, only the two-way interactions such as B*G, P*A, B*A, G*A has significant coefficients. The 4-way and all 3-way interactions have insignificant coefficients since the P-values are less than 5%. It means that removing both 4-way and 3-way interaction will not significantly affect the model.

Table 4.26: Parameter estimation(a).

Effect		Estimate	Z	P-Value	95% Confidence	
					Lower Bound	Upper Bound
B*G*P*A	1	0.107	0.961	0.336	-0.111	0.325
	2	-0.033	-0.473	0.636	-0.172	0.105
	3	0.022	0.331	0.740	-0.108	0.152
	4	-0.008	-0.102	0.919	-0.162	0.146
	5	-0.101	-0.857	0.392	-0.333	0.130
	6	-0.213	-1.191	0.234	-0.564	0.138
	7	0.105	0.766	0.444	-0.164	0.375
	8	-0.121	-0.981	0.327	-0.362	0.120
	9	-0.036	-0.239	0.811	-0.332	0.260
	10	0.120	0.439	0.661	-0.418	0.658
	11	0.180	0.705	0.481	-0.321	0.681
	12	0.096	0.631	0.528	-0.202	0.395
	13	0.126	0.866	0.386	-0.160	0.412
	14	-0.050	-0.292	0.771	-0.386	0.286
	15	-0.152	-0.671	0.502	-0.595	0.291
B*G*P	1	-0.110	-2.342	0.019	-0.203	-0.018
	2	0.059	0.646	0.518	-0.119	0.237
	3	0.138	1.373	0.170	-0.059	0.336
B*G*A	1	-0.033	-0.309	0.758	-0.242	0.176
	2	0.030	0.448	0.654	-0.102	0.163
	3	-0.040	-0.638	0.524	-0.163	0.083
	4	-0.070	-0.935	0.350	-0.216	0.077
	5	0.022	0.191	0.849	-0.202	0.246
B*P*A	1	-0.042	-0.374	0.708	-0.260	0.176
	2	0.010	0.145	0.885	-0.128	0.149
	3	-0.140	-2.106	0.035	-0.270	-0.010
	4	0.008	0.107	0.915	-0.146	0.162
	5	0.156	1.320	0.187	-0.076	0.388
	6	0.171	0.957	0.339	-0.180	0.522
	7	-0.029	-0.210	0.833	-0.299	0.241
	8	0.100	0.815	0.415	-0.141	0.341
	9	0.008	0.053	0.958	-0.288	0.304
	10	-0.441	-1.608	0.108	-0.979	0.097

Table 4.27: Parameter estimation(b).

Estimate		Estimate	Z	P-Value	95% Confidence	
					LowerBound	UpperBound
P*A	1	0.294	2.645	0.008	0.076	0.512
	2	0.301	4.258	0	0.163	0.44
	3	0.111	1.677	0.094	-0.019	0.242
	4	0.012	0.159	0.874	-0.142	0.166
	5	-0.033	-0.275	0.783	-0.264	0.199
	6	0.431	2.409	0.016	0.08	0.782
	7	0.004	0.033	0.974	-0.265	0.274
	8	0.096	0.785	0.433	-0.145	0.337
	9	-0.07	-0.465	0.642	-0.366	0.226
	10	-0.384	-1.398	0.162	-0.922	0.154
	11	0.103	0.402	0.688	-0.398	0.604
	12	0.108	0.71	0.478	-0.19	0.407
	13	-0.059	-0.408	0.684	-0.346	0.227
	14	-0.087	-0.509	0.611	-0.423	0.249
	15	-0.078	-0.343	0.731	-0.52	0.365
B	1	-1.583	-35.601	0	-1.671	-1.496
G	1	-0.107	-2.412	0.016	-0.194	-0.02
P	1	2.097	44.502	0	2.004	2.189
	2	-0.733	-8.067	0	-0.911	-0.555
	3	-1.509	-14.976	0	-1.706	-1.311
A	1	0.205	1.923	0.055	-0.004	0.414
	2	0.707	10.433	0	0.574	0.839
	3	0.524	8.336	0	0.401	0.648
	4	0.055	0.733	0.463	-0.092	0.201
	5	-0.453	-3.961	0	-0.677	-0.229

4.6 Application of Log-linear Regression using SESE

4.6.1 Cell counts and residuals

Cell counts in this section are used to check if expected cell count frequencies in the Log-linear analysis is greater than 1 and that the table have more than 20% of the expected frequencies that are less than 5.

Table 4.28: Cell counts and residuals

			Observed		Expected		Residuals	Std. Residuals
Better access to loans	**Gender**	**Population**	**Count**	**%**	**Count**	**%**		
Yes	Male	African/Black	223	0.143	228.133	0.147	-5.133	-0.340
		Coloured	13	0.008	6.473	0.004	6.527	2.565
		Indian/Asian	5	0.003	3.315	0.002	1.685	0.925
		White	7	0.005	7.578	0.005	-0.578	-0.210
	Female	African/Black	232	0.149	228.133	0.147	3.867	0.256
		Coloured	3	0.002	6.473	0.004	-3.473	-1.365
		Indian/Asian	2	0.001	3.315	0.002	-1.315	-0.722
		White	6	0.004	7.578	0.005	-1.578	-0.573
No	Male	African/Black	498	0.320	494.367	0.318	3.633	0.163
		Coloured	10	0.006	14.027	0.009	-4.027	-1.075
		Indian/Asian	11	0.007	7.185	0.005	3.815	1.423
		White	16	0.010	16.422	0.011	-0.422	-0.104
	Female	African/Black	492	0.316	494.367	0.318	-2.367	-0.106
		Coloured	15	0.010	14.027	0.009	0.973	0.260
		Indian/Asian	3	0.002	7.185	0.005	-4.185	-1.561
		White	19	0.012	16.422	0.011	2.578	0.636

Table4.28 tables is a three-way effect between Better access to loans, Gender and Population group. The observed, expected count of cells and the residuals performed by SAS version 9.2, PROC GENMOD system. The table consists of 16 cells and only 2% of cells have less 5 expected counts, the two cells are both from individuals with better access to loans from Indian/Asians population group. Majority of black business owners indicated to not have better access to loans from both genders and percentage expected count for both males

and females is 31.8%. The percentage expected counts for both blacks who indicated to have better access to loans is 14.7% for both males and females. In terms of the Residuals, the standardised residuals for all counts are between -2 and $+2$.

4.6.2 Goodness-of-fit Test

Table 4.29: Goodness-of-fit Test

	Chi-Square	**df**	**P-value**
Likelihood Ratio	15.99283537	11	0.141
Pearson	16.51719827	11	0.123

The above Table4.29 indicates the Goodness-of-Fit test from both Likelihood and Pearson Chi-square which are based on the final model after the backward elimination procedure. The P-value for both statistic tests is less than 5% level of significance, hence, we fail to reject the null hypothesis that the final model fits the data.

4.6.3 K-way and High order effects

Table 4.30: K-way and High order effects

K		df	Likelihood Ratio		Pearson	
			Chi-Square	p-value	Chi-Square	p-value
K-way and Higher Order Effects	1	15	3518.729824	0	4567.032	0.000
	2	10	15.91502122	0.102091	16.404	0.089
	3	3	7.719453976	0.05218	7.330	0.062
K-way Effects	1	5	3502.814803	0	4550.628	0.000
	2	7	8.195567239	0.315665	9.074	0.247
	3	3	7.719453976	0.05218	7.330	0.062

The purpose of K-way and High-Order effects is to test if the removal of the terms will significantly affect the model fit. Table4.30 shows two parts of K-way tests, the first part is used to test if the K-way and high order effects are zero, while the second part is used to test if only the K-way effects are zero. Likelihood and Pearson Chi-square are used to test the statistically significance of both K-way and high order effects. The above table indicates that the main effects (K=1) are significant (p-value < 0.05), which means that the individual or separate effects of gender, better access to loans, and population group is significant. The 2-way and 3-way effects were found to be statistically insignificant.

Chapter 5

Discussion and conclusion

5.1 Introduction

This section covers the discussion of the findings in Chapter 4, the conclusion based of the findings, limitations of the study and further research direction. The section also demonstrates how Logistics Regression (LR) and Log-linear Regression (LLR) model as one of the Count models were used to achieve the aim and objectives of the research. The findings of the study are then compared with the literatures reviewed by the study. The discussion and the conclusion of this study are also based on the two data sets used, which are 2017 Quarterly Labour Force Survey (QLFS) and 2017 Survey of Employers and the Self Employers (SESE).

5.2 Discussion on descriptive statistics

The 2017 QLFS which was collected from StatsSA was used to test the association between variables. The overall result from Chi-square statistics showed

a strong association between business ownership and the following categorical variables; gender, population group, marital status, age group, attended school and province.

5.3 Discussion on Multiple Logistic Regression using QLFS

One of the research objectives was to utilise Multiple Logistic regression to investigate the factors affecting business ownership in South Africa by using 2017 QLFS data. In oder to achieve this objective, the most fitted model or final model with the lower Akaike Information Criterion (AIC) was selected during the analysis. The final model consists of one binary response variables (own business) modeled against eight independent variables (gender, population group, age group, geographical type, marital status, province, education status and attended school). The main effects were found to be statistically significant, which implies that there is a high probability that each variable or factor affects business ownership in South Africa.

5.3.1 Estimates of each coefficient.

The main effect results in Section 4.18 influenced this study to further investigate the significant effect of each variable coefficient towards the response variable using MLE. Significant coefficients have higher probability of affecting business ownership, while non-significant coefficients implies lower chance of affecting business ownership in South Africa. Bellow is the breakdown of each factor in the final Multinomial Logistics Regression model;

- Marital status.

  Under marital status factor, only single/never married coefficient was found significant while coefficients for divorced, living together and mar-

ried were not significant. This implies that individuals who are single or never got married have high probability of affecting business ownership in South Africa.

- Province.

  The coefficients for Eastern Cape, Kwa-Zulu Natal and North West were found to be non-significant, which implies that those provinces had less probabilities of affecting business ownership in South Africa. The coefficients for other provinces such as Limpopo, Mpumalanga, Free State, Gauteng and Northern Cape was found significant.

- Geographical type.

  From three geographical types (Farm, Mining areas, Traditional areas) in South Africa, only the coefficient for Mining areas was found not significant while other coefficients were significant.

- Population group.

  The coefficient for Black/African was found not significant, although its P-value of 0.0583 was close to the level of significant. The other insignificant coefficient was for Indians/Asians. The only significant coefficient was for the Coloured population group.

- Education status.

  Under education status, the coefficient for individual with no school, less than primary education, completed were not significant. While, the coefficient for individuals completed secondary education were found significant.

- Gender, age group and attended school.

  All coefficients for gender , age group and attended school were found highly significant. Studies conducted by Preisendoerfer et al. (2014), Global Entrepreneurship Monitor (2017), Peters and Brijlal (2011), Maliranta

and Nurmi (2019) also indicated the attendance of school as one of the factors affecting business ownership. Findings from Chipeta et al. (2016), Van Scheers (2010) and Giandrea et al. (2008) also indicated age group and gender as some of the significant factors affecting business ownership.

5.4 Discussion on Log-linear Regression using QLFS data

The 4-way contingency table result shows that the majority of business owners are Black African men of age group $(26-35)$ and $(36-45)$, and the majority of non-business owners are Black African women within the age groups of $(16-25)$ and $(26-35)$, this finding is similar to a study conducted by Van Scheers (2010). This study also discovered in the that Coloureds and Indians/Asians population where dominated by all categories.

Log-linear Regression analysis was employed to investigate if there is a significant association between four factors namely; own business, gender, population group and age group. According to the study results, only the 3-way effect is statistically significant while the 4-way effects is not significant. This implies that the third order interaction effects have higher likelihood to improve the model than the fourth oder interaction effects.

5.5 Discussion on Log-linear Regression using SESE data

The purpose of using SESE 2017 was to analyse the financial accessibility of business owners in South Africa, hence, the usage of variable “better access to

loan" was considered, the variable had two responses(yes and no). The SESE 2017 consisted of all business owners identified in the QLFS 2017 data to investigate their financial accessibility in order to fund their businesses. The research reached the similar findings as Leshilo and Lethoko (2017), which found that the majority of business owners have no better access to loans to fund their startup businesses. Amongst the study findings, black population group showed a high dominance in terms of financial accessibility, these results supports the United State of America (USA) study conducted by Fairlie (2004), which also indicated that there is a great improvement regarding the financial accessibility by black population group. Females indicates also indicated to have better access to loans compared to males, this contradicts with the Witbooi and Ukpere (2011) findings, which indicated that female business owners had lower financial accessibility than male business owners. The overall Log-linear analysis results indicate that these factors (gender, better access to loans and population group) have a separate or individual significant influence on the model,

5.6 Overall conclusion

This study applied MLR to model business ownership from the 2017 QLFS and to identify factors affecting business ownership in South Africa. Amongst all selected factors (including, marital status, province, geographical type, population group and education status), only gender, age group and attended school were found to be the most significant factors affecting business ownership since all their categories had significant coefficients. The study further applied LLR on 2017 QLFS to investigating the possible association between business ownership, gender, age and population group. The significant (K=1,2,3)-Way interaction effects implies that the inclusion of main effects as well as the interaction between business ownership, gender and age group significantly improves

the model.

The overall conclusion reached by this study after the application of Log-linear analysis on the SESE 2017 is that, majority of business owners in South Africa do not have better access to loans in order to sustain their businesses. The study also concludes that the interaction effects for gender, better access to loans and population group is not significant, while only the main effects of the above factors significantly improve the model fit.

The study aimed to model and analyse business ownership in South Africa using the statistical models. The two models, MLR and LLR were utilised to model and analyse business ownership using the 2017 QLFS and 2017 SESE. Hence, the aim of the study is achieved. One of the objectives set by the study was to utilise LR and count models to model business ownership. MLR (which is a Logistic Regression model with more than one independent variables) and LLR (which is one of the Count models) were also applied to model business ownership. The other objective of the study was to analyse the accessibility of finance by business owners, this objective was achieved by using LLR with multi-way contingency tables to investigate the association between better access to loans, gender and population group. The other objective of the study was to perform a comparative study of LR and LLR model. This objective was achieved by applying both methods on the same dataset (2017 QLFS) and the association testing results from both methods were similar. Hence, the two methods, Logistic and Log-linear Regression model could work together to produce reliable results.

5.7 Limitations of the study

In this study, 2017 QLFS and SESE 2017, both collected from StatsSA, were used to study business ownership. The latest SESE was last conducted in 2017 by StatsSA, hence, the year was selected for the purpose of the study. The

year in which the data was collected, which is (2017) is considered as one of the study limitations as there is a little over four years difference from 2017 and the year in which the study is conducted. The second limitation identified under methodology was that under GLM, only Logistic and Log-linear regression methods were utilised, however, there are other GLM extension models such as Generalised Linear Mixed Models (GLMMs) and Generalised Additive Models (GAMs) that could be used to study business ownership.

5.8 Future research direction

For future studies, other researchers can consider studying business ownership using the latest QLFS and SESE data to improve the research quality outcomes. Future studies could also study the effect of COVID-19 pandemic on business owners in South Africa. In terms of methodology, interaction effect models under Logistic regression could be used to extensively investigate factors related to business ownership. Future studies could also utilise GLMMs to study business ownership in South Africa.

References

ABOR, J. AND QUARTEY, P. (2010). Issues in SME development in Ghana and South Africa. *International research journal of finance and economics*, **39** (6), 215–228.

ACQUAH, H. AND CARLO, M. (2010). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of an asymmetric price relationship. *Journal of Development and Agricultural Economics*, **2** (1), 1–6.

AGRESTI, A. (2003). *Categorical data analysis*, volume 482. John Wiley & Sons.

AGRESTI, A. (2018). *An introduction to categorical data analysis*. Wiley.

AKAIKE, H. (1987). Factor analysis and AIC. *Selected Papers of Hirotugu Akaike*, **9** (4), 371–386.

AKINSOMI, O., KOLA, K., NDLOVU, T., AND MOTLOUNG, M. (2016). The performance of the broad based black economic empowerment compliant listed property firms in south africa. *Journal of Property Investment & Finance*.

ATKINS, D. C., BALDWIN, S. A., ZHENG, C., GALLOP, R. J., AND NEIGHBORS, C. (2013). A tutorial on count regression and zero-altered count models for longitudinal substance use data. *Psychology of Addictive Behaviors*, **27** (1), 166.

BARBER, J. AND THOMPSON, S. (2004a). Multiple regression of cost data: use of Generalised Linear Models. *Journal of health services research & policy*, **9** (4), 197–204.

BARBER, J. AND THOMPSON, S. (2004b). Multiple regression of cost data: use of Generalised Linear Models. *Journal of health services research & policy*, **9** (4), 197–204.

BERESFORD, M. (2020). Entrepreneurship as legacy building: Reimagining the economy in post-apartheid South Africa. *Economic Anthropology*, **7** (1), 65–79.

BHORAT, H., ASMAL, Z., LILENSTEIN, K., AND VAN DER ZEE, K. (2018). SMMEs in South Africa: Understanding the constraints on growth and performance. *African Economic Outlook*.

BRIJLAL, P., NAICKER, V., AND PETERS, R. (2013). Education and smme business growth: A gender perspective from South Africa. *International Business & Economics Research Journal (IBER)*, **12** (8), 855–866.

CAMERON, A. C. AND TRIVEDI, P. K. (2013). *Regression analysis of count data*, volume 53. Cambridge university press.

CARRÈRE, C., GRUJOVIC, A., AND ROBERT-NICOUD, F. (2015). Dp10692 trade and frictional unemployment in the global economy. *Small Business Economics*, **19**, 271–290.

CARTER, S. AND WEEKS, J. (2002). Gender and business ownership: international perspectives on entrepreneurial theory and practice. *International Journal of Entrepreneurship and Innovation*, **3** (2), 81–82.

CHEN, K., HUANG, R., CHAN, N. H., AND YAU, C. Y. (2019). Subgroup analysis of zero-inflated poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics*, **86**, 8–18.

CHIPETA, E., SURUJLAL, J., AND KOLOBA, H. (2016). Influence of gender and age on social entrepreneurship intentions among university students in Gauteng province, South Africa. *Gender and Behaviour*, **14** (1), 6885–6899.

CLOVER, T. AND DARROCH, M. A. (2005a). Owners'perceptions of factors that constrain the survival and growth of small, medium and micro agribusinesses in Kwa-Zulu Natal South Africa. *Agrekon*, **44** (2), 238–263.

CLOVER, T. AND DARROCH, M. A. (2005b). Owners'perceptions of factors that constrain the survival and growth of small, medium and micro agribusinesses in Kwa-Zulu Natal South Africa. *Agrekon*, **44** (2), 238–263.

COLIN, T. (2019). The state of South African small business.
URL: *https://www.xero.com/content/dam/xero/pdf/2019-South-African-small-business-report.pdf*

DEPARTMENT OF TRADE AND INDUSTRY (2019). Republic of South Africa, Broad-Based Black Economic Empowerment Amendment Act: Codes of Good Practice Sector Code.

DOTSON, W. H., RICHMAN, D. M., ABBY, L., THOMPSON, S., AND PLOTNER, A. (2013). Teaching skills related to self-employment to adults with developmental disabilities: An analog analysis. *Research in Developmental Disabilities*, **34** (8), 2336–2350.

DUNLAVY, A., JUÁREZ, S., TOIVANEN, S., AND ROSTILA, M. (2017). Migration background characteristics and the association between unemployment and suicide andrea Dunlavy. *European Journal of Public Health*, **27** (suppl_3).

FAIRLIE, R. W. (2004). Recent trends in ethnic and racial business ownership. *Small Business Economics*, **23** (3), 203–218.

FAIRLIE, R. W. AND ROBB, A. M. (2007). Why are black-owned businesses less

successful than white-owned businesses? the role of families, inheritances, and business human capital. *Journal of Labor Economics*, **25** (2), 289–323.

FAIRLIE, R. W. AND ROBB, A. M. (2009). Gender differences in business performance: evidence from the characteristics of business owners survey. *Small Business Economics*, **33** (4), 375.

FAMOYE, F. AND SINGH, K. P. (2006). Zero-inflated generalized poisson regression model with an application to domestic violence data. *Journal of Data Science*, **4** (1), 117–130.

FATOKI, O. (2014). Immigrant entrepreneurship in South Africa: Current literature and research opportunities. *Journal of Social Sciences*, **40** (1), 1–7.

FRESE, M., HASS, L., AND FRIEDRICH, C. (2016). Personal initiative training for small business owners. *Journal of Business Venturing Insights*, **5**, 27–36.

FRESE, M., KRAUSS, S. I., KEITH, N., ESCHER, S., GRABARKIEWICZ, R., LUNENG, S. T., HEERS, C., UNGER, J., AND FRIEDRICH, C. (2007). Business owners' action planning and its relationship to business success in three African countries. *Journal of applied psychology*, **92** (6), 1481.

GEORGELLIS, Y. AND WALL, H. J. (2005). Gender differences in self-employment. *International review of applied economics*, **19** (3), 321–342.

GIANDREA, M. D., CAHILL, K. E., QUINN, J. F., ET AL. (2008). *Self-employment transitions among older American workers with career jobs*. US Department of Labor, US Bureau of Labor Statistics, Office of ....

GILL, A. AND BIGER, N. (2012). Barriers to small business growth in Canada. *Journal of Small Business and Enterprise Development*, **19** (4), 656–668.

GLOBAL ENTREPRENEURSHIP MONITOR (2017). South Africa:Can small businesses survive IN South Africa?

URL: *https://www.gemconsortium.org/report/gem-south-africa-2016-2017-report-1494860333.pdf*

GLOBAL ENTREPRENEURSHIP MONITOR (2019). Igniting startups for economic growth and social change.
URL: *https://www.usb.ac.za/wp-content/uploads/2020/06/GEMSA-2019-Entrepreneurship-Report-web.pdf*

GUERRA, G. AND PATUELLI, R. (2016). The role of job satisfaction in transitions into self–employment. *Entrepreneurship Theory and Practice*, **40** (3), 543–571.

HANSEN, B. E. (2007). Least squares model averaging. *Econometrica*, **75** (4), 1175–1189.

HATFIELD, I. (2015). Self-employment in europe. Technical report, IPPR London.

HENNING, S. AND AKOOB, K. (2017). Motivational factors affecting informal women entrepreneurs in North-West province. *The Southern African Journal of Entrepreneurship and Small Business Management*, **9** (1), 1–10.

HIKIDO, A. (2018). Entrepreneurship in South African township tourism: The impact of interracial social capital. *Ethnic and Racial Studies*, **41** (14), 2580–2598.

HIRZEL, A. H., HELFER, V., AND METRAL, F. (2001). Assessing habitat-suitability models with a virtual species. *Ecological modelling*, **145** (2-3), 111–121.

HOSMER JR, D. W., LEMESHOW, S., AND STURDIVANT, R. X. (2013). *Applied logistic regression*, volume 398. John Wiley & Sons.

HUIS, M., LENSINK, R., VU, N., AND HANSEN, N. (2019). Impacts of the Gender and Entrepreneurship Together Ahead (GET Ahead) training on em-

powerment of female microfinance borrowers in Northern Vietnam. *World Development*, **120**, 46–61.

HUNDLEY, G. (2000). Male/female earnings differences in self-employment: The effects of marriage, children, and the household division of labor. *ILR Review*, **54** (1), 95–114.

INTERNATIONAL FINANCE CORPORATION (2019). Banking on SMEs:trends and challenges.
URL: *https://www.ifc.org/wps/wcm/connect/dd06b824-c38b-4933-9108-0c834f182fee/IFC+on+Banking+SMEs+Publication+June+2019.pdf?MOD=AJPERESCVI*

IRENE, B. N. (2017). Women entrepreneurship in South Africa: Understanding the role of competencies in business success. *The Southern African Journal of Entrepreneurship and Small Business Management*, **9** (1), 1–9.

IVERSEN, J., MALCHOW-MØLLER, N., AND SØRENSEN, A. (2010). Returns to schooling in self-employment. *Economics Letters*, **109** (3), 179–182.

JUSTO, R. AND DETIENNE, D. R. (2008). Gender, family situation and the exit event: Reassessing the opportunity-cost of business ownership. *Madrid IE Business School Working Paper, WP08*, **26** (1).

KELLEY, D. J., BRUSH, C. G., GREEN, P., AND LITOVSKY, Y. (2011). Global entrepreneurship monitor (GEM): 2010 women's report. *Babson Park: Babson College*.

KENGNE, B. D. S. (2016). Mixed-gender ownership and financial performance of SMEs in South Africa. *International Journal of Gender and Entrepreneurship*.

KEYSER, M., DE KRUIF, M., AND FRESE, M. (2000). The psychological strategy process and sociodemographic variables as predictors of success for micro-and small-scale business owners in Zambia. *Journal of Management*.

KONGOLO, M. (2010). Job creation versus job shedding and the role of SMEs in economic development. *African journal of business management*, **4** (11), 2288–2295.

LADZANI, W. (2010). Historical perspective of small business development initiatives in South Africa with special reference to Limpopo province. *Problems and Perspective in Management*, **8** (3), 68–79.

LEFORT, V., LONGUEVILLE, J.-E., AND GASCUEL, O. (2017). Sms: smart model selection in PhyML. *Molecular biology and evolution*, **34** (9), 2422–2424.

LEONI, T. AND FALK, M. (2010). Gender and field of study as determinants of self-employment. *Small Business Economics*, **34** (2), 167–185.

LESHILO, A. AND LETHOKO, M. (2017). The contribution of youth in local economic development and entrepreneurship in Polokwane municipality, Limpopo Province. *Skills at Work: Theory and Practice Journal*, **8** (1), 45–58.

LIAO, J., YU, S., YANG, F., YANG, M., HU, Y., AND ZHANG, J. (2016). Short-term effects of climatic variables on hand, foot, and mouth disease in mainland China, 2008–2013: a multilevel spatial poisson regression model accounting for overdispersion. *PLoS One*, **11** (1).

LIDDLE, A. R. (2007). Information criteria for astrophysical model selection. *Monthly Notices of the Royal Astronomical Society: Letters*, **377** (1), L74–L78.

LIU, X. S., LOUDERMILK, B., AND SIMPSON, T. (2014). Introduction to sample size choice for confidence intervals based on t statistics. *Measurement in Physical Education and Exercise Science*, **18** (2), 91–100.

MALIRANTA, M. AND NURMI, S. (2019). Business owners, employees, and firm performance. *Small Business Economics*, **52** (1), 111–129.

MAMMAN, A., BAWOLE, J., AGBEBI, M., AND ALHASSAN, A.-R. (2019). Sme policy formulation and implementation in Africa: Unpacking assumptions as opportunity for research direction. *Journal of Business Research*, **97**, 304–315.

MARLOW, S., HENRY, C., AND CARTER, S. (2009). Exploring the impact of gender upon women's business ownership: Introduction. *International small business journal*, **27** (2), 139–148.

MARTINEZ-GRANADO, M. (2002). Self-employment and labour market transitions: A multiple state model. *African Economic Journal*.

MASERUMULE, M. H. (2015). A legacy of perseverance-nafcoc: 50 years of leadership in business, Kwandiwe Kondlo: book review. *New Agenda: South African Journal of Social and Economic Policy*, **2015** (58), 57–58.

MBOKO, S. AND SMITH-HUNTER, A. (2010). Zimbabwe women business owners: Survival strategies and implications for growth. *Journal of Applied Business & Economics*, **11** (2).

MELTON, D., BENTING, S., BEYER, G., AND VENABLES, J. (2019). Women entrepreneurs in Cape Town, South Africa: Challenges and opportunities. *In ICGR 2019 2nd International Conference on Gender Research*. Academic Conferences and publishing limited, p. 392.

MITCHELL, B. (2004). Motives of entrepreneurs: A case study of South Africa. *The Journal of Entrepreneurship*, **13** (2), 167–183.

MOY, J. W. AND LUK, V. W. (2003). The life cycle model as a framework for understanding barriers to sme growth in hong kong. *Asia Pacific Business Review*, **10** (2), 199–220.

MUDENDA, P. M. (2013). *Business transformation in Durban: perceptions of*

black entrepreneurs in the context of black economic empowerment. Ph.D. thesis, University of Kwa-Zulu Natal.

MUSARA, M. AND GWAINDEPI, C. (2014). Factors within the business regulatory environment affecting entrepreneurial activity in South Africa. *Mediterranean Journal of Social Sciences*, **5** (6), 109.

NDHLOVU, T. P. AND SPRING, A. (2009). South african women in business and management: Transformation in progress. *Journal of African Business*, **10** (1), 31–49.

NELDER, J. A. AND WEDDERBURN, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, **135** (3), 370–384.

NEMAENZHE, P. P. ET AL. (2011). *Retrospective analysis of failure causes in South African small businesses*. Ph.D. thesis, University of Pretoria.

NYAKUDYA, F. W., SIMBA, A., AND HERRINGTON, M. (2018). Entrepreneurship, gender gap and developing economies: the case of post-apartheid South Africa. *Journal of Small Business & Entrepreneurship*, **30** (4), 293–324.

ODEKU, K. O. AND RUDOLF, S. S. (2019). An analysis of the transformative interventions promoting youth entrepreneurship in South Africa. *Academy of Entrepreneurship Journal*, **25** (4).

OKPARA, J. O. AND WYNN, P. (2007). Determinants of small business growth constraints in a sub-Saharan African Economy. *SAM advanced management journal*, **72** (2), 24.

PEDELI, X. AND VARIN, C. (2018). Pairwise likelihood estimation of latent autoregressive count models. *arXiv preprint arXiv:1805.10865*.

PENNY, W. D. (2012). Comparing dynamic causal models using AIC, BIC and free energy. *Neuroimage*, **59** (1), 319–330.

PETERS, R. M. AND BRIJLAL, P. (2011). The relationship between levels of education of entrepreneurs and their business success: a study of the province of KwaZulu-Natal, South Africa. *Industry and higher education*, **25** (4), 265–275.

PHILLIPS, M., MOOS, M., AND NIEMAN, G. (2014). The impact of government support initiatives on the growth of female businesses in Tshwane South Africa. *Mediterranean Journal of Social Sciences*, **5** (15), 85–85.

POHAR, M., BLAS, M., AND TURK, S. (2004). Comparison of logistic regression and linear discriminant analysis: a simulation study. *Metodoloski zvezki*, **1** (1), 143.

PONTE, S., ROBERTS, S., AND VAN SITTERT, L. (2007). 'black economic empowerment', business and the state in South Africa. *Development and Change*, **38** (5), 933–955.

PREISENDOERFER, P., BITZ, A., AND BEZUIDENHOUT, F. J. (2014). Black entrepreneurship: A case study on entrepreneurial activities and ambitions in a South African township. *Journal of Enterprising Communities: People and places in the global economy*.

PRETORIUS, M., NIEMAN, G., AND VAN VUUREN, J. (2005). Critical evaluation of two models for entrepreneurial education. *International Journal of Educational Management*.

REPUBLIC OF SOUTH AFRICA (2003). Bbbee act - republic of south africa: Pretoria, 53. *Act, Broad-Based Black Economic Empowerment*.

RIAZ, I. AND BATOOL, M. (2018). Unemployment; factors affect educated youth of multan, pakistan. *In 16th International Conference on Statistical Sciences*. Taylor & Francis, p. 139.

RIDOUT, M., HINDE, J., AND DEMÉTRIO, C. G. (2001). A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives. *Biometrics*, **57** (1), 219–223.

ROODT, J. (2005). Self-employment and the required skills. *Management Dynamics: Journal of the Southern African Institute for Management Scientists*, **14** (4), 18–33.

ROUSE, J. AND JAYAWARNA, D. (2006). The financing of disadvantaged entrepreneurs. *International Journal of Entrepreneurial Behavior & Research*, **12** (6).

S.B.A, U. S. (2012). Growing your business.
URL: *http://www.sba.gov/content/ideas-growing-your-business*

SCHWARZ, G. ET AL. (1978). Estimating the dimension of a model. *The annals of statistics*, **6** (2), 461–464.

SEDA (2021). The global competitiveness report.
URL: *https://edse.org.za/wp-content/uploads/2021/04/SMME-Quarterly-2020-Q3*$_0$8032021.*pdf*

SEEKINGS, J. AND NATTRASS, N. (2008). *Class, race, and inequality in South Africa*. Yale University Press.

SILVA, J. S. AND TENREYRO, S. (2010). On the existence of the maximum likelihood estimates in poisson regression. *Economics Letters*, **107** (2), 310–312.

SITHARAM, S. AND HOQUE, M. (2016). Factors affecting the performance of small and medium enterprises in kwazulu-natal, south africa. *Problems and perspectives in Management*, **14** (2), 277–288.

SIXABA, Z. AND ROGERSON, C. M. (2019). Black economic empowerment and south african tourism: The early pioneers. *African Journal of Hospitality, Tourism and Leisure*, **8** (4), 1–10.

STATSSA (2019). Quarterly labour force survey, quarter 1: 2019.

TEMTIME, Z. T. AND PANSIRI, J. (2004). Small business critical success/failure factors in developing countries: some evidences from botswana. *American Journal of Applied Sciences*, **1** (1), 18–25.

TERVO, H. (2008). Self-employment transitions and alternation in finnish rural and urban labour markets. *Papers in Regional Science*, **87** (1), 55–76.

THURIK, A. R., CARREE, M. A., VAN STEL, A., AND AUDRETSCH, D. B. (2008). Does self-employment reduce unemployment? *Journal of Business Venturing*, **23** (6), 673–686.

VALDEZ, Z. AND ROMERO, M. (2018). *Intersectionality and Ethnic Entrepreneurship*. Routledge.

VAN PRAAG, C. M. (2003). Business survival and success of young small business owners. *Small business economics*, **21** (1), 1–17.

VAN SCHEERS, L. (2010). The role of ethnicity and culture in developing entrepreneurs in south africa. *Problems and perspectives in Management*, **8** (4), 20–28.

WANG, Y. AND LIU, Q. (2006). Comparison of Akaike information criterion (AIC) and Bayesian information criterion (BIC) in selection of stock–recruitment relationships. *Fisheries Research*, **77** (2), 220–225.

WITBOOI, M. AND UKPERE, W. (2011). Indigenous female entrepreneurship: Analytical study on access to finance for women entrepreneurs in South Africa. *African Journal of Business Management*, **5** (14), 5646.

WORLD BANK (2017). Smes finance.

URL: *https://www.worldbank.org/en/topic/smefinance*

WORLD ECONOMIC FORUM (2019). The global competitiveness report.

URL: *http://www3.weforum.org/docs/WEF*$_{T}heGlobalCompetitivenessReport2019.pdf$

ZEILEIS, A., KLEIBER, C., AND JACKMAN, S. (2008). Regression models for count data in r. *Journal of statistical software*, **27** (8), 1–25.

ZHENG, D., THOMPSON, C., REICHL, P., WHITE, J., SETHI, H., ET AL. (2016). Tamping prediction using Generalised Linear Models with gamma distribution. *CORE 2016: Maintaining the Momentum*, 724.

ZOU, G. AND DONNER, A. (2013). Extension of the modified poisson regression model to prospective studies with correlated binary data. *Statistical methods in medical research*, **22** (6), 661–670.

APPENDIX A:STATISTICAL ANALYSIS SOFTWARE VERSION 9.2 (SAS) SCRIPTS FOR DISCRIPTIVE STATISTICS

proc freq data = work.soon; table q24bownbusns*q13gender /chisq measures; run; proc freq data = work.soon; table q24bownbusns*q15population /chisq measures; run; proc freq data = work.soon; table q24bownbusns*q19atte /chisq measures; run; proc freq data = work.soon; table q24bownbusns*q16maritalstatus /chisq measures; run; proc freq data = work.soon; table q24bownbusns*Agegroup /chisq measures; run; proc freq data = work.soon; table q24bownbusns*education_status /chisq measures; run; proc freq data = work.soon; table q24bownbusns*province /chisq measures; run; proc freq data = work.soon; table q24bownbusns*q /chisq measures; run; proc freq data = work.soon; table q24bownbusns*q24apdwrk

/chisq measures; run; proc freq data = work.soon; table q24bownbusns*geo_type /chisq measures; run; ods pdf close;

APPENDIX B:STATISTICAL ANALYSIS SOFTWARE VERSION 9.2 (SAS) SCRIPTS FOR MODEL FITTING

proc logistic data= work.csv;/* the model i ran on 13/08/2021*/ class q13gender q15population Q14AgeGr province education_status q16maritalstatus; model q24bownbusns(event = 'Yes') = q13gender q15population Q14AgeGr province education_status q16maritalstatus/ aggregate scale= none selection=backward lackfit rsquare; output out = Vester0821 p = pred_prob lower=low upper=upp; run;

APPENDIX D:R-STUDIO SCRIPTS FOR MODEL FITTING

$model.null\ =\ glm(as.factor(Q24BOWNBUSNS)\ 1, data\ =\ data.omit, family\ =\ binomial)model.full\ =\ glm(as.factor(Q24BOWNBUSNS)\ +\ Q13GENDER\ +\ Q14AgeGr+Q15POPULATION+Q16MARITALSTATUS+Education\_Status+\ Province+Geo\_type+Q19ATTE+Q24APDWRK+Status, data = data.omit, family = binomial)step(model.null, scope = list(upper = model.full), direction = "both", test = "Chisq", data = data.omit)$

APPENDIX E:CROSS-TABLE RESULTS

Better access to loans

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Yes	491	31.6	31.6	31.6
	No	1064	68.4	68.4	100.0
	Total	1555	100.0	100.0	

Figure 5.1: Distribustion of business owners with Better Access to loans.

Age group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	16-25	93	6.0	6.0	6.0
	26-35	379	24.4	24.4	30.4
	36-45	425	27.3	27.3	57.7
	46-55	386	24.8	24.8	82.5
	56-65	212	13.6	13.6	96.1
	66-75	44	2.8	2.8	99.0
	More than 75	16	1.0	1.0	100.0
	Total	1555	100.0	100.0	

Figure 5.2: Distribustion of business owners by Age group.

Population group

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	African/Black	1445	92.9	92.9	92.9
	Coloured	41	2.6	2.6	95.6
	Indian/Asian	21	1.4	1.4	96.9
	White	48	3.1	3.1	100.0
	Total	1555	100.0	100.0	

Figure 5.3: Distribustion of business owners by Population group.

Gender

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	783	50.4	50.4	50.4
	Female	772	49.6	49.6	100.0
	Total	1555	100.0	100.0	

Figure 5.4: Distribustion of business owners by Gender.

Province

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Western Cape	80	5.1	5.1	5.1
	Eastern Cape	183	11.8	11.8	16.9
	Northern Cape	17	1.1	1.1	18.0
	Free State	109	7.0	7.0	25.0
	KwaZulu-Natal	217	14.0	14.0	39.0
	North West	96	6.2	6.2	45.1
	Gauteng	305	19.6	19.6	64.8
	Mpumalanga	210	13.5	13.5	78.3
	Limpopo	338	21.7	21.7	100.0
	Total	1555	100.0	100.0	

Figure 5.5: Distribution of business owners by Population.